U.PORTO

FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

Utilização de *Big Data* na Estimação Rápida de Indicadores Macroeconómicos

José Pedro Ribeiro Silva

Dissertação

Mestrado em Economia e Administração de Empresas

Orientado por

Professor Doutor Pedro Moreira de Campos

Coorientado por

Professora Doutora Cecília Castro

2020

Nota Biográfica

José Pedro Ribeiro Silva nasceu em Vila Nova de Famalicão, distrito de Braga, a 10 de Fevereiro de 1996.

Licenciou-se em Economia em 2018 na Faculdade de Economia da Universidade do Porto. Com objetivo de complementar a sua formação, no mesmo ano ingressou no Mestrado em Economia e Administração de Empresas, também na Faculdade de Economia da Universidade do Porto.

Em termos profissionais, iniciou-se em 2019 como Consultor de CRM (*Customer Relationship Management)* na Hitachi Solutions Europe, Ltd, onde desempenha funções até à data.

Agradecimentos

Ao longo desde ano contei com o apoio de várias pessoas, especialmente importante em termos motivacionais e de suporte, que permitiu tornar esta dissertação possível. Desta forma, gostaria de lhes apresentar os meus mais sinceros agradecimentos.

Gostaria de começar pelas duas pessoas que mais diretamente tiveram impacto nesta dissertação, nomeadamente ao Professor Doutor Pedro Moreira de Campos e à Professora Doutora Cecília Castro, marcado por toda a preocupação, acompanhamento e partilha de conhecimento durante este período. Por esse motivo deixo um sincero agradecimento pela orientação.

Aos meus pais, Teresa e Domingos, que toda a vida motivaram o meu percurso académico e me proporcionaram todas as condições para perseguir os meus objetivos. Mais recentemente, a ingressar no desafio de realizar mestrado, e principalmente no último ano, por todo o apoio incondicional para a realização deste trabalho.

À minha namorada, Luísa, agradeço toda a paciência, preocupação e motivação nos dois anos do mestrado. Um agradecimento especial pelo último ano letivo cujo suporte me permitiu ultrapassar todos os desafios académicos e profissionais.

Por último, aos meus amigos que me acompanharam neste percurso e me motivaram e ajudaram a superar todos os desafios. Um especial obrigado aqueles que me acompanharam nesta última etapa com palavras de motivação e entusiasmo.

Resumo

As necessidades de informação dos agentes económicos para apoiarem as suas decisões em contexto macroeconómico tem evoluído nos últimos anos. Esta preocupação tem atingido especialmente entidades como os bancos centrais, os quais têm traçado como objetivo a redução do prazo publicação de estimativas rápidas preliminares do Produto Interno Bruto (PIB) de 45 para 30 dias após o término do período de referência. No entanto, é essencial que esta redução não implique perda na qualidade da informação, que consequentemente teria efeitos negativos na economia.

Assim, um dos desafios nesta tarefa trata-se de encontrar novos indicadores (fontes de informação) que permitam fazer uma correta avaliação do PIB e estejam disponíveis para utilização antes dos indicadores convencionais. Desta forma, bases de *Big Data* como dados de transações eletrónicas podem ajudar a resolver este problema.

Neste trabalho, dados de compras em terminais de pagamento automático, levantamentos em caixas multibancos e dados provenientes do e-fatura serão utilizados como *proxys* do PIB para prever a sua variação da taxa de crescimento. Serão utilizados nesta análise três modelos com propriedades distintas, nomeadamente o modelo ARIMAX, o modelo VAR e o modelo NNARX.

Os resultados revelam que a melhor previsão é obtida pelo modelo ARIMAX utilizando os dados provenientes da rede multibanco. Foi ainda obtida uma estimativa rápida para o último trimestre com dados disponíveis, a qual foi consistente com o valor real observado do PIB, corroborando a hipótese de incorporação de dados de pagamento eletrónico na estimação rápida do PIB trimestral português como um indicador indireto.

Palavras-chave: Estimativas Rápidas; Dados de Multibanco; E-fatura; ARIMAX; VAR; NNARX.

Classificação JEL: C32, C53, E01, E27

Abstract

The information needs of economic agents to support their decisions in a macroeconomic context have evolved in recent years. This concern has particularly affected entities such as central banks, which have aimed to reduce the publication deadline for preliminary rapid estimates of Gross Domestic Product (GDP) from 45 to 30 days after the end of the reference period. However, it is essential that this reduction does not imply loss of information quality, which consequently would have negative effects on the economy.

Thus, one of the challenges in this task is to find new indicators (sources of information) that allow for a correct assessment of GDP and are available for use before conventional indicators. In this way, Big Databases like electronic transaction data can help to solve this problem.

In this work, data from purchases at automatic payment terminals, withdrawals from ATMs, and data from the e-invoice will be used as proxies for GDP to predict its growth rate variation. Three models with different properties will be used in this analysis, namely the ARIMAX model, the VAR model, and the NNARX model.

The results reveal that the best forecast is obtained by the ARIMAX model using data from the ATM network. A flash estimate was also obtained for the last quarter with available data, which was very consistent with the actual observed value of GDP, corroborating the hypothesis of incorporating electronic payment data in the rapid estimation of Portuguese quarterly GDP as an indirect indicator.

Keywords: Flash Estimates; ATM Data; E-Fatura; ARIMAX; VAR; NNARX.
JEL classification: C32, C53, E01, E27

Índice

Nota Biográfica i

Agradecimentos ii

Resumo iii

Abstract iv

Índice de Figuras vii

Índice de Tabelas viii

Lista de Acrónimos ix

1. Introdução 1

2. Revisão de Literatura 5

2.1 Estimativas Rápidas 5

2.1.1 Informação no curto prazo 5

2.1.2 Definição de estimativas rápidas 6

2.1.3 Estimativas rápidas do PIB trimestral 9

2.1.4 Seleção de indicadores 11

2.1.5 Abordagens direta e indireta 11

2.2 *Big Data* na estimação macroeconómica 15

2.2.1 Novas fontes de informação 15

2.2.2 Dados de Pagamentos Eletrónicos 16

3. Exposição do Problema, Dados e Metodologia 19

3.1 Exposição do Problema 19

3.2 Dados 20

3.2.1 Considerações sobre os dados 21

3.3 Metodologia 22

3.3.1 Séries Temporais 22

3.3.2 Decomposição de séries temporais 23
3.3.3 Estacionariedade das séries temporais 24
3.3.4 Modelo ARIMA 26
3.3.5 Modelo de Vetor Autorregressivo (VAR) 30
3.3.6 Auto Regressão de Rede Neuronal (NNAR) 32
3.3.7 Seleção dos Modelos 35
4. Análise e Discussão dos Resultados 36
4.1 Análise Exploratória dos Dados 36
4.1.1 Análise e Descrição dos Dados 36
4.1.2 Análise Multivariada 38
4.2 Transformação dos Dados 42
4.3 Identificação dos Modelos 44
4.3.1 Modelo ARIMAX 45
4.3.2 Modelo VAR 48
4.3.1 Modelo NNARX 51
4.4 Seleção do Modelo 53
Conclusão 57
Bibliografia 60
Anexos 65

Índice de Figuras

Figura 1 - Sequência de produtos estatísticos associados às estimativas rápidas.....7
Figura 2 - Abordagem bottom-up - componentes desagregados.....10
Figura 3 - Taxas de crescimento do PIB (em %) da EA e da EU em t+30, t+45 e t+65 14
Figura 4 - Rede Neuronal com quatro nós inferiores, um nó na camada superior e quatro nós na camada intermédia.....33
Figura 5 - Arquitetura Paralela e Paralela em série do modelos NNAR.....34
Figura 6 - Séries temporais individuais e a sua sazonalidade decomposta 37
Figura 7 - Séries temporais individuais corrigidas de sazonalidade e a sua decomposição...37
Figura 8 – Representação do Conjunto das Séries Temporais.....39
Figura 9 - Gráficos de Dispersão em Pares 40
Figura 10 - Gráficos Comparativos de PIB com Compras, Levantamentos e Fatura.....41
Figura 11 - Diagramas de Caixa após a primeira diferença 42
Figura 12 – Representação das séries após transformação.....43
Figura 13 – Ajustamento do Modelo 1 46
Figura 14 – Resíduos do Modelo 1 46
Figura 15 – Previsões do Modelo 1 46
Figura 16 - Ajustamento do Modelo 2 47
Figura 17 - Resíduos do Modelo 2.....48
Figura 18 – Ajustamento do Modelo 3 49
Figura 19 – Resíduos do Modelo 3 49
Figura 20 – Previsões do Modelo 3 49
Figura 21 – Ajustamento do Modelo 4 50
Figura 22 – Resíduos do Modelo 4 50
Figura 23 – Previsões do Modelo 4 50
Figura 24 – Ajustamento do Modelo 5 51
Figura 25 – Resíduos do Modelo 5 52
Figura 26 – Previsões do Modelo 5 52
Figura 27 – Ajustamento do Modelo 6 53
Figura 28 – Resíduos do Modelo 6 53
Figura 29 – Previsões do Modelo 6 53
Figura 30 - Ajustamento do Modelo Final.....54
Figura 31 - Previsão do Modelo Final 56

Índice de Tabelas

Tabela 1 - Sumário Estatístico das Séries Temporais Individuais.....38
Tabela 2 – Coeficientes de Correlação de Pearson.....44
Tabela 3 - Definição dos modelos estimados.....44
Tabela 4 - Critério RMSE.....54
Tabela 5 - Comparação dos valores originais e estimados para o Modelo Final.....55

Lista de Acrónimos

ACF – Função de Autocorrelação (*Autocorrelation Function*)

ADF – Dickey-Fuller Aumentado (Augmented Dickey-Fuller)

AIC – Critério de Informação Akaike

AICc – Critério de Informação Akaike Corrigido

AR – Autorregressivo (*Autoregressive*)

ARIMA – Modelo Autorregressivo Integrado de Médias Móveis (*Autoregressive Integrated Moving Average*)

ARMA – Modelo Autorregressivo de Médias Móveis (*Autoregressive Moving Average*)

BIC – Critério de Informação Bayesian

EA – Zona Euro (*Euro Area*)

EU – União Europeia (*European Union*)

EUA – Estados Unidos da América

GDP – *Gross Domestic Product*

INE – Instituto Nacional de Estatística

MA – Médias Móveis (*Moving Average*)

NNAR – Auto Regressão de Rede Neuronal (Neural Network Autoregressive)

OECD – Organização para a Cooperação e Desenvolvimento Econômico (*Organisation for Economic Cooperation and Development)*

PACF – Função de Autocorrelação Parcial (*Partial Autocorrelation Function*)

PIB – Produto Interno Bruto

PV – Postos de Venda

RMSE - Raiz Quadrada do Erro Médio (*Root Mean Square Error*)

SEE – Sistema Estatístico Europeu

TPA – Terminais de Pagamento Automático

VAR – Vetor Autorregressivo (*Vector Autoregressive*)

1. Introdução

Ter acesso a informações corretas no momento mais oportuno sempre foi uma preocupação na tomada de decisões. Assim, tal como identificado por Bell (1976), uma sociedade pós-industrial é constituída com base em conhecimento e na sua capacidade de o gerar, transformar e distribuir informação rapidamente.

Desde então, é possível encontrar uma vasta gama de literatura ligada à questão de os agentes económicos que operam nos mercados (sejam eles financeiros, governamentais, bancos centrais ou outros) valorizarem cada vez mais o acesso a informação em tempo real que lhes permita fazer uma correta avaliação das condições económicas nos seus contextos de atuação (Charpin, Mathieu, & Mazzi, 2008; Galbraith & Tkacz, 2007; Giannone, Reichlin, & Small, 2008; Golinelli & Parigi, 2008; Modugno, Soybilgen, & Yazgan, 2016; Radovan, 2017).

Nesse sentido, as estimativas rápidas surgem como um produto estatístico para a antecipação da estimação de tendência de indicadores macroeconómicos quando o valor presente dos mesmos ainda não está disponível (Buono, Mazzi, Kapetanios, Marcellino, & Papailias, 2017). Os diferentes produtos obtidos contam no seu cálculo com metodologias e dados diferentes, criando um *trade-off* entre a pontualidade e relevância/precisão das estimativas (Barcellan, 2010)

O Produto Interno Bruto (PIB) apresenta-se como um dos mais relevantes indicadores económicos para as entidades políticas e financeiras (Eurostat, 2016b), uma vez que permite analisar facilmente o estado de saúde de uma economia. Nesse sentido as principais autoridades estatísticas internacionais têm trabalhado para reduzir o prazo de publicação deste e outros indicadores (como a taxa de desemprego ou inflação) através de estimativas rápidas.

Esta preocupação chamou a atenção do Eurostat (2016b) (entidade que coordena as publicações do PIB da União Europeia (EU) e Zona Euro (EA)), sendo que em maio de 2013 foi formada uma *task force*, tendo como objetivo *"to assess whether it is feasible to estimate a sufficiently reliable flash GDP for the euro area and the European Union, based on the information available at t+30 days*[1]*, including internal national estimates"*.

[1] A notação $(t + x)$ significa que as estimativas estão disponíveis x dias após o fim do período de referência t

O resultado da *task force* consiste em dois *working papers* onde são expostos os principais métodos e técnicas de estimação, estrutura de dados disponíveis e uma série de opiniões e sugestões deixadas pelo grupo de trabalho de forma a capacitar as contas nacionais dos diferentes países da União Europeia a estimar o PIB num menor prazo. Ainda, foram efetuadas estimativas rápidas de teste em t+30 de forma a confirmar a aplicabilidade da estratégia sugerida (Eurostat, 2016a). Estes resultados serão alvo de discussão futura no decorrer deste trabalho.

No caso português, a partir do 1º trimestre de 2007 iniciou-se a divulgação de estimativas rápidas para o PIB trimestral em t+45 (Instituto Nacional de Estatística, 2018).

Devido à estratégia proposta pelo Eurostat (2016a), atualmente o INE está a estudar os resultados da publicação da primeira estimativa preliminar do PIB em t+30, sendo que a primeira publicação da mesma foi efetuada apenas para o segundo trimestre de 2020 (Instituto Nacional de Estatística, 2020c). No relatório publicado com a estimativa rápida em t+30 foi também divulgada a informação de que o INE tem feito testes para a publicação desta estimativa desde o primeiro trimestre de 2017 até ao quarto trimestre de 2019, sendo que o valor médio absoluto de revisão da taxa da variação homóloga do PIB em volume com a publicação a t+60 foi de 0.1 pontos percentuais e nunca excedendo 0.3 pontos percentuais para os 12 trimestres de análise.

Assim, analisando a primeira divulgação desta estimativa em t+30, foi registado uma variação de -14.1% do PIB em cadeia (Instituto Nacional de Estatística, 2020c). A t+45, esta estimativa rápida foi revista para -13.9%, registando apenas uma diferença de 0.2 pontos percentuais (Instituto Nacional de Estatística, 2020a), sendo que à data da divulgação das Contas Nacionais Trimestrais em t+60 a estimativa rápida manteve-se no valor -13.9% (Instituto Nacional de Estatística, 2020b).

De notar que estas estimativas rápidas diferem da divulgação das contas trimestrais em 60 dias em relação ao período de referência (t+60), sendo que a última goza de informação mais completa, precisa e detalhada e incluído informação sobre os restantes agregados (Instituto Nacional de Estatística, 2018).

A estimação trimestral do PIB tem por base a estimação independente de cada agregado macroeconómico, os quais por sua vez são obtidos recorrendo a indicadores. Segundo Esteves e Rua (2012) os indicadores podem ser distinguidos em dois grupos. Enquanto que os indicadores diretos são usados para estimar diretamente o agregado

objetivo (são uma medida do próprio), os indicadores indiretos estão relacionados com a variável a medir apesar de não serem medidas diretas da mesma.

Assim, um dos principais desafios para adiantar a estimação do PIB consiste em encontrar indicadores que apresentem uma relação de correlação forte com a variável a estimar (Charpin et al., 2008). Este é também um dos métodos sugeridos pelo Eurostat (2016b) na prossecução do objetivo principal de reduzir o tempo de estimação preliminar do PIB.

Por outro lado, os avanços tecnológicos das últimas décadas têm permitido uma crescente captação de dados de diferentes fontes e disciplinas, inclusive de cariz financeiro, chamados de *Big Data*. No trabalho de (Buono et al., 2017) são exploradas diversas aplicações de *Big Data* a áreas económicas, tais como finanças e marketing assim como a utilização destes dados na previsão de indicadores macroeconómicos relevantes como o PIB ou o Consumo Privado. Nesse sentido, os dados de pagamentos eletrónicos são uma das aplicações apresentadas para a temática em estudo.

Estes dados podem ter diferentes origens tais como: transações com cartões de débito/crédito, pagamentos por depósitos eletrónicos diretos, pagamentos de débito direto iniciados pelas empresas (que debitam diretamente das contas dos consumidores), transferências de crédito, pagamentos com contas eletrónicas em lojas online, entre outros (Buono et al., 2017).

Exemplos de estudos de exploração deste tipo de dados são encontrados para diversos países, incluindo Portugal (Duarte, Rodrigues, & Rua, 2017; Esteves, 2011), confirmando a sua validade como possíveis fontes de informação para a estimação de variáveis macroeconómicas.

Tendo em conta o contexto descrito, este trabalho terá como objetivo a obtenção de previsões para o Produto Interno Bruto nacional recorrendo a fontes de *Big Data*. Estes dados são provenientes de meios de pagamento, nomeadamente de compras através de terminais de pagamento automático e levantamentos nacionais em caixas de multibanco em Portugal, e ainda de uma nova fonte de dados, ainda em construção, proveniente do sistema e-Fatura.

Uma vez que os dados estão disponíveis nos primeiros dias de cada mês, tal viabiliza a sua utilização na obtenção de estimativas rápidas como indicadores indiretos ou *proxys* da variável em estudo.

Para tal, serão usados três modelos de séries temporais distintos, nomeadamente o modelo ARIMA com inclusão de variáveis exógenas (ARIMAX), o modelo VAR e o modelo NNAR com inclusão de variáveis exógenas (NNARX).

O capítulo 2 deste trabalho contempla a revisão de literatura, onde serão estudados a importância da utilização de estimativas rápidas e a necessidade de informação em tempo real e a clarificação da definição de estimativas rápidas e os diferentes produtos estatísticos provenientes das mesmas. Ainda, serão apresentadas as metodologias atuais de cálculo das estimativas do PIB e restantes agregados tais como a utilização de diferentes indicadores e métodos na estimação indireta do PIB trimestral. Por último, é discutida a utilização de novas fontes de *Big Data* na estimação de indicadores macroeconómicos à luz de trabalhos desenvolvidos no seio desta temática.

No capítulo 3 será exposto o problema que motiva a realização deste estudo. Ainda, serão apresentados os dados utilizados e algumas considerações sobre os mesmos, terminando o capítulo com a exposição da metodologia utilizada para responder ao problema. Neste capítulo serão exploradas as formulações dos modelos ARIMAX, VAR e NNARX assim como a definição de critérios de seleção e avaliação dos modelos.

No capítulo 4 o trabalho recai num primeiro momento sobre a análise exploratória e transformação dos dados temporais utilizados. Posteriormente, serão constituídos conjuntos de treino e teste com os dados e à luz dos modelos descritos serão apresentados os resultados dos ajustamentos obtidos com os dados de treino, assim como feita a validação dos mesmos. Os dados de teste serão utilizados para avaliar a capacidade preditiva dos diferentes modelos. Por último, será identificado entre os modelos a melhor solução para o problema e fazer previsões para o último período com base no modelo selecionado.

Por último, é apresentado um capítulo com as conclusões do trabalho, onde serão discutidas as ideias gerais do trabalho e salientadas as principais ideais e resultados da análise dos dados, assim como sugeridas propostas para futuros trabalhos nesta temática.

2. Revisão de Literatura

Este capítulo é dedicado ao enquadramento teórico e análise de literatura relacionada com a temática em estudo, nomeadamente, a utilização de *Big Data* na estimação rápida de indicadores macroeconómicos.

Na primeira parte deste capítulo será feita uma abordagem às estimativas rápidas e à importância de informação pontual. Na segunda parte, é explorada a utilização de *Big Data* na estimação de indicadores, com foco em trabalhos anteriormente desenvolvidos no seio desta temática e nos dados utilizados nos mesmos.

2.1 Estimativas Rápidas

2.1.1 Informação no curto prazo

Os agentes económicos e investidores financeiros, que atuam e afetam frequentemente os mercados, necessitam cada vez mais de aceder a informação em tempo oportuno de forma a suportar as suas decisões políticas e estratégias de investimento num ambiente macroeconómico ativo (Esteves, 2011; Modugno et al., 2016)

Para melhor anteciparem o comportamento dos mercados a procura por novas fontes de informação financeira com precisão, credibilidade e principalmente, disponíveis num mais curto prazo, é uma tendência da atualidade como resposta a este problema (Duarte et al., 2017).

Esta informação é tanto mais valiosa quanto mais em tempo útil estiver disponível e melhor a sua capacidade preditiva. De outra forma, a formulação de políticas assente em dados temporalmente desfasados dos respetivos períodos/acontecimentos de referência pode trazer efeitos colaterais negativos e ações inadequadas, gerando resultados desfavoráveis para a economia (Radovan, 2017).

No entanto, segundo Giannone et al. (2008), as decisões de politica monetária tem por base indicadores muitas vezes sustentados em inquéritos do estado atual e futuro da economia pelo que os dados podem estar incompletos, apesar de apresentarem resultados historicamente positivos para o efeito (Haluška, 2006).

Segundo Buono et al. (2017), as estimativas rápidas são uma resposta a este problema, surgindo como uma previsão do valor presente de indicadores macroeconómicos antes da

sua publicação oficial, salientando a possibilidade da utilização de fontes de dados alternativas na estimação dos indicadores pretendidos.

Segundo o *working* paper *"Overview of GDP flash estimation methods"*, publicado pelo Eurostat (2016b), apesar de alguns países europeus, tais como o Reino Unido, Bélgica, Lituânia e Espanha, e mais recentemente, Áustria, Letónia e França, já fazerem à data a publicação de estimativas rápidas do PIB em t+30 ou antes (em t+25 no caso do Reino Unido), os restantes países ainda faziam esta divulgação em t+45 dias.

2.1.2 Definição de estimativas rápidas

Como descrito por Barcellan (2010) em "*Flash Estimates of GDP*", as estimativas rápidas devem ser entendidas como uma sequência de produtos estatísticos. Estendendo esta definição, no site de *Statistics Explained* (Eurostat, 2012), encontra-se um glossário de estimativas rápidas, onde estas são consideradas como avaliações oportunas de uma variável económica de interesse, usando geralmente dados recolhidos em tempo real[2], para captar uma melhor imagem de um futuro próximo (*forecasting*), do presente (*nowcasting*) ou do passado recente (*flash estimates*).

Para melhor entendimento da matéria em estudo, torna-se necessário fazer uma distinção entre os produtos estatísticos mencionados. Segundo o mesmo glossário, podemos definir:

- *Forecasting*: processo de fazer declarações sobre eventos futuros no momento presente, os quais o resultado ainda não foi observado;
- *Nowcasting*: processo de avaliação rápida que acontece no período de referência corrente T (presente) de uma determinada variável de interesse no mesmo momento T. O processo de *nowcasting* faz uso de dados em tempo real, captados com base no período ente T e T-1;
- *Flash Estimate:* conceito mais comumente associado a estimativas rápidas, trata-se da estimação para uma variável macroeconómica de interesse após o fim do respetivo período de referência. Geralmente, estas estimativas são calculadas a partir de modelos econométricos e tem uma divulgação anterior relativamente à primeira estimativa real da variável em consideração.

[2] Definição de dados em tempo real: Informação disponível para os utilizadores num momento específico no tempo. **Fonte:** *Eurostat Statistics Explained* (Eurostat, 2012).

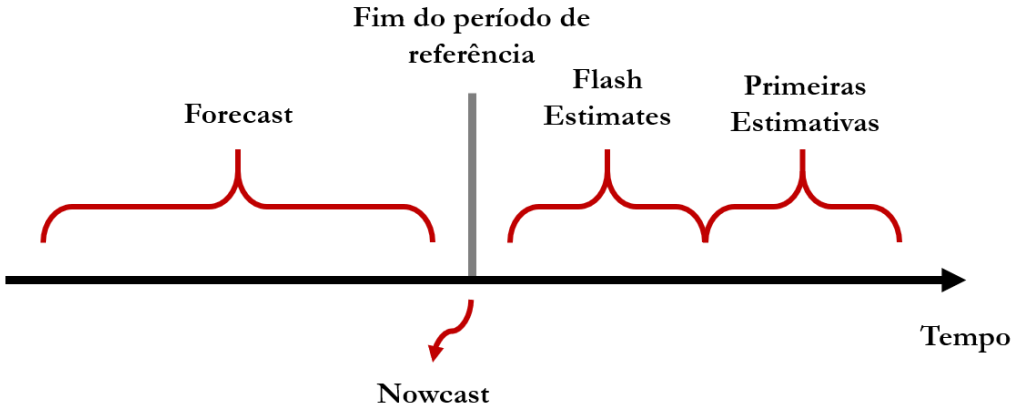
Tomando como exemplo a estimação do crescimento trimestral do PIB, a sequência de produtos estatísticos pode atravessar as seguintes fases (Barcellan, 2010):

1. *Forecasting*;
2. *Nowcasting*;
3. *Flash estimates*/estimativas rápidas;
4. Primeiras estimativas oficiais;
5. Sucessivas estimativas e revisões com novos dados.

Na Figura 1 é ilustrada a distribuição temporal das estatísticas mencionadas, tendo o final do período em questão como referência. Analisando a Figura 1, podemos inferir que:

- Quanto mais avançado no eixo temporal se encontra o ponto de observação, maior a quantidade de informação estatística disponível relativamente ao período de referência;
- Quanto mais avançado no eixo temporal se encontra o ponto de observação, menor é a base de métodos e técnicas econométrica/estatísticas incluídos no processo;
- Quanto mais próximo o ponto de observação se encontra do fim do período de referência, maior a importância do *nowcasting* em função do *forecasting*;
- Quanto mais afastado o ponto de observação se encontra do fim do período de referência, maior a importância das estimativas reais em favor das *flash estimates*.

Figura 1 - Sequência de produtos estatísticos associados às estimativas rápidas



Fonte: Autoria própria, adaptado de "*Flash Estimates of GDP*" (Barcellan, 2010)

De acordo com Barcellan (2010), o principal objetivo destes produtos trata-se de capacitar o utilizador da informação a prever ou mais rapidamente perceber o rumo da

variável económica de interesse (por exemplo, o PIB) com base na informação disponível à data do seu cálculo. Assim, as estimativas rápidas surgem como um mecanismo de análise económica composta por uma sequência de diferentes estimativas associadas ao indicador em questão.

Esta sequência de produtos é caraterizada principalmente pela diferente informação que o cálculo inclui, podendo variar quanto à sua quantidade, relevância e disponibilidade, o que consequentemente, requer a utilização de métodos e técnicas de estimação diferenciados. Mazzi, Moauro, Montana e Cannata (2016) salientam que apesar dos beneficios de usar dados em tempo real para estimar indicadores macroeconómicos, estes correm o risco de sofrer futuras revisões o que implica por sua vez revisões dos indicadores. Desta forma, a sensibilidade dos modelos devem ser testados de forma a incorporar estes dados sem prejuizos significavos dos valores estimados em futuras revisões.

Daí, resultam diferenças na qualidade dos seus apuramentos uma vez que existe um *trade-off* entre a pontualidade da informação e a precisão da mesma (Castle, Fawcett, & Hendry, 2009), sendo que atualmente o foco do Eurostat recai sobre a pontualidade das estimativas (Mazzi et al., 2016).

No entanto, continuando a ter como exemplo a estimação rápida do PIB trimestral, o progressivo aumento de informação estatística disponível ao longo do tempo, permite reforçar o cálculo do mesmo (Barcellan, 2010). Ainda, revisões a esta estimação vão sendo efetuadas aquando da disponibilização de nova informação sobre os principais indicadores macroeconómicos relacionados (Cardoso & Rua, 2011; Castle et al., 2009).

Os diferentes produtos estatísticos relativos ao momento anterior e posterior ao período de referência são, normalmente, direcionados a diferentes entidades com os mais variados fins e destinatários da informação. Tal como descrito no *working paper* do Eurostat (2016b), o principal utilizador das estimativas do PIB é a Comissão Europeia e o Banco Central Europeu com o fim de apoiar na estruturação de políticas e controlo do estado atual da economia. No entanto, existem outras entidades interessadas nesta informação, tais como autoridades estatísticas nacionais, organizações privadas (por exemplo, analistas económicos e jornalistas) e decisores políticos (Chamberlin, 2009).

No que se refere às autoridades e institutos nacionais de estatística, o foco encontra-se nos últimos produtos da sequência, nomeadamente, as estimativas *flash* e posteriores (Instituto Nacional de Estatística, 2018).

No trabalho de Barcellan (2010), este aponta o limite de interesse dos institutos nacionais de estatística entre o *nowcasting* e *flash estimates* devido ao carácter oficial das publicações destas entidades, sendo que dificulta a aceitação de novas fontes de informação como válidas para a estimação mais oportuna de agregados macroeconómicos.

No entanto, Chamberlin (2009) apesenta os decisores políticos que tem de tomar decisões em tempo real como das entidades mais interessadas nas estimativas rápidas. Desta forma, tendo em conta o *trade-off* entre a pontualidade e qualidade da informação, considera mais importante a tomada de decisão com dados limitados do que na total ausência dos mesmos.

Neste trabalho, adotaremos a terminologia de estimativas rápidas referente às estimativas obtidas após o período de referência e antes da publicação das estatísticas oficiais.

2.1.3 Estimativas rápidas do PIB trimestral

Segundo o Eurostat (2016b), o PIB trimestral é provavelmente o indicador de curto prazo mais relevante produzido pelo Sistema Estatístico Europeu (SEE). Devido ao valor que a estimativa trimestral do PIB traz aos agentes económicos, este torna-se o indicador no qual recai a maior parte da literatura referente às estimativas rápidas e previsão.

A publicação da taxa de crescimento do PIB trimestral na União Europeia é coordenada pelo próprio Eurostat, sendo que acontece em vários momentos distintos (Eurostat, 2020). Recentemente, a t+30 é passou a ser divulgada uma estimativa rápida preliminar para a União Europeia e Zona Euro. A t+45 é publicada a estimativa rápida, a t+65 as primeiras estimativas reais para o PIB e os principais agregados e em t+90 as estimativas finais dos indicadores anteriores (incluindo revisões às divulgações passadas) (Eurostat, 2016a).

Apenas em 2003 o Eurostat (2016a) publicou pela primeira vez estimativas rápidas do PIB trimestral, com um desfasamento de 45 dias após a data de término do período de referência (t+45). Apesar dos esforços para a redução deste prazo para o crescimento do PIB para t+30, é de notar que os restantes agregados económicos são estimados para t+65.

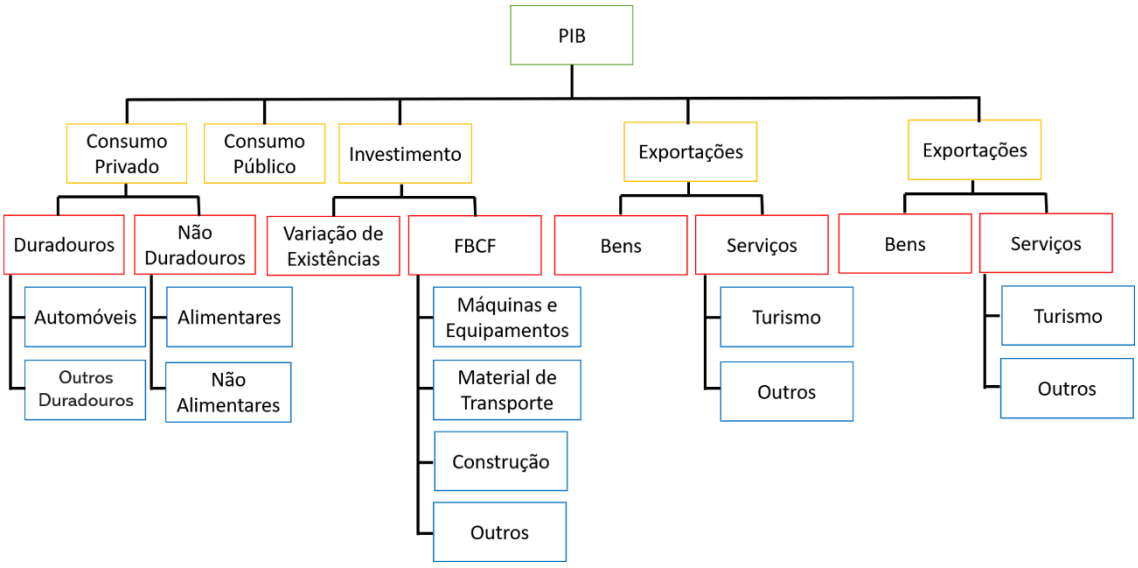
Apesar de os decisores políticos e investidores justificarem a necessidade de estimativas rápidas devido à necessidade de informação atempada, foi principalmente com a recente crise económica que cresceram as motivações para tornar esta estimação mais eficiente (Golinelli & Parigi, 2008).

Outro dos motivos trata-se do alinhamento do lançamento do PIB trimestral com os EUA, onde este indicador já é calculado em t+30/60/90 dias (Eurostat, 2016b; Mazzi et al., 2016). Também o alinhamento entre os vários países da União Europeia e Zona Euro é uma das motivações importantes a considerar.

Em Portugal, as primeiras estimativas rápidas em t+45 foram publicadas pelo INE para o 1º trimestre de 2007 (Instituto Nacional de Estatística, 2018). No sentido de se alinhar com o Eurostat, está atualmente a ser estudada a publicação de estimativas preliminares em t+30, sendo que a primeira divulgação ocorreu para o 2º trimestre de 2020 (Instituto Nacional de Estatística, 2020c).

De acordo com Esteves e Rua (2012) a estimação rápida do PIB português é baseada numa abordagem *bottom-up* em dois níveis, em que para cada componente na ótica da despesa (consumo privado, consumo público, investimento, exportações e importações) é necessária uma previsão para posterior agregação de forma a obter a previsão para o PIB. Num segundo nível é considerada uma desagregação superior dos componentes referidos (Dias, Pinheiro, & Rua, 2016). Na Figura 2 está representada de forma simplificada a desagregação dos componentes pela abordagem *bottom-up* que o Banco de Portugal considera para nesta estimação.

Figura 2 - Abordagem *bottom-up* - componentes desagregados



Fonte: Autoria própria, adaptado de "Previsões Económicas de Curto Prazo para Portugal: uma Síntese Metodológica" de Esteves e Rua (2012)

2.1.4 Seleção de indicadores

Na ausência de informação para todos os indicadores no momento de estimação do PIB trimestral uma das soluções apresentadas é a utilização de indicadores indiretos ou *proxys* (que estão relacionados com a componente a estimar, mas não é uma medida direta da mesma) que reflitam o comportamento das variáveis macroeconómicas de forma semelhante (Esteves & Rua, 2012).

Num estudo feito por Charpin, Mazzi e Moauro (2016), usando modelos de correção de erros para estimar as taxas de crescimento do PIB trimestral dos principais países europeus, esta relação é denominada de cointegração.

Os indicadores devem ser selecionados de entre o conjunto de dados que leiam o mesmo fenómeno da variável macroeconómica em questão. Para além disso, é necessário estudar se o comportamento destes indicadores após transformação (por exemplo, transformação logarítmica) é similar ao do relativo componente em estudo (Eurostat, 2016a).

Segundo Mazzi et al. (2016) estes são chamados de indicadores coincidentes, em que a sua principal vantagem reside no facto de estarem sujeitos a menos restrições do que as estimativas rápidas.

Esteves and Rua (2012) salientam que na escolha de indicadores deve ter-se em atenção dois aspetos – apresentar uma relação com o fenómeno da variável macroeconómica objetivo e ainda apresentar uma relação estatística forte dentro dos dados da amostra entre os indicadores e a variável.

Assim, no trabalho desenvolvido por Rua (2004) encontramos a proposta de um indicador coincidente para o PIB utilizando oito séries temporais com indicadores distintos. Como resultado, apesar de nem todas as séries apresentarem uma relação forte com o PIB, foi possível construir um indicador mensal consistente com a tendência de crescimento do PIB.

2.1.5 Abordagens direta e indireta

A metodologia utilizada é diferente para os indicadores disponíveis, sendo que têm de ser estimados através de métodos diretos ou indiretos, baseados em modelos econométricos adequados (Eurostat, 2016a):

- Abordagem direta: o último trimestre de um determinado indicador é

estimado diretamente;

- Abordagem indireta: o último trimestre é extrapolado usando procedimentos de desagregação temporal. Esta divide-se em duas etapas, sendo que inicialmente o indicador do último trimestre é extrapolado e depois dá-se a desagregação temporal para a variável de interesse.

Como em Charpin et al. (2016), os métodos utilizados na abordagem indireta exploram as relações de correlação entre um ou vários indicadores e o agregado relacionado do qual se pretende prever o comportamento (por exemplo, o PIB ou o consumo privado).

Ainda, no trabalho Charpin et al. (2016), as abordagens diretas e indiretas foram comparadas na estimação da taxa de crescimento do PIB trimestral da zona euro (com base nos resultados do cálculo referido anteriormente). Os resultados indicaram que a abordagem indireta obteve ligeiramente melhores estimativas (cerca de 0.02 pontos percentuais) em relação à abordagem direta.

Este método (abordagem indireta) foi utilizado pela força de trabalho do Eurostat (2016a) para produzir estimativas rápidas de teste de indicadores para a União Europeia e zona euro no sentido de estimar o PIB trimestral em t+30.

Um dos modelos propostos para a estimação em t+30 é o modelo ARIMA (Eurostat, 2016b). Este modelo é considerado adequado para estimar séries temporais com caraterísticas de tendência e sazonalidade encontradas nas séries temporais económicas. No modelo ARIMA (p, d, q), os valores da série são descritos com base no parâmetro p (função da ordem autorregressiva), no parâmetro d (integrado) e q (médias móveis), onde nenhum dos parâmetros assume valores negativos.

Ainda, é possível adicionar regressores estacionários exógenos quando estes melhoram os ajustamentos do modelo, conhecidos como modelos ARIMAX. Este foi um dos modelos utilizados pela *task force* do Eurostat, como descrito em Eurostat (2016a), de forma a calcular as estimativas rápidas do PIB em t+30 devido à falta de informação para estimar o PIB de forma direta.

Outra das metodologias sugeridas em Eurostat (2016b) é o modelo de Vetores Autorregressivos (VAR) proposto inicialmente por Sims (1980), que segundo a análise de Christiano (2012) identifica como principais objetivos na utilização destes modelos:

- A previsão de séries temporais económicas;
- Construção e avaliação de modelos económicos;
- Avaliação das consequências das ações de política económica.

Na análise dos resultados obtidos pelo Eurostat (2016a), devemos notar que entre as 16 estatísticas de teste calculadas (correspondentes ao período entre 2012Q1-2014Q4[3]), as amostras foram divididas entre 2012Q1-2013Q4 e 2014Q1-2015Q4, uma vez que as últimas foram calculadas em tempo real.

Os valores estimados em t+30, t+45, t+65, t+100 e t+3m+65d (3 meses e 65 dias) e as respetivas revisões são apresentados no Anexo 1 e 2 para a EU e EA, respetivamente. A estimativa para o último trimestre cobriu o PIB de cerca de 91% dos 28 países da União Europeia (EU) e 94% dos 19 países da Zona Euro (EA). Considerando o *trade-off* entre a pontualidade das estimativas rápidas e a sua precisão, os resultados foram positivos, tal que:

- A revisão prevista das estimativas em t+30 da taxa de crescimento do PIB comparativamente a t+45 era de ± 0.1 ponto percentual (p.p.). Os resultados mostraram que para 9 das estimativas a revisão foi de 0.0 p.p., enquanto para os restantes 7 foi de ± 0.1.

- As revisões foram testadas contra critérios definidos *a priori*, sendo que a média das revisões ao PIB estimado em t+30 quando publicadas as estimativas em t+45 foi de 0.00 p.p. para a EA e -0.01 p.p. para a EU (critério de qualidade definido para ± 0.05 p.p.). A média das revisões absolutas ao PIB estimado em t+30 quando publicadas as estimativas em t+45 foi de 0.06 p.p. para a EA e 0.05 para a EU (critério de qualidade definido para valores abaixo de 0.10 p.p.). Quando comparado com as estimativas produzidas em t+65 (média absoluta de revisão), foi igualmente de 0.06 p.p. para a EA e 0.05 para a EU (critério de qualidade definido para valores abaixo de 0.13 p.p.).

Na Figura 3 estão representados graficamente os resultados obtidos pelo Eurostat (2016a) descritos anteriormente. Apesar de os resultados favorecerem a estimativa em t+45 em relação à estimativa em t+30, estes são considerados como positivos devido à redução de 15 dias na sua publicação.

No que toca a Portugal, segundo Cardoso e Rua (2011), a metodologia usada no cálculo das Contas Nacionais Trimestrais é indireta para alguns componentes, tendo por base a relação de indicadores indiretos com as Contas Nacionais Anuais. Esta metodologia, utilizada pelo INE, em termos matriciais pode ser equacionada (1.1) como:

[3] A notação AAAAQT será usada tal que AAAA representa o ano e T representa o trimestre

$$Y = X\beta + \varepsilon \tag{1.1}$$

onde Y corresponde a uma série de estimativas trimestrais desconhecidas de determinado componente do PIB, X corresponde às séries trimestrais dos indicadores relacionados e ε é o vetor de resíduos que podem não ser independentes no tempo. Existe, no entanto, uma restrição (1.2) a considerar tal que:

$$\bar{Y} = CY \tag{1.2}$$

onde $\bar{Y}$ corresponde ao valor anual observado do componente do PIB enquanto C trata-se de uma matriz que converte os valores trimestrais deste componente em valores anuais.

Em suma, o valor das contas anuais tem necessariamente de ser reproduzidos pelas contas trimestrais (Instituto Nacional de Estatística, 2018).

Assim, os métodos indiretos, se por um lado permitem a estimação dos valores trimestrais quando os valores anuais são conhecidos, por outro é uma ferramenta que permite a extrapolação dos valores trimestrais quando ainda não estão disponíveis os resultados das contas anuais e a utilização de indicadores indiretos na estimação.

Figura 3 - Taxas de crescimento do PIB (em %) da EA e da EU em t+30, t+45 e t+65



Fonte: Eurostat Statistic Explained – "*Preliminary GDP flash estimate in 30 days for Europe*" (Eurostat, 2019)

2.2 *Big Data* na estimação macroeconómica

2.2.1 Novas fontes de informação

Os esforços das estimações de muito curto prazo centram-se em encontrar indicadores coincidentes que captem informação similar à dos agregados macroeconómicas em questão, mas que permita o seu acesso e utilização mais cedo (Gil, Pérez, Sanchez Fuentes, & Urtasun, 2018). Tal como descrito em Charpin et al. (2016) e Mazzi et al. (2016), tal implica encontrar uma relação de cointegração entre os indicadores.

Por outro lado, os rápidos avanços tecnológicos das últimas décadas tem permitido a captura e tratamento de novos dados em todas as áreas, incluindo a nível de transações financeiras (Dias et al., 2016).

Segundo Buono et al. (2017), o recurso a novas bases de *big data* na produção de estimativas rápidas de indicadores macroeconómicos tem como objetivos:

- Melhorar a qualidade das atuais estimativas rápidas;
- Antecipar a data das divulgações das estimativas;
- Complementar a informação convencional.

Ainda, uma vez que estes dados provêm de uma origem diferente dos convencionalmente utilizados na estimação dos indicadores, tal permite completar a análise dos mesmos sobre uma perspetiva mais granular, tanto na dimensão temporal como secional. No trabalho de Buono et al. (2017), estas vantagens são explicadas, sendo que no que toca à dimensão temporal, a informação de alta frequência possibilita o utilizador de rever as estimativas de várias frequências, como mensal, semanal ou até diária, capacitando-o de informação mais precisa na tomada de decisão.

Por outro lado, na dimensão secional, as bases de dados podem conter informações relevantes sobre as unidades de estudo, tais como regiões ou setores, que podem não ser cobertas pelos indicadores tradicionais.

Idealmente, as bases de dados devem estar disponíveis logo após o término do período de forma a que a sua incorporação imediata no modelo, ou seja, a sua divulgação deve ser em prazo inferior aos indicadores convencionais (Esteves, 2011).

2.2.2 Dados de Pagamentos Eletrónicos

Com o desenvolvimento dos meios de pagamento eletrónicos e informatizados, existe uma crescente captação de dados de transações financeiras de elevada frequência, especialmente em Portugal (Duarte et al., 2017; Esteves, 2011).

Estes dados são essencialmente provenientes de Terminais de Pagamento Automático/Multibanco (TPA) e Pontos de Venda (PV), que garantem um grande volume de dados com informação relevante e precisa (livre de erros de medição) (Esteves, 2011).

Assim, de acordo com Duarte et al. (2017) estes tratam-se dos indicadores mais destacados na literatura como covariante do consumo privado que, para além da abundância de registos, apresentam dois grandes motivos que qualificam esta escolha:

- Os levantamentos de multibanco e os pagamentos cartão de débito/crédito estão diretamente associados ao consumo privado;
- Esta informação está disponível de forma muito rápida.

Segundo Duarte et al. (2017), o caso português é conhecido mundialmente, tanto a nível de inovação e funcionalidades, como da própria rede de TPAs e PVs. Um fator a salientar é de que a lei portuguesa proíbe os bancos e os mercadores de cobrarem taxas pelas operações em TPAs e PVs. Os dados retirados do mesmo trabalho relatam que em 2013, em Portugal, o número de Terminais de Pagamento Automático era de 1.540 e o número de Pontos de Venda 24.471. Ainda, salienta que o montante de dinheiro em circulação por estes meios por residentes era à data cerca de 50% do PIB e 70% do Consumo Privado.

Um dos primeiros estudos sobre a aplicação destes dados na estimação do consumo privado em Portugal foi o trabalho de Esteves (2011), onde é proposta a utilização de dados provenientes de TPAs/PVs na estimação do consumo privado. Os resultados deste estudo suportam o uso desta informação como meio de melhorar as estimativas do consumo privado não duradouro (assim como o uso do indicador de vendas no comércio a retalho).

Mais recentemente, Duarte et al. (2017), com base no trabalho de (Esteves), confirmaram a relevância da utilização de dados de sistemas de pagamento eletrónico como um indicador do consumo privado trimestral em Portugal, utilizando o modelo de regressão MIDAS. Concluíram ainda que a incorporação destes dados na estimação da variável em questão permite obter melhores resultados na estimação rápida e previsão do que a maioria dos indicadores convencionais. Ainda, destacou que a utilização dos dados das transações eletrónicas a nível diário não permite obter resultados robustos.

A nível internacional, encontram-se ainda outros trabalhos pioneiros neste tema. No Canada, Galbraith e Tkacz (2007) utilizaram dados de transações económicas como indicadores de alta frequência da atividade económica. Como resultado, encontraram evidência que os dados de cartões de débito conduzem a melhoramentos na previsão da taxa de crescimento do PIB do Canada, assim como prever possíveis revisões feitas ao PIB. Estes resultados também se confirmam para o consumo privado.

No caso de Espanha, Gil et al. (2018) utilizaram dados dos indicadores tradicionais, dados de cartões de crédito e dados de pesquisa online (*Google Trends)* para fazer estimação rápida do consumo privado utilizando modelos de frequência mista. Os resultados salientam que apesar de através de indicadores tradicionais se obterem boas estimativas, os dados de pagamentos com cartões e de pesquisas online (*Google Trends)* permitem melhorar as mesmas, dando maior relevância ao primeiro.

Na Índia, foram utilizados os dados de TPA e PV assim como de um sistema de liquidação bruta em tempo real (novo sistema de pagamentos) de forma a fazer previsão do PIB com recurso a modelos ARIMA (Raju & Balakrishnan, 2019). Apesar de o estudo ter sido efetuado num momento em que ocorreu um grande desenvolvimento dos sistemas de pagamento no país, as conclusões foram positivas no que toca à capacidade destes indicadores preverem o crescimento do PIB.

Ainda, na Colômbia, León e Ortega (2018) utilizando um modelo de redes neural com variáveis exógenas, sendo elas dados do número e valor de transferências eletrónicas entre indivíduos, firmas e o banco central e cheques processados, com objetivo de prever um indicador de atividade económica no curto prazo. Os resultados sugerem que estes dados permitem a obtenção de previsões boas dois meses antes da publicação dos valores oficiais.

Os resultados dos trabalhos para os países enunciados são também consistentes com outros estudos como para Itália (Aprigliano, Ardizzi, & Monteforte, 2017) e Holanda (Verbaan, Bolt, & van der Cruijsen, 2017). Na Dinamarca (Carlsen & Storgaard, 2010) os dados de pagamentos eletrónicos foram usados para prever o indicador de vendas do comércio de retalho.

As conclusões destes estudos salientam o potencial da utilização de novas fontes informação para efeitos de modelização e previsão de variáveis macroeconómicas. No entanto, apesar dos resultados positivos, a informação destes indicadores não é incluída na estimação rápida do PIB em Portugal.

Também, percebe-se que a escolha de modelos adequados para a estimação com base na frequência destes dados é um fator de extrema relevância para o sucesso e valor acrescentado desta estimação.

3. Exposição do Problema, Dados e Metodologia

Efetuada a revisão de literatura, importa esclarecer as questões de investigação a abordar neste trabalho, assim como descrever a metodologia e dados utilizados no estudo.

3.1 Exposição do Problema

Conforme discutido no capítulo anterior, a importância do acesso a informação de cariz macroeconómico em tempo oportuno tem preocupado as diferentes entidades estatísticas, sobretudo na União Europeia, motivando a procura por novas fontes de informação (Eurostat, 2016a, 2016b).

Por outro lado, os avanços tecnológicos têm vindo a permitir uma maior capacidade na captação dados disponíveis em tempo real, inclusive de transações financeiras, para aplicação em temáticas económicas (Duarte et al., 2017; Esteves, 2011; Gil et al., 2018; Raju & Balakrishnan, 2019)

Assim, assente nestas duas premissas, está formulado o problema da falta de pontualidade da informação em função da oportunidade de exploração de novos dados financeiros.

Este trabalho tem como objetivo principal a obtenção de previsões de estimativas rápidas para o PIB, Produto Interno Bruto nacional, robustas o suficiente que corroborem a utilidade dos dados utilizados para a estimação deste indicador. De forma a atingir esse fim, são utilizados e comparados três famílias de modelos de séries temporais, nomeadamente o modelo ARIMA (Autoregressive Integrated Moving Average) com inclusão de variáveis exógenas, ARIMAX, o modelo VAR (Vector Autoregressive) e o modelo NNAR (Neural Network AutoRegressive) com inclusão de variáveis exógenas (NNARX). Para a obtenção dos parâmetros destes modelos, são utilizadas novas bases de dados financeiras, caraterizadas por potenciarem uma elevada frequência e volume de observações disponibilizadas num curto espaço de tempo, usadas como variáveis exógenas dos modelos, sempre que se aplique.

Os dados que neste estudo são essencialmente provenientes de dois meios de pagamento e do e-Fatura, caraterizados pela elevada frequência de observações (mensais).

Numa fase inicial será estudada a validade destes indicadores como uma *proxy* na estimação do PIB, tendo em consideração que os indicadores em estudo são diretamente

relacionados com o consumo privado, agregado que segundo Gil et al. (2018) representa para os países da OECD cerca de 60 a 80% do valor do PIB.

Para a definição do modelo mais adequado, o procedimento utilizado neste trabalho de investigação, passa por recorrer a conjuntos de dados de treino e de teste, na identificação do modelo que permite o melhor ajustamento e previsões para, posteriormente, tendo em conta os dados em estudo, se obter o modelo final, cujos parâmetros serão estimados usando todas as observações disponíveis.

3.2 Dados

Neste estudo serão utilizados novas fontes de informação que convencionalmente não são aproveitados no cálculo das contas nacionais. No entanto, uma vez que todos os dados apresentados sugerem uma aparente relação direta com o fenómeno do consumo privado, considera-se importante, à semelhança de outros estudos, confirmar a validade da sua utilização na procura por fontes de dados alternativas.

São utilizados neste trabalho dados do valor de Compras através de terminais de pagamento automático (TPA) em Portugal (em milhões de euros), do valor de Levantamentos nacionais em caixas multibanco em Portugal (em milhões de euros) e os dados provenientes do e-fatura agrupados a nível nacional (em milhões de euros).

No restante trabalho, para facilidade de leitura, a designação adotada para cada variável é:

- *PIB* – Produto Interno Bruto;
- *Compras* - Compras através de terminais de pagamento automático em Portugal;
- *Levantamentos* - Levantamentos nacionais em caixas multibanco em Portugal;
- *Fatura* - Dados do e-Fatura agrupados a nível nacional.

Os dados de *PIB*, a variável dependente no estudo, encontram-se divulgados em frequência trimestral, sendo que existem observações desde janeiro de 1995.

Por outro lado, para os dados de *Compras, Levantamentos* e *Fatura*, a frequência de divulgação é mensal, pelo que foi efetuada uma agregação dos dados de forma a igualar a frequência trimestral da variável *PIB*.

No que toca aos dados de *Compras* e *Levantamentos,* existem observações publicamente disponíveis desde Setembro de 2000 até à atualidade. Apesar destes dados ficarem disponíveis para utilização para as entidades estatísticas alguns dias depois de ocorrerem (em frequência

diária, inclusive) os dados de *Compras* e *Levantamentos* são divulgados publicamente em t+20 dias pelo Banco de Portugal. No entanto, é ainda possível obter publicamente estes dados diretamente na secção de estatísticas do *website* oficial do Multibanco[4], sendo que os dados são atualizados trimestralmente, normalmente 5 ou 6 dias depois do fim do trimestre.

As observações utilizadas destas duas variáveis descritas vão assim entre os trimestres 2000Q4 – 2020Q2 (79 observações), pelo que será também o período que será utilizado para a série *PIB*.

Uma vez que o e-fatura é uma nova base de dados, esta não apresenta o mesmo número de observações das restantes. Os dados do *Fatura* existem desde Janeiro de 2016 a Junho de 2020, em frequência mensal. Devido à transformação para obter a mesma frequência do *PIB*, fica-se com 18 observações no total para a série temporal (2016Q1 – 2020Q2).

O método proposto neste estudo passa por utilizar os dados provenientes de todas as variáveis referidas como *proxys* do *PIB* para prever a tendência do mesmo. O *Fatura* trata-se duma fonte especial neste estudo uma vez que é pouco foi utilizada em estudos científicos (de conhecimento público) até à data, embora os dados agregados estejam disponíveis no site do e-Fatura[5].

3.2.1 Considerações sobre os dados

Apesar dos benefícios apontados na utilização de novas fontes de informação, é relevante fazer uma reflexão sobre os cuidados a ter no processo.

Um dos problemas que deve ser considerado é a data inicial de registos na base de dados pois, tratando-se de um sistema recente (como o caso do e-fatura, criado em 2013 e com dados acessíveis a partir de 2016), o número de observações temporais é reduzido quando agregado em frequência mensal ou trimestral.

A baixa amostragem temporal dificulta a comparação e identificação de padrões com a variável macroeconómica em estudo, prejudicando a avaliação dos resultados obtidos (Buono et al., 2017).

Outro dos pontos sensíveis comumente referidos tem a ver com a presença de irregularidades e de padrões sazonais nos dados. Apesar de este fenómeno também estar

[4] https://www.multibanco.pt/estatisticas/
[5] https://faturas.portaldasfinancas.gov.pt/home.action

presente nas estimativas tradicionais, o seu efeito é mais sentido em amostras de alta frequência (Buono et al., 2017).

3.3 Metodologia

Nesta secção, serão apresentados três modelos utilizados nas estimações do capítulo 4 deste trabalho. Também é feita uma introdução a séries temporais, assim como as análises e cuidados que se deve ter no tratamento e na estimação dos parâmetros dos respetivos modelos.

Por último, serão apresentados critérios de seleção de modelos, que devem ser utilizados na escolha dos modelos que melhor se ajustam aos dados.

3.3.1 Séries Temporais

Uma série temporal consiste numa sequência de observações de uma variável, registadas ao longo de um período de tempo, em intervalos regulares (Box, Jenkins, Reinsel, & Ljung, 2015). Ao contrário dos dados transversais, em que a ordem das observações é irrelevante, em séries temporais a ordenação dos dados é fundamental (Wooldridge, 2012).

De acordo com Box et al. (2015), uma das características que as séries temporais apresentam normalmente, inerente ao processo de recolha dos dados, é a dependência entre as observações vizinhas (autocorrelação). Esta dependência deve ser considerada e incluída no processo de modelação, no qual as séries são vistas como realizações parciais de modelos processos estocásticos em tempo discreto.

Um processo estocástico (ou aleatório) é uma família de variáveis aleatórias indexadas no tempo (Wooldridge, 2012). Séries temporais multivariadas são realizações de processos estocásticos multivariados, ou seja, famílias de vetores aleatórios indexados no tempo (discreto).

Num processo de modelação de séries temporais, em que a relação entre várias séries pretende ser avaliada e incluída num modelo, deve ter-se em atenção que, para muitos problemas de diversas áreas, como frequentemente em economia, dados de séries temporais encontram-se disponíveis para várias variáveis de interesse, sendo a análise destes dados, regra geral, mais informativa e efetiva quando as séries temporais são consideradas como componentes multivariadas e analisadas todas em conjunto (Box et al., 2015).

Segundo Chatfield (2000) enquanto que os métodos univariados permitem fazer previsões dependendo apenas do valor presente e passado da série em análise (possivelmente aumentado por uma função do tempo, como a tendência), os métodos multivariados, por outro lado, permitem fazer previsões de uma variável dependente com base em valores de uma ou mais variáveis adicionais, chamadas de variáveis explicativas.

Ora, com vista à previsão de uma variável em função de outras, podem surgir problemas se se utilizarem métodos para dados seccionais, não incluindo a componente temporal.

De facto, usando um modelo de regressão clássico, em que, supondo dados de séries temporais $x_t \ e \ y_t$, em que y_t é a variável dependente, determinada por x_t, ambas datadas contemporaneamente, a relação entre as duas séries é dada pela equação (3.1):

$$y_t = \beta_0 + \beta_1 x_t + u_t, \ t = 1, 2, \dots, n \tag{3.1}$$

Onde t e indica o período à qual a observação se refere e u_t representa a componente estocástica ou aleatória do modelo de regressão (Wooldridge, 2012).

Neste modelo, o problema que eventualmente surge, e que inviabiliza a sua utilização, deve-se à componente estocástica u_t não se comportar de acordo com um ruído branco Gaussiano, uma vez que as observações das séries envolvidas podem estar correlacionadas.

As previsões multivariadas podem envolver mais do que uma equação se as variáveis forem dependentes entre si (o que acontece com os modelos VAR).

Box et al. (2015) identificam como duas das maiores vantagens da análise multivariada das séries temporais a compreensão das relações dinâmicas entre as séries ao longo do tempo e a melhoria na precisão das previsões de uma série individual usando informação disponível adicional de outras séries relacionadas.

3.3.2 Decomposição de séries temporais

Antes de se proceder à modelação e previsão de séries temporais, é importante fazer uma análise inicial dos dados, de forma a perceber as suas principais características e comportamento, identificando possíveis transformações necessárias para que seja possível estimar os parâmetros dos modelos considerados. Para isso, a análise gráfica das séries temporais é uma ferramenta essencial. Esta análise pode passar pela decomposição aditiva

das séries temporais (ou multiplicativa se as séries apresentarem variabilidade crescente com o tempo), de forma a detetar padrões de comportamento dos dados. Os métodos de decomposição identificam esses padrões individualmente e podem ser extraídos da série para uma melhor compreensão ou apenas para melhorar as previsões Chatfield (2000).

Existem quatro componentes essenciais a considerar (Chatfield & Xing, 2019):

- Variação sazonal: fenómeno de periodicidade geralmente anual, que pode ser encontrado em séries de qualquer frequência, sempre que padrões de comportamento similares são influenciados pelos fatores sazonais (por exemplo trimestre do ano, o mês, o dia). A sazonalidade é sempre de um período fixo e conhecido (Hyndman & Athanasopoulos, 2018). Caso não seja de interesse para a estimação, esta componente é removida dos dados;
- Variação cíclica: para além das variações sazonais, variações cíclicas regulares podem existir para períodos superiores a um ano com origem em fenómenos diferentes da sazonalidade;
- Tendência: esta variação é apresentada quando a série exibe um comportamento de crescimento positivo ou negativo persistente (pelo menos num período considerável de observações sucessivas), ou seja, apresenta uma alteração nos valores observados a longo prazo.
- Flutuações irregulares: variações que permanecem depois da tendência, sazonalidade e outros efeitos sistemáticos serem removidos, originando uma série de resíduos que podem ou não ser aleatórios. As flutuações aleatórias irregulares podem ser explicadas com base em modelos de probabilidades como as médias móveis (MA) ou autorregressivos (AR).

3.3.3 Estacionariedade das séries temporais

A estacionariedade é uma suposição básica importante na modelação de séries temporais. De facto, não é possível estimar parâmetros com o valor médio, variância ou covariância em conjuntos de dados temporais, se não for verificada a estacionariedade das observações.

Em termos simples, um processo estocástico (ou série temporal) é estacionário se as leis de probabilidade que atuam no processo não mudam com o tempo, isto é, o processo mantém o equilíbrio estatístico. (Chatfield & Xing, 2019).

Formalmente, uma série é diz-se estritamente estacionária se a distribuição conjunta de qualquer vetor n-variado $X(t_1), \dots, X(t_n)$ é igual à distribuição conjunta do vetor sujeito a uma translação no tempo, $X(t_1 + \tau), \dots, X(t_n + \tau)$ para todo $t_1, \dots, t_n, \tau$. Por outras palavras, mudando o tempo de origem pelo montante τ não se observa um efeito nas distribuições conjuntas, as quais devem depender apenas dos intervalos entre $t_1, \dots, t_n$. Esta propriedade deve ser verificável para qualquer valor de n.

Usualmente considera-se apenas estacionariedade de segunda ordem ou para a covariância (estacionariedade fraca) da série. Este tipo de estacionariedade pressupõe a existência do segundo momento finito.

Numa série fracamente estacionária, a função valor médio é constante e a covariância entre duas variáveis $X(t)\ e\ X(t_n + \tau)$ depende apenas de τ. Como as variáveis $X(t)\ e\ X(t_n + \tau)$ são observações da mesma série temporal, fala-se em autocovariância entre as observações. Neste trabalho, considera-se sempre estacionariedade para a covariância.

Apesar de a análise gráfica dar uma ideia do comportamento dos dados, devem-se efetuar testes estatísticos de forma a avaliar a estacionariedade das séries. A maioria deste testes designam-se por testes de raiz unitária (por se basearem nas propriedades das raízes dos polinómios autorregressivos), sendo um dos mais utilizados o teste de Dickey-Fuller Aumentado (ou ADF) proposto por Dickey e Fuller (1979) cujas hipóteses são:

H_0: Existe pelo menos uma raiz unitária - a série não é estacionária

H_1: Não existem raízes unitárias - a série é estacionária

Quando a série é não estacionária por exibir tendência estocástica, o recurso às diferenças dos valores da série, sucessivamente até que se consiga estacionariedade, é um método que integra os procedimentos popularizados por Box e Jenkins (1970) na análise de séries temporais, (Chatfield & Xing, 2019).

Considerando dados sem sazonalidade, quando o crescimento (ou decrescimento) é linear, as diferenças de primeira ordem são normalmente suficientes para obter a estacionariedade. Os valores $y_1, \dots, y_{n-1}$ são obtidos a partir da série original $x_1, \dots, x_n$ condorme (3.2):

$$y_t = x_{t+1} - x_t = \Delta x_{t+1} \tag{3.2}$$

No entanto, no caso da diferença de primeira ordem (3.2) não corrigir a não estacionariedade da série, procede-se a segundas diferenças, ou seja, utilizando operador Δ^2 tem-se (3.3):

$$\Delta^2 x_{t+2} = \Delta x_{t+2} - \Delta x_{t+1} = x_{t+2} - 2x_{t+1} + x_t \tag{3.3}$$

É importante perceber o conceito de operador de retrocesso tanto para fazer diferenciação das séries como para os modelos na próxima secção. Assim, o operador de retrocesso trata-se de uma notação para quando se está a tratar com desfasamentos nas séries temporais (Hyndman & Athanasopoulos, 2018). Considerando B como o operador, uma variável desfasada um período pode ser escrita como $By_t = y_{t-1}$. Quando é aplicado um desfasamento de dois períodos a notação indicada é obtida por $B(By_t) = B^2 y_t = y_{t-2}$.

3.3.4 Modelo ARIMA

Segundo Zhang (2003) um dos mais conhecidos e importantes modelos de séries temporais é o modelo autorregressivo integrado de medias moveis (ARIMA) popularizados por Box e Jenkins (1970). Estes modelos podem ser estendidos a dados com sazonalidade.

Os modelos ARIMA são considerados bastante flexíveis uma vez que representam vários tipos de modelos de séries temporais tais como regressão sobre o passado da série, autorregressivo (AR), ou de médias moveis (MA) ou ainda processos AR e MA combinados (ARMA). No entanto estes modelos apresentam a limitação de serem modelos lineares, ou seja, os dados atuais são descritos através de dependência linear com o seu próprio passado e com o passado de erros aleatórios (inovações) não sendo possível capturar qualquer padrão não linear (Zhang, 2003).

Antes da formulação do modelo ARIMA, torna-se assim necessário conhecer primeiro os modelos Autorregressivos e de Médias Móveis. Estes modelos assentam nos denominados processos de ruído branco (Box et al., 2015). Um processo de ruído branco é constituído por uma sucessão de variáveis aleatórias $e_t, e_{t-1}, e_{t-2}, \ldots$ identicamente distribuídas, de média μ e variância σ_e^2 constantes, e não correlacionadas, cujas realizações são designadas por choques aleatórios. Se a distribuição dos choques aleatórios é Normal, o

processo de ruído branco diz-se Gaussiano. Os processos de ruído branco são peças fundamentais na definição de modelos ARMA.

Modelo Autorregressivo (AR):

Ao contrário dos modelos de regressão múltipla, no modelo de auto regressão, a variável dependente é estimada e prevista com base numa combinação de valores anteriores dessa mesma variável. O termo auto regressão enfatiza o facto de se tratar de uma regressão de uma variável sobre si própria (Hyndman & Athanasopoulos, 2018) . Assim, um modelo autorregressivo de ordem p pode ser escrito como (3.4):

$$y_t = c + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + + \beta_p y_{t-p} + \varepsilon_t \quad (3.4)$$

Onde ε_t é ruído branco, c uma constante e β_t os coeficientes para os valores desfasados de y_t. Apesar de o modelo admitir múltiplos regressores estes são valores desfasados de y_t. Este modelo é referido como modelo autorregressivo de ordem p, ou AR(p).

Modelo de Médias Móveis (MA):

Ao contrário dos modelos autorregressivos que usam valores anteriores da variável estimada, os modelos de médias móveis usam uma média móvel dos choques aleatórios passados num modelo do tipo regressão (Hyndman & Athanasopoulos, 2018). O modelo é assim descrito como (3.5):

$$y_t = c + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \ldots + + \alpha_p \varepsilon_{t-q} \quad (3.5)$$

Sendo ε_t o ruído branco, c uma constante e α_t os coeficientes para os valores desfasados de ε_t.. Este modelo é referido como modelo de médias móveis de ordem q, ou MA (q). Como os valores de ε_t não são observados o modelo não é considerado uma regressão no sendo comum.

Modelo Autorregressivo de Médias Móveis:

Os modelos autorregressivos de médias móveis, modelos ARMA(p, q), combinam os processos AR(p) e MA(q) descritos, sendo que explicam os valores atuais de uma série temporal, y_t, em função dos próprios valores passados (equação 3.4) e de erros aleatórios

presentes e passados (equação 3.5). É admitido que os erros são não correlacionados entre si, mas podem estar correlacionados com a variável dependente (Box et al., 2015).

O modelo ARMA($p,\ q$) pode assim ser descrito em (3.6) como uma agregação das equações (3.4) e (3.5) tal que:

$$y_t = c + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_p \varepsilon_{t-q} + \varepsilon_t \quad (3.6)$$

Onde ε_t é um processo de ruído branco Gaussiano. Os parâmetros p e q são estimados com base nas observações da série.

O modelo autorregressivo de ordem p, AR(p), pode ser escrito usando o operador de retrocesso B na seguinte forma (3.7):

$$\left(1 - \beta_1 B - \beta_2 B^2 - \ldots - \beta_p B^p\right) Y_t = \varepsilon_t \quad (3.7)$$

Ou ainda (3.8):

$$\beta_p(B) Y_t = \varepsilon_t \quad (3.8)$$

Onde $\beta_1 B - \beta_2 B^2 - \ldots - \beta_p B^p$ é designado por polinómio autorregressivo de ordem p com ε_t um processo de ruído branco Gaussiano. A estacionariedade do processo exige que as p raízes de $\beta_p(B)$ seja, em módulo, superiores a 1.

A ACF (Função de AutoCorrelação) de um processo AR(p) é dada por (3.9):

$$\rho_k = \beta_1 \rho_{k-1} + \beta_2 \rho_{k-2} + \ldots + \beta_p \rho_{k-p}\ , k = 1, 2, \ldots \quad (3.9)$$

Consequentemente, a ACF tem um descaimento exponencial ou sinusoidal amortecido para zero, dependendo das raízes de $\beta_p(B)$.

Modelo Autorregressivo Integrado de Médias Móveis:

Como já foi referido, a maioria dos procedimentos de análise de séries temporais supõe que estas sejam estacionárias, pelo que será necessário transformar os dados originais se estes não formam uma série estacionária. A transformação mais comum consiste em

tomar diferenças sucessivas da séries original, até se obter uma série estacionária (Hyndman & Athanasopoulos, 2018).

Quando se combina os modelos autorregressivos e de médias móveis com o processo de diferenciação d (recurso a d diferenças sucessivas) explicado na secção 3.3.3 é obtido um modelo ARIMA em que I representa "integrado" no sentido de reverso de diferenciado discreto, ou seja, somas parciais, uma vez que se tratam de diferenças (Hyndman & Athanasopoulos, 2018).

O modelo completo pode assim ser descrito em (3.10) como uma agregação das equações (3.4) e (3.5) tal que:

$$y'_t = c + \beta_1 y'_{t-1} + \ldots + \beta_p y'_{t-p} + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_p \varepsilon_{t-q} + \varepsilon_t \qquad (3.10)$$

Em que y'_t é o valor diferenciado das séries (pode ser diferenciado mais do que uma vez). As variáveis independentes do modelo são assim os próprios valores desfasados de y_t e os erros também desfasados. O modelo é então designado por ARIMA (p, d, q) onde:

- p = ordem da parte autorregressiva;
- d = ordem das diferenças envolvidas;
- q = ordem da parte de médias móveis.

Todos os modelos (AR, MA, ARMA e ARIMA) seguem condições de estacionariedade e invertibilidade (Hyndman & Athanasopoulos, 2018).

Gráficos ACF e PACF:

Os termos ACF e PACF significam, respetivamente, Função de AutoCorrelação e Função de AutoCorrelação Parcial. Estes gráficos são uteis para prever os valores apropriados para os parâmetros p e q.

O gráfico ACF mostra a autocorrelação medida na relação entre y_t e y_{t-k} para diferentes valores de k. Se y_t e y_{t-1} estiverem correlacionados então y_{t-1} e y_{t-2} também estão correlacionados. No entanto, se y_t e y_{t-2} estiverem correlacionados, não é possível observar essa correlação uma vez que ambos estão relacionados com y_{t-1}. Assim, recorre-se paralelamente na análise ao gráfico PACF para medir a relação entre y_t e y_{t-k},uma vez que é removido o efeito dos desfasamentos de y_t com y_{t-k}, permitindo perceber se ambas as observações estão parcialmente correlacionadas (Hyndman & Athanasopoulos, 2018).

Os gráficos ACF e PACF tornam-se assim fundamentais não só na escolha dos valores dos parâmetros para um modelo ARIMA mas também na análise dos resíduos da estimação de um modelo, permitindo avaliar e comparar os ajustamentos de diferentes modelos (Box et al., 2015).

ARIMAX – modelo ARIMA com inclusão de variáveis exógenas:

Como referido por Kongcharoen e Kruangpradit (2013), quando o objetivo se trata de previsões de uma variável macroeconómica, a inclusão de indicadores relacionados com esta pode melhorar a capacidade preditiva do modelo. Assim, o modelo ARIMA simples estendido para um modelo ARIMA com variáveis explicativas (X) é chamado de modelo ARIMAX (p, d, q). Tendo em conta a equação (3.10), um modelo ARIMAX(p, d, q) com uma variável explicativa pode ser definido como (3.11):

$$y'_t = c + \gamma_1 x_t + \beta_1 y'_{t-1} + \ldots + \beta_p y'_{t-p} + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_p \varepsilon_{t-q} + \varepsilon_t \quad (3.11)$$

Em que x_t representa a variável exógena do modelo enquanto que γ_1 representa o respetivo coeficiente.

3.3.5 Modelo de Vetor Autorregressivo (VAR)

Uma das limitações da maior parte dos modelos de regressão temporal tem a ver com o facto de estes considerarem uma relação unidirecional entre a variável dependente e as variáveis independentes. Desta forma, apenas a variável dependente é influenciada pelas variáveis explicativas, e não vice-versa, o que pode não ser verdade com variáveis macroeconómicas (Hyndman & Athanasopoulos, 2018).

De facto, é comum existir uma relação entre as variáveis, onde todas têm poder explicativo, o que pode afetar a previsão das restantes, permitindo-se explicar entre si. Esta situação é frequente quando se lida com variáveis económicas, uma vez que o valor de uma variável, pode não estar apenas relacionado com o seu passado, como também depender do passado de outras variáveis (Lütkepohl, 2005).

Ora, esta relação bidirecional, deve ser incluída na estimação dos parâmetros dos modelos, melhorando a qualidade das previsões.

O modelo de vetor autorregressivo, VAR, surge assim no sentido de explorar as relações identificadas entre as variáveis do conjunto das séries temporais em que todas as variáveis são tratadas de forma simétrica. Assim, todas as variáveis são modeladas com a premissa de que se afetam entre si de forma igual, ou seja, todas as variáveis são endógenas (Zivot E. & Wang J., 2003).

Consideremos, como, exemplo um modelo VAR com duas variáveis, sendo elas $y_{1,t}$ e $y_{2,t}$, e um desfasamento 1 (desfasagem temporal $t-1$). O modelo VAR é constituído por uma equação por variável, em que no lado direito das mesmas se pode incluir uma constante (c) e as desfasagens de todas as variáveis utilizadas. O modelo para duas séries (modelo bidimensional) é escrito como (3.12 e 3.13):

$$y_{1,t} = c_1 + \beta_{11} y_{1,t-1} + \beta_{12} y_{2,t-1} + \varepsilon_{1,t} \quad (3.12)$$

$$y_{2,t} = c_2 + \beta_{21} y_{2,t-1} + \beta_{22} y_{2,t-1} + \varepsilon_{2,t} \quad (3.13)$$

Onde $\varepsilon_{1,t}$ e $\varepsilon_{2,t}$ são variáveis de ruído branco que podem estar contemporaneamente correlacionados.

Em geral, o coeficiente $\beta_{ii,l}$ captura a influência da ordem de l desfasamentos da variável y_i em si própria enquanto que $\beta_{ij,l}$ captura a influência da ordem de l desfasamentos da variável y_j em y_i.

Os modelos são estimados equação por equação usando o método dos mínimos quadrados, isto é, para cada equação os parâmetros são estimados minimizando a soma dos quadrados dos valores de $\varepsilon_{i,t}$.

No momento de estimação com o modelo VAR há duas decisões que devem ser consideradas, nomeadamente o número de variáveis (K) e quantos desfasamentos (p) devem ser incluídos. Assim, o número de coeficientes a ser estimados pelo modelo VAR é igual a $K + pK^2$. Por exemplo, para um modelo com $K = 3$ e $p = 2$, serão estimados 21 coeficientes no total ou 7 coeficientes por equação. Uma vez que um grande número de coeficientes pode afetar a qualidade das previsões, devem ser apenas incluídas variáveis com forte correlação entre elas. No que toca ao valor de p, critérios de informação são normalmente usados para selecionar o número de desfasamentos ótimos a incluir.

Os modelos VAR estão incluídos no `package vars` no R. Este `package` contém a função `Varselect()` para a seleção do número de desfasamentos ótimos p usando

quatro critérios de informação diferentes: AIC, HQ, SC (também chamado de BIC) e FPE (Zivot E. & Wang J., 2003).

Teste de Causalidade de Granger:

O teste proposto por Granger em 1969 tem por base superar as limitações do uso de correlações entre as variáveis, uma vez que a correlação por si só não implica uma relação de causa-efeito entre elas (Granger, 1969).

Desta forma, uma variável X causa Y no sentido de Granger se a observação de X ajuda a prever os valores futuros de Y. Considerando o modelo VAR bidimensional, o conceito de causalidade de Granger pode ser caraterizado sob a hipótese de que os valores presentes e passados das variáveis endógenas X e Y do modelo contenham toda a informação necessária para a previsão dos seus valores futuros, podendo-se mostrar que a variável X causa Y no sentido de Granger se na equação de Y algum dos coeficientes $\beta_{i,l}$ é diferente de zero. Em alternativa, X não causa Y se todos os coeficientes são nulos, tal que:

$$H_0: \beta_{i,l} = 0$$

3.3.6 Auto Regressão de Rede Neuronal (NNAR)

Segundo Hill, O'Connor, e Remus (1996), os modelos de rede neuronal artificial são modelos matemáticos inspirados na funcionalidade de neurónios biológicos. Apesar das suas inúmeras variações, o princípio base destes modelos é permitir relações complexas não lineares entre a variável dependente e as variáveis explicativas.

Segundo Abounoori, Mohammadali, Gandali Alikhani e Naderi (2012), os diferentes modelos de rede neuronal artificial procuram simular a mente humana através de processos de aprendizagem baseados em métodos computacionais com o objetivo de resolver problemas grandes e complexos.

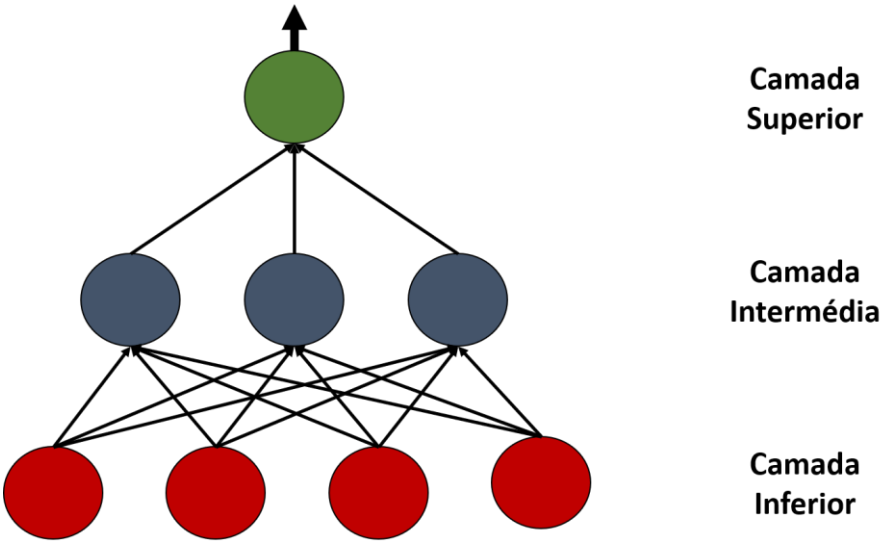
Uma rede neuronal pode ser entendida como uma rede de "neurónios" organizados em camadas de nós (Zhang, Patuwo, & Hu, 1998) onde os preditores (dados das séries) formam a camada inferior da rede e as camadas superiores (resultados) formam a camada superior onde a solução para o problema é obtida. Estas camadas podem estar separadas por uma camada intermediária no processo contendo "neurónios ocultos". Um exemplo visual simplificado desta estimação é dado na Figura 4, onde quatro nós na camada inferior são

representadas tendo como resultado um nó na camada superior de resultados. No meio destas camadas existe ainda a camada intermédia com quatro nós.

No entanto, na variação mais simples podem não existir camadas intermédias sendo que o modelo é equivalente a um modelo de regressão linear, sendo que nesse caso os valores estimados são obtidos diretamente pela combinação linear das variáveis independentes (Abounoori et al., 2012).

A representação na Figura 4 é conhecida como uma rede multicamada *feed-foward* em que em camada os nós são "alimentados" pela informação das camadas anteriores (os resultados dos nós da camada anterior são utilizados pelos nós da camada seguinte) (Zhang et al., 1998). Os dados que chegam a cada nó são obtidos através de uma combinação linear ponderada. O resultado é de seguida calculado por uma função não linear.

Figura 4 - Rede Neuronal com quatro nós inferiores, um nó na camada superior e quatro nós na camada intermédia



Fonte: Autoria própria, adaptado de "*Forecasting with artificial neural network*" de Zhang et al. (1998)

Segundo Abounoori et al. (2012) um dos modelos mais usados entre os modelos de rede neural dinâmicos são os modelos NNAR (Auto Regressão de Rede Neuronal). Os modelos NNAR são construídos adicionando um componente AR ao processo do modelo de rede neuronal, tendo assim o modelo uma componente linear e não linear (a parte não linear é estimada pelo método artificial de rede neural *feed-foward* com camadas ocultas enquanto que a componente linear é incluída pelo modelo autorregressivo). Na presença de variáveis explicativas exógenas o modelo toma a designação de NNARX, tal como descrito

por Abounoori et al. (2012) e Karbasi, Laskukalayeh e Fahimifard (2009). De forma geral, a estimação para o modelo NNARX é descrita por:

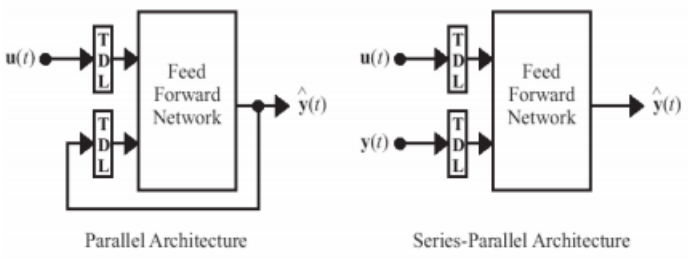
$$y_t = f(u_t, u_{t-1}, \dots, u_{t-n_u}, y_{t-1}, y_{t-2}, \dots, y_{t-n_y}) \tag{3.14}$$

Em que f representa o mapeamento da informação entre os nós nas diferentes camadas da rede neural, sendo que o valor para a variável dependente (resultado) dado por y_t é estimado com base nos seus valores passados (desfasados) e nos valores da variável independente (exógena) u_t. Segundo a arquitetura do modelo NNARX, o valor do resultado obtido para y_t pode retornado como variável explicativa na componente do processo *feed-foward* da rede neuronal, como ilustrado na Figura 5 (*Parallel Architecture)*. No entanto, na presença dos valores reais de y_t, estes podem ser imputados diretamente durante o treino da rede, utilizando como variável explicativa os valores originais em vez dos valores estimados (Figura 5 – *Series-Parallel Architecture*).

No que toca à abordagem *Series-Parallel Architecture* Karbasi et al. (2009) identificam como principais vantagens uma maior precisão do modelo estimado e o resultado de uma rede puramente *feed-foward* onde os dados passados podem ser usados para o treino.

Os modelos de auto regressão de rede neuronal utilizados neste trabalho estão implementados no `package forecast` no R, sendo `nnetar` a usada na estimação dos parâmetros do modelo NNAR (ou NNARX) (Hyndman & Athanasopoulos, 2018).

Figura 5 - Arquitetura Paralela e Paralela em série do modelos NNAR



Fonte: *"Comparison of NNARX, ANN and ARIMA Techniques to Poultry Retail Price Forecasting"* de Karbasi et al. (2009)

3.3.7 Seleção dos Modelos

Apesar de os testes de autocorrelação e autocorrelação parcial poderem ser bastante uteis na identificação dos melhores ajustamentos dos modelos, por vezes estes podem ser ambíguos e levar a erros na decisão(Box et al., 2015).

Uma das propostas para a seleção de modelos em Box et al. (2015) é o uso de critérios de informação tais como o AIC proposto por Akaike (3.15) (ou o critério AIC na versão corrigida AICc (3.16)) e o critério BIC (Bayesian Information Criteria) de Schwarz (3.17):

$$AIC = -2\ln(L) + 2k \tag{3.15}$$

$$AICc = AIC + \frac{2k(k-1)}{k-p-1} \tag{3.16}$$

$$BIC = -2\ln(L) + 2\ln(N)\,k \tag{3.17}$$

Onde L representa o valor que minimiza a função log-verosimilhança da amostra, n é o número de medições registadas e k o número de parâmetros estimados.

Uma das medidas utilizados para medir a qualidade da previsão através dos erros é o indicador RMSE (*Root Mean Square Error* ou Raiz Quadrada do Erro-médio) e pode ser calculado pela equação (3.18) (Hyndman & Athanasopoulos, 2018):

$$RMSE = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.18}$$

Em que $\hat{y}_i$ representa os valores previstos e $y_i$os valores atuais ou observados.

4. Análise e Discussão dos Resultados

Nesta secção procede-se à análise dos dados, e apresentam-se e discutem-se os principais resultados obtidos, usando a metodologia apresentada.

Inicialmente será efetuada uma breve descrição e análise univariada das séries temporais, que levam às primeiras conclusões a partir da observação gráfica e quantitativa das observações. Seguidamente, a fim de se estabelecerem relações entre os dados, serão efetuadas análises multivaridadas, a partir de vários modelos propostos para posterior discussão e comparação dos resultados.

4.1 Análise Exploratória dos Dados

Antes de proceder à modelação dos dados temporais, importa proceder ao seu processamento, efetuando correções e eventuais transformações, de modo a ser possível efetuar comparações e a que se verifiquem os pressupostos da metodologia utilizada.

4.1.1 Análise e Descrição dos Dados

Na Figura 6 (à esquerda) estão representadas graficamente todas as séries em estudo, em frequência trimestral. A observação gráfica dos dados temporais permite retirar algumas conclusões iniciais relativamente ao seu comportamento da sua dinâmica.

Sazonalidade:

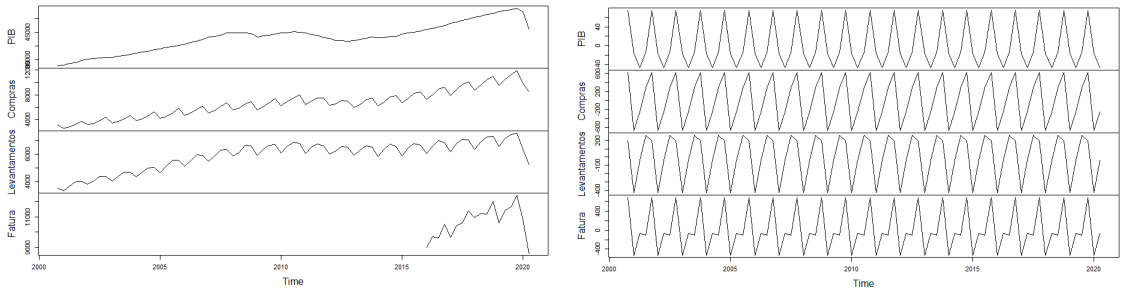
A série temporal *PIB* é publicada com ajustamento de sazonalidade e efeitos de calendário, pelo que se deve aplicar a mesma correção às restantes séries.

De facto, as séries *Compras, Levantamentos* e *Fatura* , apresentam sazonalidade anual regular, com comportamento semelhante nas três variáveis, sugerindo que os fenómenos que eventualmente afetam estas variáveis são comuns a todas, conforme se pode constatar graficamente a partir de decomposições aditivas das séries que evidenciam a componente sazonal (ver gráfico da direita na Figura 6).

A correção dos efeitos identificados nas séries *Compras, Levantamentos* e *Fatura* é assim efetuada, usando o procedimento para ajustamento sazonal e de efeitos de calendário

aplicado pelo Instituto Nacional de Estatística chamado X-13 ARIMA (Cabral & Duarte, 2016).

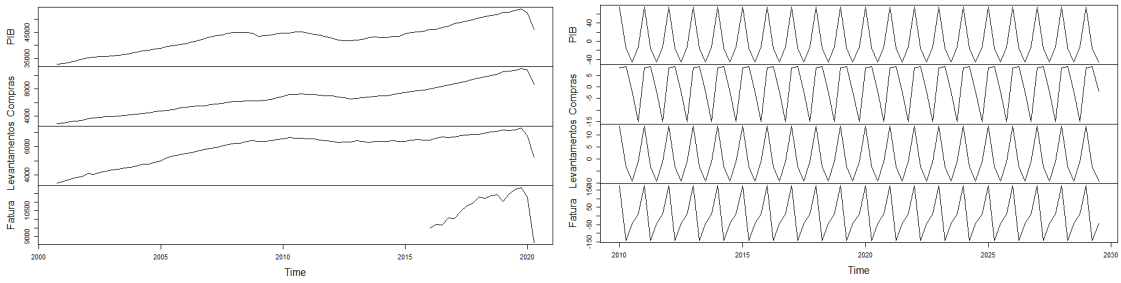
Figura 6 - Séries temporais individuais e a sua sazonalidade decomposta



Fonte: Autoria própria

Este ajustamento é possível recorrendo a uma ferramenta disponibilizada pelo Eurostat para uso público de nome JDemetra+ (Grudkowska, Buono, Palate, & Ciebiera, 2013). Este procedimento não é aplicado a *PIB* uma vez que é admitido que a série já foi submetida ao mesmo antes da sua publicação. Os resultados destas correções são visíveis na Figura 7.

Figura 7 - Séries temporais individuais corrigidas de sazonalidade e a sua decomposição



Fonte: Autoria própria

Nas Figura 7 pode-se verificar que o efeito de sazonalidade foi corrigido nas séries *Compras, Levantamentos* e *Fatura*, tendo diminuído drasticamente, permitindo agora uma análise e comparação das séries mais fiel e precisa em relação aos seus comportamentos.

Na tabela 1 estão representadas algumas estatísticas sumárias quantitativas referentes a cada uma das séries originais, após a correção da sazonalidade.

A variável *PIB* apresenta-se com média e mediana mais elevada (43 154 e 43 520 milhões de euros, respetivamente), sendo este quase quatro vezes superior do observado em Fatura (10 689 e 10 985 milhões de euros).

Tabela 1 - Sumário Estatístico das Séries Temporais Individuais

	Mínimo	Q1	Mediana	Média	Q3	Máximo
PIB	32 831	40 017	43 520	43 154	45 113	53 702
Compras	2 799	5 037	6 718	6 608	7 723	11 062
Levantamentos	3 415	5 296	6 367	5 911	6 565	7 303
Fatura	8 641	10 037	10 985	10 689	11 343	11 841

Fonte: Autoria própria

No que toca a *Compras* a grande amplitude entre os valores mínimo e máximo (2 799 e 11 062 milhões de euros, respetivamente) é compatível com o elevado crescimento observado ao longo do período observado, nomeadamente a partir de 2013 (Figura 8).

4.1.2 Análise Multivariada

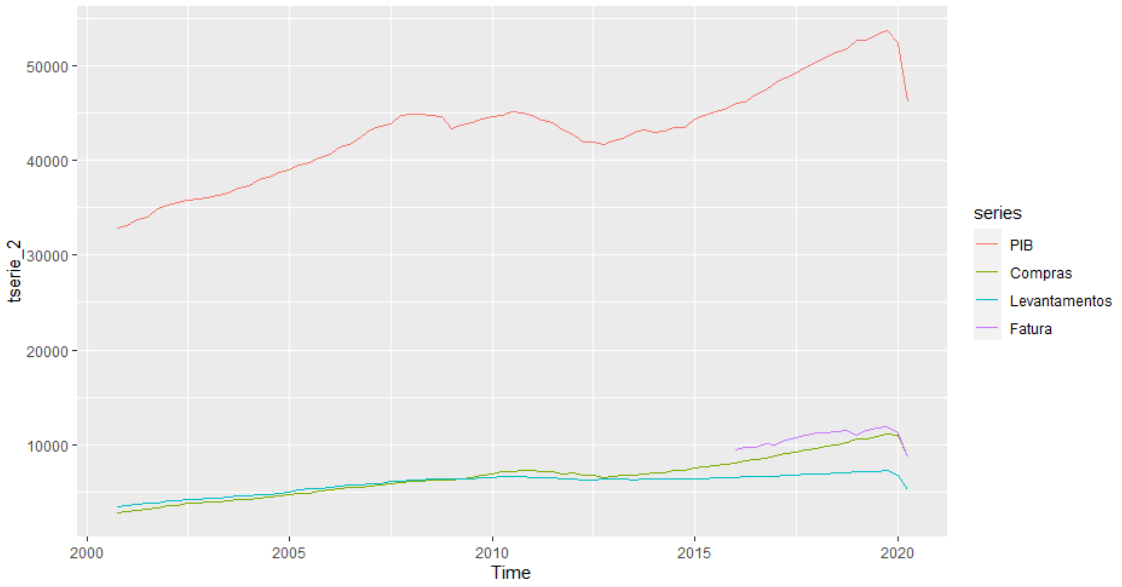
A crise financeira internacional de 2008, impôs imediatamente uma reação negativa no *PIB* (Figura 8), registando-se uma queda abrupta no seu crescimento, invertendo o crescimento acentuado que vinha sendo observado até à data. Assim, após atingir no segundo trimestre de 2008 o valor de 44 897.1 milhões de euros, é registado no primeiro trimestre de 2009 o valor de 43 346.9 milhões de euros.

A partir de final de 2010, percebe-se um decréscimo em todas as séries, mais acentuado em *PIB*, sendo que após no terceiro trimestre de 2010 atingir o valor de 45 117.6 milhões de euros, atinge no último trimestre de 2012 o valor mínimo para toda a série de 441 698.9 milhões de euros. Só após este período foi verificado um novo crescimento, com taxa mais ou menos constante, na série dos valores de *PIB*.

Este comportamento de decréscimo, desde final de 2010 até 2013 também se encontra bem patente em *Compras* que, como já foi referido anteriormente, a partir deste ano de 2013, tem um crescimento mais acentuado, e em linha com *PIB,* o que não é evidente em

Levantamentos, com taxa de crescimento inferior ao período pré-crise. A análise para *Fatura* não é possível por falta de observações neste período.

Figura 8 – Representação do Conjunto das Séries Temporais



Fonte: Autoria própria

A recessão observada no Produto Interno Bruto do país é retratada por Lourtie (2011) como consequência da crise de dívida soberana na União Europeia, afetando a estabilidade financeira da zona euro e por efeito de contágio atingiu gravemente a economia portuguesa.

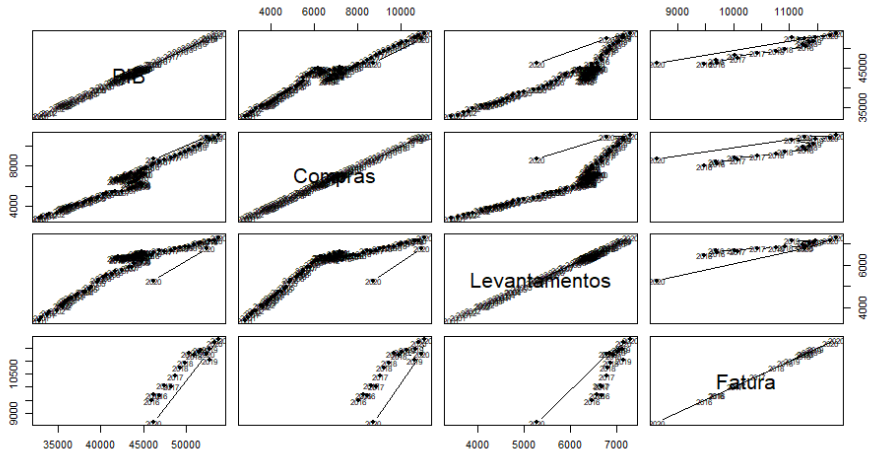
Mais recentemente, nomeadamente, nos dois trimestres registados para o ano de 2020, um recessão mais drástica foi sentida em *PIB* como consequência da pandemia do Coronavírus (Banco de Portugal, 2020). É visível na Figura 8 que a queda é mais acentuada e rápida do que a verificada durante o período da crise de 2010. Assim, o valor de *PIB* desceu de 53 702 para 46 138.4 milhões de euros do último trimestre de 2019 para o segundo trimestre de 2020, uma queda de 7 563.6 milhões de euros, o correspondente a 14.1% em apenas um semestre.

À semelhança de *PIB*, as séries *Compras, Levantamentos* e *Fatura* apresentam também uma queda violenta nos dois primeiros trimestres de 2020. Em termos comparativos o decréscimo de valor para o mesmo período para *Compras, Levantamentos* e *Fatura* foi de 21.3%, 27.8% e 27%, respetivamente.

De forma a melhor compreender a relação entre as variáveis em estudo, na Figura 9 estão representados graficamente os diagramas de dispersão, em pares, do conjunto de dados.

De acordo com estes diagramas, todas as séries apresentam relações fortes positivas com PIB. Note-se que a leitura destes gráficos deve ser cuidadosa uma vez que as séries não são estacionárias, com comportamento crescente ao longo do tempo, pelo que estas relações podem ser falaciosas nesta primeira análise.

Figura 9 - Gráficos de Dispersão em Pares



Fonte: Autoria própria

Na Figura 10 são apresentados gráficos comparativos entre a série *PIB* e cada uma das restantes séries temporais a fim de se identificarem os ritmos de crescimento dos valores em milhões de euros das séries. De notar que no gráfico com *PIB* e *Fatura*, apenas é visível a parte da série para as quais existem observações para *Fatura.*

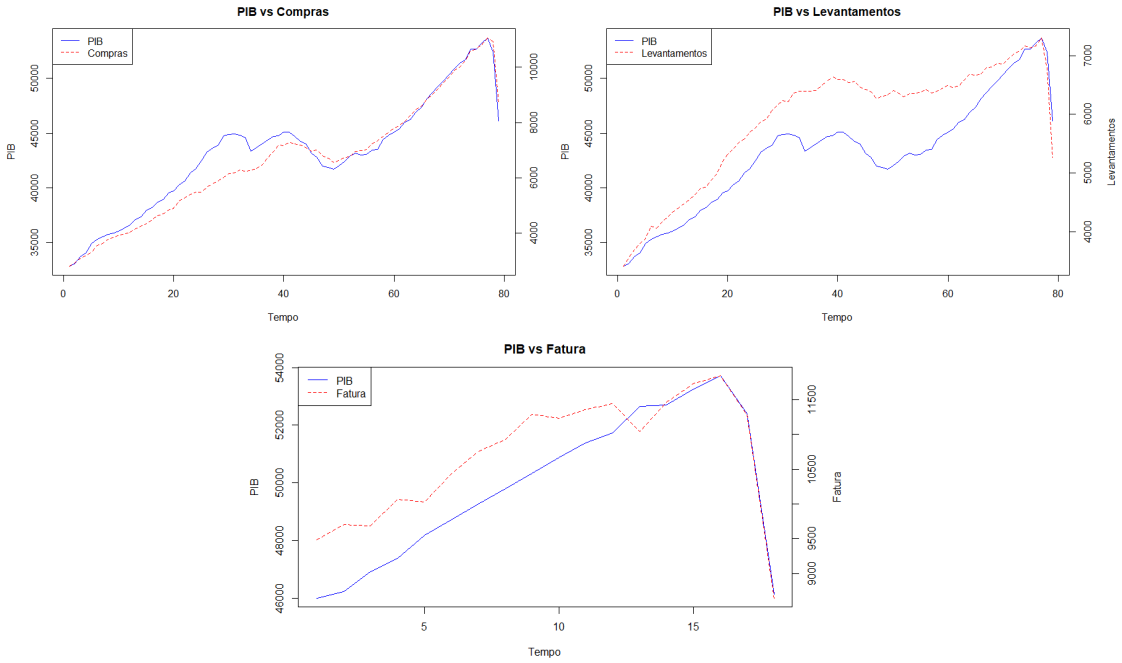
Relativamente a *PIB* e *Compras*, no período pré-crise financeira de 2008, os valores de *PIB* apresentam uma razão de crescimento superior à razão de crescimento de crescimento de *Compras.* No entanto, o comportamento de ambas as séries torna-se muito idêntico durante e após este marco, registando-se uma queda abrupta no período inicial com uma recuperação semelhante até ao final de 2019. A partir do primeiro trimestre de 2020 é de notar a inversão drástica no crescimento de ambas as séries, sendo esta mais acentuada para *PIB.*

No caso de *Levantamentos*, esta comparação apresenta conclusões inversas à anterior, apresentando até 2008 um crescimento semelhante a *PIB.* Também durante o período da crise de 2010 é visível uma quebra no crescimento em valor da variável (apesar de não tão acentuada como no caso de *Compras*) apesar de no período posterior e até final de 2019 o

crescimento ter abrandado. Também, a partir de 2020, a recessão visível é mais acentuada para *Levantamentos* do que *PIB,* inversamente à série *Compras.*

No que toca a *Fatura,* apesar de numa fase inicial ambas as séries crescerem de forma semelhante, existem observações com tendências contrárias em ambas. No entanto, as séries respondem de forma idêntica à quebra do PIB resultante da pandemia de Covid-19. De salientar que *Fatura* é uma fonte de dados recente, que se encontra em construção pelo que torna-se difícil captar a tendência correta da série.

Figura 10 - Gráficos Comparativos de *PIB* com *Compras*, *Levantamentos* e *Fatura*



Fonte: Autoria própria

Estacionariedade:

Um aspeto a ter em conta é a queda na última observação, que representa um *outlier* de todas as séries. Segundo o Boletim Estatístico de agosto publicado pelo Banco de Portugal (2020), este decréscimo severo é resultado da quebra da procura interna como consequência do confinamento devido à recente pandemia de Covid-19.

Outro ponto importante, e comum a todas as séries, é a ausência de estacionariedade, sendo esta essencial quando se trata de modelação de séries temporais. Para testar formalmente a hipótese de estacionariedade, procede-se ao teste de Dickey-Fuller, em que a hipótese nula é não a estacionariedade. O procedimento adotado passa por excluir o *outlier*

identificado pois, como é evidente, esta observação extrema tem um efeito negativo na sensibilidade do teste.

Os resultados do teste de Dickey-Fuller Aumentado apresentam os valores de *p-value* de 0.54, 0.60, 0.81 e 0.96 para as séries *PIB, Compras, Levantamentos* e *Fatura,* respetivamente, levando à não rejeição da hipótese de não estacionariedade das séries em análise.

O pressuposto de estacionariedade é essencial na modelação de séries temporais. Assim, torna-se necessário proceder a uma transformação dos dados (Dickey & Fuller, 1979). O cálculo de diferenças sucessivas da série original é a solução compatível com a metodologia de Box-Jenkins mais comum para o problema identificado.

4.2 Transformação dos Dados

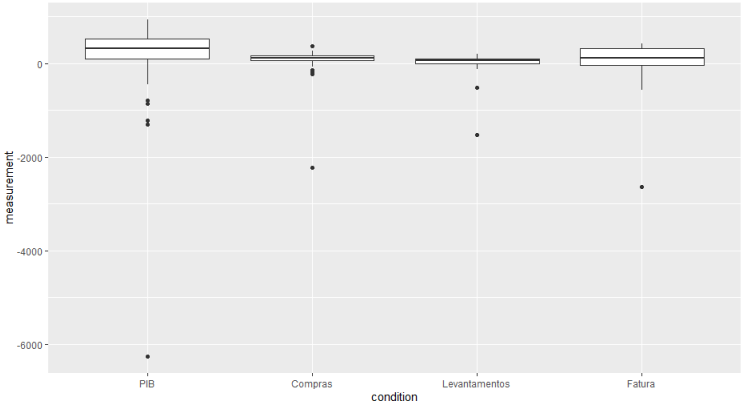
Com base na análise exploratória feita, a potencial relação forte entre as séries *PIB* com *Compras, Levantamentos* e *Fatura* deve ser explorada de forma a ser possível prever a tendência de crescimento das séries.

Para que esta metodologia seja viável, deverá ser feita uma transformação dos dados de modo a obter estacionariedade das séries temporais. Assim, começa-se por se aplicar uma diferença de primeira ordem.

Nos diagramas de caixas-de-bigodes da Figura 11 são visíveis alguns *outliers* para as variáveis após transformação, nomeadamente a última observação de todas as séries temporais (sendo estes *outliers* severos).

Como referido anteriormente, uma vez que estas observações extremas tem um efeito negativo na sensibilidade do teste Dickey-Fuller, estas serão excluídas do mesmo. No

Figura 11 - Diagramas de Caixa após a primeira diferença



Fonte: Autoria própria

entanto, o *p-value* para todas as séries diferenciadas apresenta-se ainda superior a 0.05, pelo que uma nova diferença é aplicada.

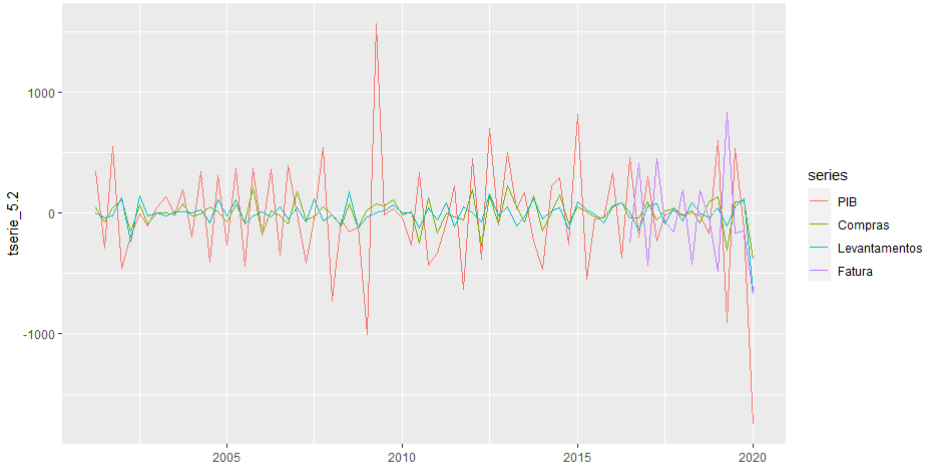
O teste de Dickey-Fuller para as novas séries transformadas indicam que todas apresentam um comportamento estacionário com um *p-value* inferior a 0.05.

Para uma melhor leitura, as variáveis após a transformação referida passam a ser designadas por:

- *dPIB* – Segunda diferença de *PIB;*
- *dCompras* - Segunda diferença de *Compras;*
- *dLevantamentos* - Segunda diferença de *Levantamentos;*
- *dFatura* – Segunda diferença de *Fatura.*

Na Figura 12 estão representadas todas as séries temporais após a referida transformação dos dados. A variável *dPIB* apresenta-se com maior amplitude de valores em relação às restantes séries. À exceção de *dFatura* que aparenta ter um comportamento inverso a *dPIB*, *dCompras* e *dLevantamentos* tem uma distribuição com uma tendência parecida a *dPIB* (apesar de com menor amplitude de valores).

Figura 12 – Representação das séries após transformação



Fonte: Autoria própria

Na Tabela 3 apresentam-se os coeficientes de correlação de Pearson das variáveis transformadas. As séries *dCompras* e *dLevantamentos* apresentam uma correlação positiva, no entanto os valores não são muito altos, como era de esperar após as diferenças de segunda ordem. No entanto é importante destacar a correlação negativa (e baixa) de *dFatura* para com *dPIB*.

Tabela 2 – Coeficientes de Correlação de Pearson

	dPIB	***dCompras***	***dLevantamentos***	***dFatura***
dPIB	1.00	0.50	0.43	-0.13*
dCompras	0.50	1.00	0.65	-0.26*
dLevantamentos	0.43	0.65	1.00	0.17*
dFatura	-0.13*	-0.26*	0.17*	1.00*

Fonte: Autoria própria
* correlação das séries entre 2016Q3-2020Q1

4.3 Identificação dos Modelos

Tal como mencionado em secções anteriores, a metodologia utilizada é baseada em modelos ARIMAX, VAR e NNARX.

O objetivo centra-se na obtenção de estimativas consistente com os valores de *PIB* ou *dPIB*, pelo que, num modelo de regressão, esta variável é dependente (endógena) sendo utilizadas como variáveis independentes (exógenas) *dCompras,* de *dLevantamentos* e *dFatura*.

Devido à disparidade do número de observações disponíveis entre as variáveis *Fatura* relativamente às restantes, são considerados dois grupos para cada modelo incluindo:

- Grupo 1 - *dCompras* + *dLevantamentos;*
- Grupo 2 - *dFatura* + *dCompras* + *dLevantamentos;*

Para facilidade de interpretação, na Tabela 3 são designados os modelos para as diferentes metodologias e os dois grupos.

Tabela 3 - Definição dos modelos estimados

	ARIMAX	VAR	NNARX
Grupo 1	**Modelo 1**	**Modelo 3**	**Modelo 5**
Grupo 2	**Modelo 2**	**Modelo 4**	**Modelo 6**

Fonte: Autoria própria

Uma vez que o número de observações é reduzido, a maior parte das observações será utilizada na estimação dos modelos. Assim, para o Grupo 1, das 77 observações restantes após transformação das séries, as primeiras 71 (2001Q2-2018Q4) serão alocadas ao conjunto

de treino dos modelos, enquanto as restantes 6 observações (2019Q1-2020Q2) serão usadas como teste dos mesmos.

No que toca ao Grupo 2, das 16 observações restantes, as primeiras 14 serão incluídas no conjunto de treino e as restantes 2 no conjunto de teste.

4.3.1 Modelo ARIMAX

Ajustamento do Modelo 1:

Os parâmetros do Modelo 1 são estimados usando como variável endógena d*PIB* e variáveis exógenas d*Compras* e d*Levantamentos*. O Modelo 1 obtido é um ARIMAX (0,0,1), isto é, os erros da regressão de d*PIB* sobre d*Compras* e d*Levantamentos* comportam-se como um MA(1).

O valor dos critérios de informação, são obtidos os valores de 1 016.72 para o critério AIC, 1 017.32 para AICc e 1 024.77 para BIC.

Na Figura 13, a vermelho estão indicados os valores ajustados pelo modelo, em comparação com os pontos a preto (valores originais). As linhas nas mesmas cores permitem perceber a tendência de crescimento com mais clareza de forma a facilitar a comparação (este esquema será utilizado para gráficos dos restantes modelos).

Pela análise visual da Figura 13 é percetível que o modelo ARIMAX (0,0,1) apresenta um ajustamento muito bom aos dados originais, uma vez que consegue captar todas as variações (quedas) de *dPIB* em quase todos os trimestres, à exceção ao período da crise financeira de 2008.

Análise dos Resíduos do Modelo 1:

Na Figura 14 são apresentados os gráficos de dispersão dos resíduos e os gráficos das funções de autocorrelação, ACF, e de autocorrelação parcial PACF do Modelo 1.

Os resíduos apresentam um comportamento adequado, dado que a média oscila aleatoriamente em torno de zero com variância constante, de acordo com o comportamento que se espera num processo de ruído branco.

Os gráficos das ACF e PACF evidenciam a não autocorrelação ou autocorrelação parcial dos resíduos, pelo que se assume que correlação entre as observações foi capturada pelo modelo de regressão ARIMAX(0, 0, 1).

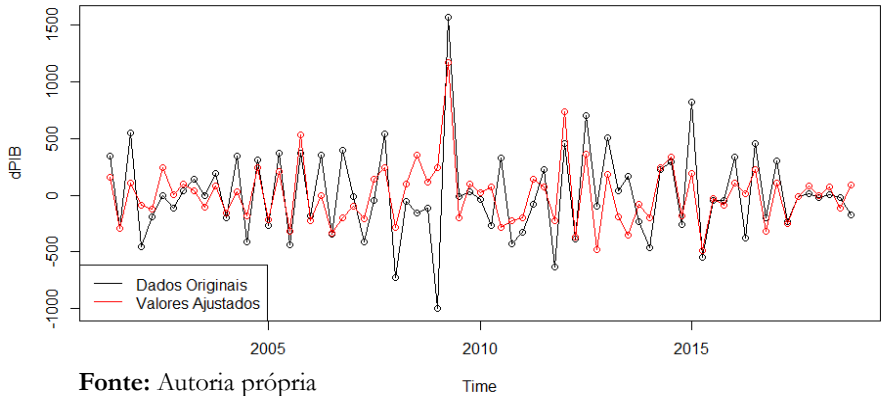
Para confirmar a análise gráfica é efetuado um teste de Box-LJung. O *p-value* do teste tem o valor de 0.1, o que leva à não rejeição da hipótese nula de que o modelo apresenta resíduos não correlacionados, conforme um processo de ruído branco.

Previsão do Modelo 1:

O resultado da previsão do Modelo 1 para o período 2019Q1 a 2020Q2 está representada no gráfico da Figura 15. A previsão parece ser bastante boa pois, para além de prever com exatidão as estimativas de *dPIB* para o período pré-Covid, também devolve valores da variável resposta quer para 2020Q1, quer para 2020Q2 com bastante exatidão. Assim, fazendo comparação dos valores, para 2020Q2 o valor real observado foi de -4 950.8 milhões de euros enquanto que o valor estimado para o mesmo período foi de -4 704.2 milhões de euros.

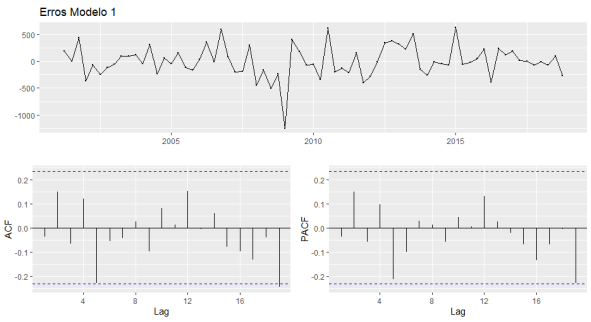
É importante ter em conta que o valor para 2020Q2 é considerado um *outlier* da série pelo que o modelo parece ser bastante robusto a prever em situações extremas.

Figura 13 – Ajustamento do Modelo 1



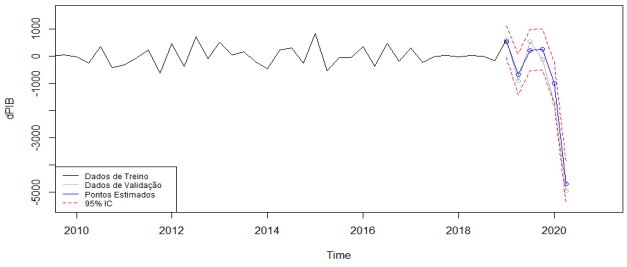
Fonte: Autoria própria

Figura 14 – Previsões do Modelo 1



Fonte: Autoria própria

Figura 15 – Resíduos do Modelo 1



Fonte: Autoria própria

Ajustamento do Modelo 2:

Os parâmetros do Modelo 2 são estimados usando novamente como variável endógena d*PIB*, sendo agora as variáveis exógenas *dCompras, dLevantamentos* e *dFatura*. O Modelo 2 é um ARIMAX (1,0,0) cujos valores dos critérios de informação bastante inferiores aos observados no Modelo 1. O valor obtido para AIC é 187.09, para AICc 194.59 e para BIC 190.28. Estes valores devem ser analisados com muita cautela, uma vez que houve uma redução drástica do número de dados usados na aprendizagem do modelo. Salienta-se que a estimação do Modelo 2 inclui apenas as observações desde 2016Q3 a 2019Q4.

Na Figura 16 está representada a série de valores originais em conjunto com os valores estimados pelo modelo, no mesmo esquema de cores do apresentado no Modelo 1.

O Modelo 2 apresenta um ajustamento bom para as observações disponíveis, sendo que apresenta até alguns valores previstos coincidentes com os reais, o que pode ter a ver com um sobreajustamento do modelo aos dados.

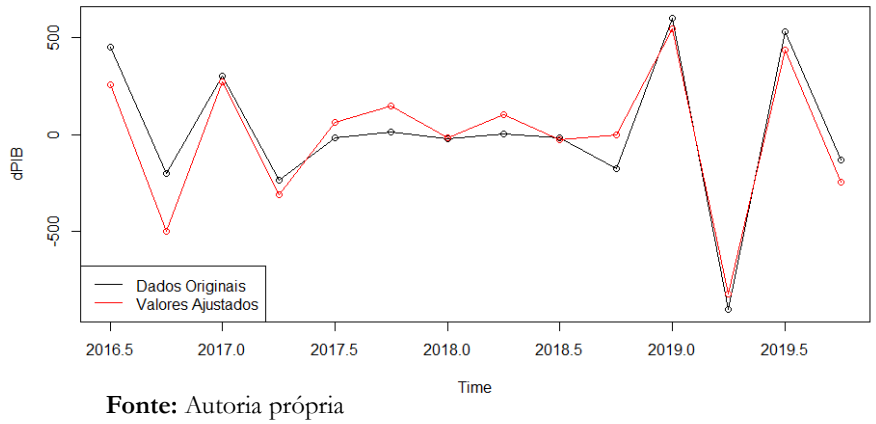
Análise dos Resíduos do Modelo 2:

Na Figura 17 são apresentados os gráficos de dispersão dos resíduos e os gráficos ACF e PACF do Modelo 2. O comportamento dos resíduos não parece ser compatível com um modelo com variância constante (existe uma grande variabilidade dos valores).

Os gráficos das funções ACF e PACF são compatíveis com a não existência de autocorrelação ou autocorrelação parcial entre os resíduos, mas o gráfico de dispersão dos resíduos reflete uma grande queda em 2008 e o teste Box-LJung para o Modelo 2 apresenta um *p-value* de 0.03, o que leva à rejeição da hipótese nula de que os resíduos do modelo se comportam de acordo com um processo de ruído branco.

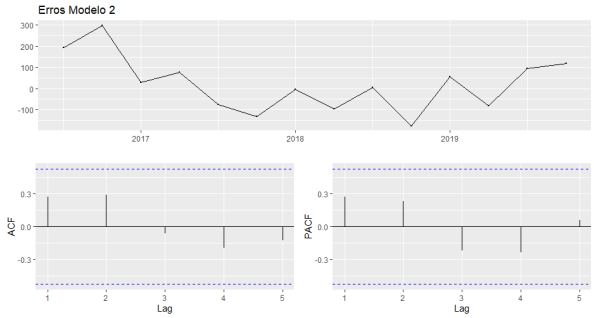
Desta forma, Modelo 3 não será considerado para previsão.

Figura 16 - Ajustamento do Modelo 2



Fonte: Autoria própria

Figura 17 - Resíduos do Modelo 2



Fonte: Autoria própria

4.3.2 Modelo VAR

Ajustamento do Modelo 3:

Na estimação dos parâmetros do Modelo 3 serão utilizados os valores das variáveis após transformação *dPIB, dCompras* e *dLevantamentos.*

A função `VARselect` retorna o valor de 4 desfasamentos como o valor ótimo para o Modelo 3 em função dos dados disponíveis pelos critérios de informação.

Testes de Causalidade de Granger são efetuados para as três variáveis. Os *p-values* para *dPIB, dCompras* e *dLevantamentos* são respetivamente 0.005, 0.060 e 0.226, pelo que se concluí que *dPIB* causa *dCompras* e *dLevantamentos*, *dCompras* causa dPIB e *dLevantamentos* para um nível de significância de 10% e *dLevantamentos* não causa dPIB e *dCompras*.

Assim, o modelo deve incluir *dPIB* e *dCompras,* podendo *dLevantamentos* não levar a uma melhor qualidade em termos de ajustamento (e de previsão). O modelo é estimado novamente excluindo *dLevantamentos* conforme as conclusões dos testes anteriores, sendo o ajustamento do modelo à variável *dPIB* é representado na Figura 18.

O Modelo 3, pela análise da Figura 18, permite captar a maior parte da tendência nas variações de *dPIB*. No entanto, percebe-se que o modelo não consegue reagir de forma adequada a eventuais choques, positivos ou negativos, tal como nos períodos de queda e recuperação de *PIB*, observadas na crise financeira de 2008.

Análise dos Resíduos do Modelo 3:

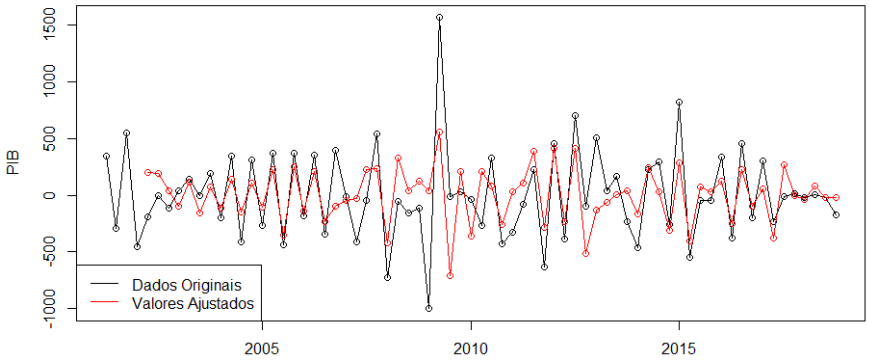
Pela análise dos gráficos dos resíduos (Figura 19) verifica-se que apresentam média em torno de zero e variância constante e os gráficos das funções ACF e PACF são compatíveis com ausência de autocorrelação e autocorrelação parcial. O teste de Box-LJung

apresenta um *p-value* de 0.25, concluindo-se que os resíduos são caraterizados por um processo de ruído branco.

Previsão do Modelo 3:

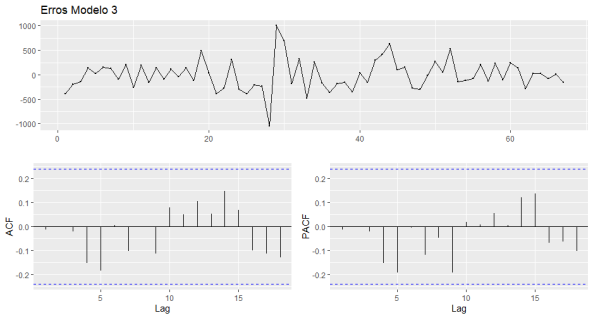
Na Figura 20 encontra-se a previsão do modelo. Verifica-se que a precisão das estimativas não é boa em nenhum trimestre comparativamente aos valores reais, mas principalmente nos dois últimos trimestres

Figura 18 – Ajustamento do Modelo 3



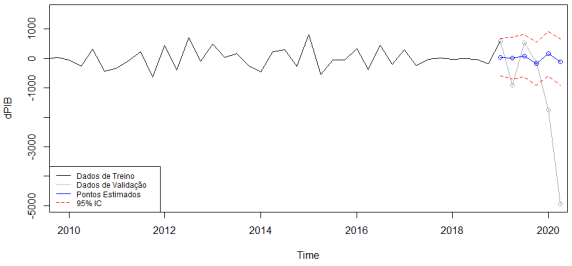
Fonte: Autoria própria

Figura 19 – Resíduos do Modelo 3



Fonte: Autoria própria

Figura 20 – Previsões do Modelo 3



Fonte: Autoria própria

Ajustamento do Modelo 4:

Na estimação do Modelo 1 são utilizados os valores das variáveis após transformação, nomeadamente, *dPIB, dCompras*, *dLevantamentos* e *dFatura.*

A função `VARselect` retorna o valor de 1 desfasamentos como o valor ótimo para o Modelo 4. Assim, o ajustamento do modelo é representado na Figura 21.

O ajustamento do Modelo 4 é bastante semelhante ao do Modelo 2, apresentando valores muito semelhantes e captando corretamente a tendência em todos os períodos.

Relativamente ao teste de Causalidade de Granger, os testes falharam para todas as variáveis, salientando que não existe uma relação de Granger causa no Modelo 4.

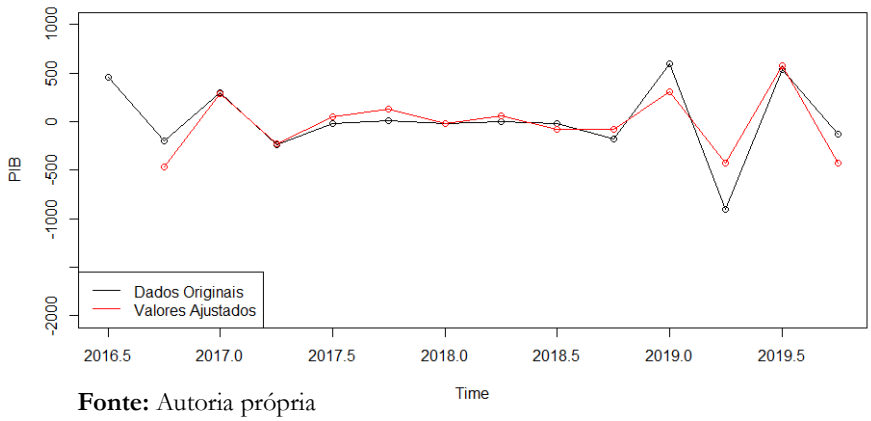
Análise dos Resíduos do Modelo 4:

Relativamente à análise de resíduos, na Figura 22 os resíduos apresentam-se como não correlacionados pelos gráficos ACF e PACF, o que é confirmado pelo teste de Box-LJung apresenta um *p-value* de 0.20.

Previsão do Modelo 4:

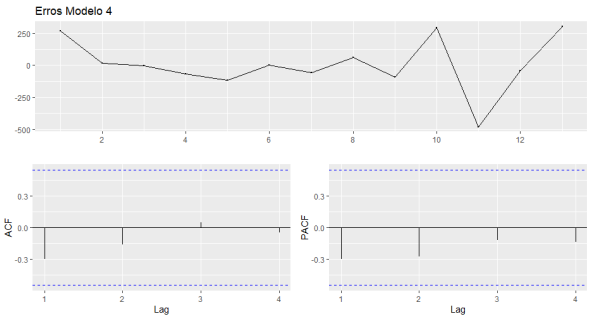
À semelhança do Modelo 3, o modelo VAR mostra-se novamente incapaz de prever corretamente a queda observada em *dPIB* no segundo trimestre de 2020 (Figura 23). Tais resultados eram previstos resultante da falta de correlação entre as variáveis identificada anteriormente.

Figura 21 – Ajustamento do Modelo 4



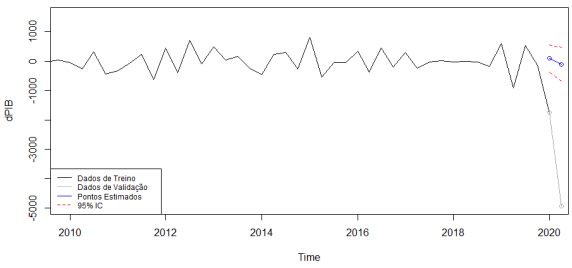
Fonte: Autoria própria

Figura 22 – Resíduos do Modelo 4



Fonte: Autoria própria

Figura 23 – Previsões do Modelo 4



Fonte: Autoria própria

4.3.1 Modelo NNARX

Ajustamento do Modelo 5:

Na estimação do Modelo 5 são utilizadas como variáveis independentes *dCompras* e *dLevantamentos*.

O Modelo 5 é um NNARX(1,2), contando com um desfasamento de 1 trimestre na variável *dPIB* e com 2 nódulos na camada intermédia, tendo a rede uma estrutura 3-2-1. A comparação dos valores ajustados e originais é representada na Figura 24.

Verifica-se que o ajustamento é bastante bom, conseguindo captar todas variações de *PIB* corretamente, sendo a única exceção o período de queda da crise de 2008.

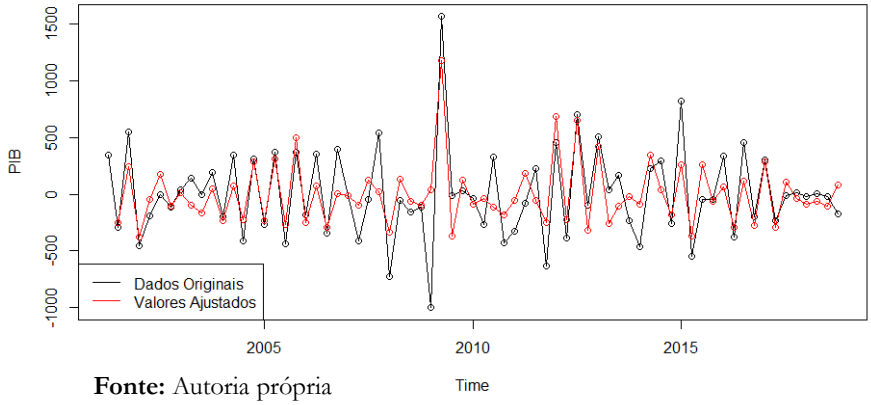
Análise dos Resíduos do Modelo 5

Relativamente à análise de resíduos do Modelo 5, o gráfico de dispersão correspondente apresenta um valor atípico (relativo ao ano de 2008), ver Figura 25, pelo que não é de estranhar que o teste de Box-LJung apresente um *p-value* de 0.045, próximo do patamar 0.05 normalmente considerado. De qualquer maneira há dúvidas que os resíduos da estimação do Modelo 5 se comportem de acordo com um processo de ruído branco.

Previsão do Modelo 5:

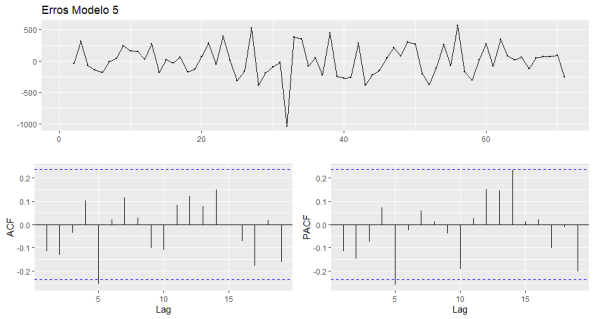
As previsões do Modelo 5 estão representadas na Figura 26. Apesar de o modelo estimar corretamente os valores para os trimestres anteriores a 2020, para o período 2020Q1-2020Q2 as estimativas são inconsistentes com os valores reais. Este modelo deve ser abandonado.

Figura 24 – Ajustamento do Modelo 5



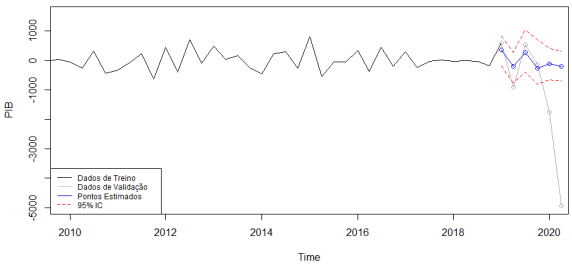
Fonte: Autoria própria

Figura 25 – Resíduos do Modelo 5



Fonte: Autoria própria

Figura 26 – Previsões do Modelo 5



Fonte: Autoria própria

Ajustamento do Modelo 6:

Na estimação do Modelo 6 são utilizadas como variáveis independentes *dCompras, dLevantamentos* e *dFatura.* O Modelo 6 é NNARX(1,2), contando também com um desfasamento de 1 trimestre na variável *PIB* e com 2 nódulos na camada intermédia, constituindo uma rede de estrutura 4-2-1.

Na Figura 27 está representado o ajustamento do Modelo 6 para o período 2016Q3 – 2019Q4. Este ajustamento é praticamente perfeito, sendo que o modelo foi capaz de prever tanto a tendência como os valores da variável dependente corretamente.

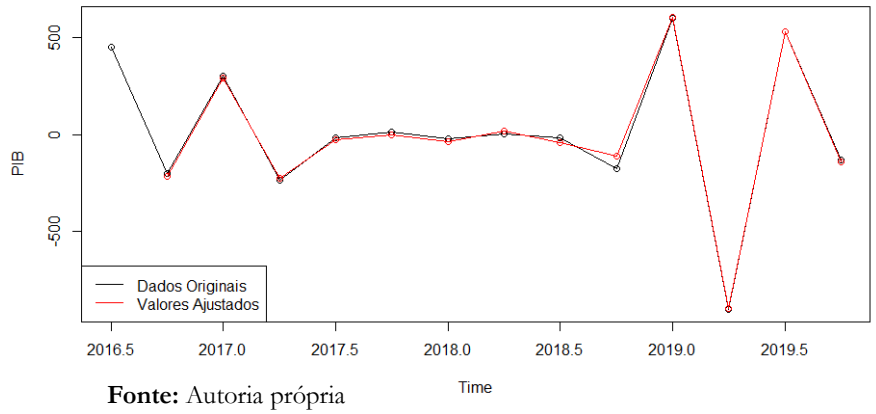
Análise dos Resíduos do Modelo 6:

No que toca aos resíduos do Modelo 6, o comportamento exibido das funções ACF (e PACF) na Figura 28 é compatível com ausência de correlação (e de correlação parcial) dos resíduos, ausência confirmada com o valor 0.2 do *p-value* do teste de Box-Ljung (apesar da existência de uma observação atípica relacionada com 2008).

Previsão do Modelo 6:

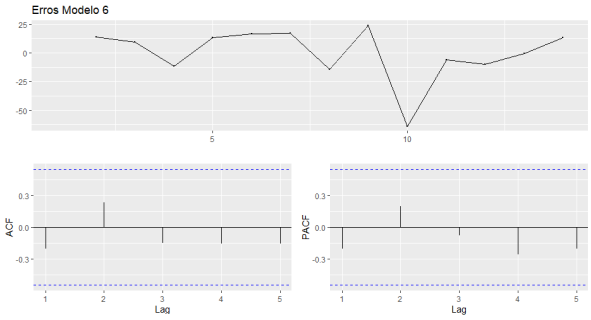
Por último, as previsões do Modelo 6 para os períodos 2020Q1 – 2020Q2 estão representadas na Figura 29. Apesar do bom ajustamento, as previsões apresentam um erro elevado (valor do RMSE na Tabela 4).

Figura 27 – Ajustamento do Modelo 6



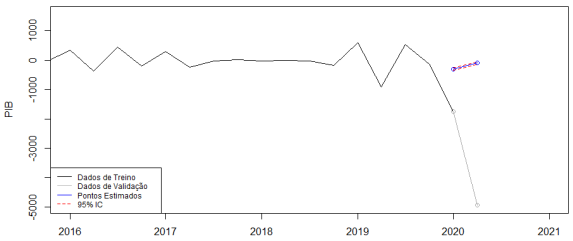
Fonte: Autoria própria

Figura 28 – Resíduos do Modelo 6



Fonte: Autoria própria

Figura 29 – Previsões do Modelo 6



Fonte: Autoria própria

4.4 Seleção do Modelo

Por último, importa fazer a seleção do modelo adequado e que se comporta melhor entre os seis modelos considerados anteriormente. Além das considerações que têm vindo a ser efetuadas, calcula-se a capacidade preditiva dos modelos a partir dos valores do erro RMSE para as previsões de todos os modelos. Os resultados (Tabela 4) mostram que o RMSE mais baixo é o apresentado pelo Modelo 1.

O resultado do critério RMSE é, portanto, consistente com as considerações que foram feitas na secção anterior.

Nesta secção vai-se ajustar um modelo ARIMAX aos dados, tendo como variável endógena *PIB* e variáveis exógenas *Compras* e *Levantamentos*.

No processo de estimação dos parâmetros do modelo serão utilizadas todas as observações disponíveis à exceção das do último trimestre, com as quais se testará a capacidade dos modelos.

Tabela 4 - Critério RMSE

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6
RMSE	392.53	N/A	2 179.07	3 664.79	2 080.19	3 589.052

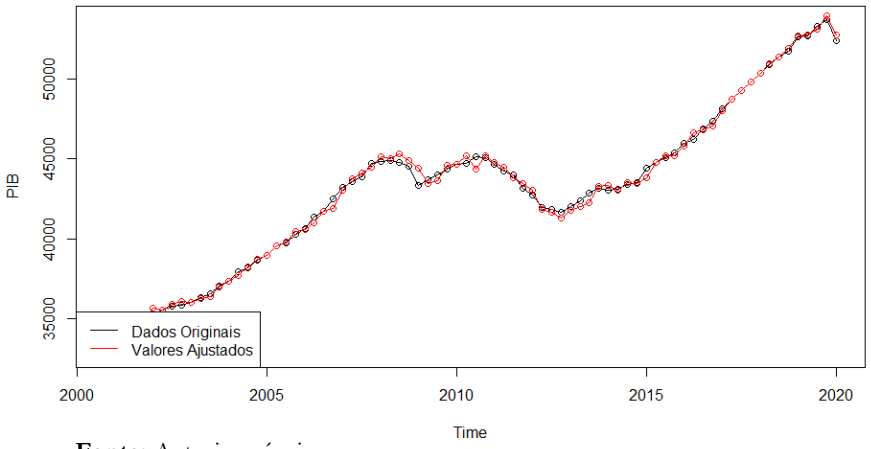
Fonte: Autoria própria

Quanto ao parâmetro d, o valor é 2, uma vez que só após duas diferenças é que as séries são estacionárias.

O modelo ARIMAX (1,2,2), escolhido para o ajustamento foi o que apresentou um menor valor do AICc. Encontra-se representado na Figura 30 o ajustamento do modelo aos dados do PIB.

O modelo ARIMAX (1,2,2) apresenta um ajustamento muito bom aos dados originais, sendo que consegue captar todas as variações (quedas) de *PIB* e inclusive apresenta alguns valores coincidentes com os valores originais.

Figura 30 - Ajustamento do Modelo Final



Fonte: Autoria própria

Para efeitos de comparação, na Tabela 5 estão os valores estimados em milhões de euros para *PIB* para os últimos 10 trimestres comparativamente aos valores originais.

Na última coluna da Tabela 5 são apresentados os resíduos da estimação. Estes podem ser encarados como a as diferenças entre os valores originais e estimados. A diferença de valores para maior parte dos trimestres é muito reduzida sendo que apenas se verifica diferenças mais substanciais (e negativas) nos períodos mais próximos da pandemia de Covid-19 (sendo esta considerada uma perturbação irregular).

Tabela 5 - Comparação dos valores originais e estimados para o Modelo Final

	Originais	Estimados	Resíduos
2017Q4	49 804.6	49 826.2	-21.7
2018Q1	50 333.6	50 331.0	2.6
2018Q2	50 867.1	50 941.3	-74.2
2018Q3	51 382.4	51 387.3	-4.9
2018Q4	51 721.7	51 927.0	-205.4
2019Q1	52 660.8	52 666.9	-6.1
2019Q2	52 695	52 737.4	-42.4
2019Q3	53 262.8	53 116.6	146.2
2019Q4	53 702	53 969.1	-267.1
2020Q1	52 395.6	52 770.0	-374.5

Fonte: Autoria própria

No gráfico da Figura 31, o último valor da série original corresponde a 2020Q1, sendo que a partir de 2020Q2 são visíveis os valores previstos. O valor para o trimestre 2020Q2 é estimado com base em dados em tempo real de *Compras* e *Levantamentos*. Os restantes períodos são previstos com base em observações passadas.

Assim, é notável que o valor para 2020Q2, apesar de ser um considerado um *outlier* como visto na secção 4.2, é previsto com bastante exatidão, sendo que é notória a queda de *PIB* na série para esse período.

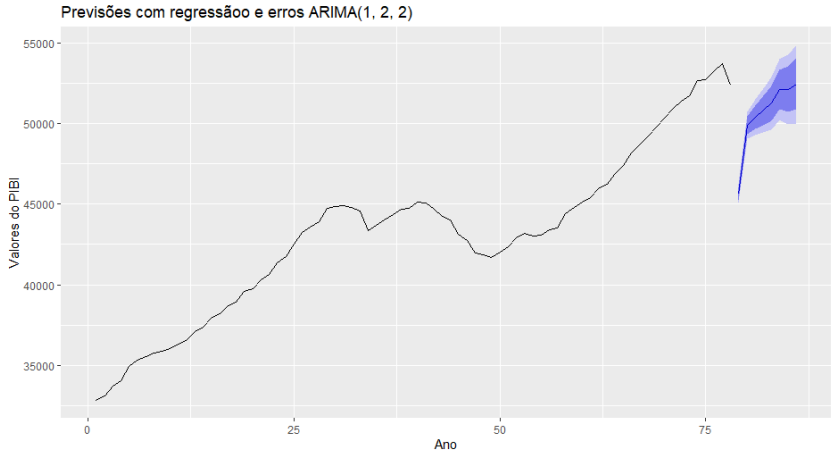
Desta forma o valor previsto para 2020Q2 pelo modelo para o PIB é de 45 644.7 milhões de euros. Uma vez que o valor original encontra-se em 46 138.4 milhões de euros, tal traduz-se numa diferença de apenas 493.7 milhões de euros.

Para os restantes períodos prevê-se uma recuperação económica, voltando ao ritmo de crescimento pré-Covid 19.

O modelo final é desta forma estimado por:

$$PIB = -0.9265ar1 - 0.0175ma1 - 0.7100ma2 + 1.9585Compras + 1.5202Levantamentos$$

Figura 31 - Previsão do Modelo Final



Fonte: Autoria própria

Conclusão

O presente trabalho dedica-se a estudar as necessidades e respetivas possibilidades da utilização de novas fontes de informação, nomeadamente, *Big Data,* como resposta à necessidade de estimativas rápidas de indicadores macroeconómicos, que cada vez mais tem preocupado as mais variadas instituições e entidades.

Para tal, tirou-se proveito de dados captados pela rede multibanco em Portugal, que, para além de serem dados acessíveis a qualquer interessado, tem um grande número de registos e informação diversificada em termos secionais. Estes dados são provenientes de dois meios de pagamento distintos, nomeadamente de compras com cartão em postos de venda e de levantamentos em terminais de pagamento automático.

Ainda, estudou-se a hipótese de incluir os dados de uma fonte de informação muito recente, o e-Fatura, na tentativa de melhorar a qualidade das estimativas produzidas com os dados de multibanco.

Assim, ambos os conjuntos de dados, representantes do Consumo Privados, foram encarados como indicadores ou *proxys* do Produto Interno Bruto português, com o objetivo de obter estimativas rápidas do seu valor com dados que estão disponíveis antes da sua divulgação oficial.

Para estudar esta relação, foram utilizadas duas metodologias sugeridas pelo Eurostat (2016b), indicadas para fazer previsão do PIB e dos restantes indicadores macroeconómicos, nomeadamente os modelos ARIMA(X) e os modelos VAR. Foi também estudada uma terceira metodologia, NNAR(X), com motivação em outros trabalhos realizados na área das estimativas rápidas (León & Ortega, 2018).

Relativamente aos dados utilizados, apesar dos dados de compras em PV ter um crescimento muito semelhante ao PIB, o mesmo não se pode dizer em relação a levantamentos e a e-Fatura. Se por um lado o crescimento do valor de levantamentos parece estar a estagnar, por outro os dados do e-Fatura mostram-se ruidosos e de difícil tratamento.

Desta forma, seis ajustamentos, divididos em dois grupos, foram estudados de forma a captar o máximo de observações de cada série. Os resultados, de forma geral, não favoreceram a utilização dos dados do Grupo 2, grupo modelado incluindo os dados do e-Fatura transformados. Apesar de os Modelos 2, 4 e 6 fazerem ajustamentos muito precisos com base nos dados do conjunto de treino, nenhum dos modelos conseguiu prever valores perto dos valores observados para o conjunto de teste, sendo que no caso do Modelo 6

(NNARX), a tendência prevista foi crescente contrariamente à recessão observada para o período. Os valores do RSME para os Modelos 4 e 6 são respetivamente 3 664.79 e 3 589.05, pelo que são considerados como muito elevados. Não foi obtido o valor do critério RSME para Modelo 2 uma vez que este foi rejeitado antes de se proceder à previsão devido aos resíduos da estimação não seguirem um processo de ruído branco.

Em relação ao Grupo 1, as conclusões são ligeiramente diferentes. Segundo o critério RMSE o pior ajustamento foi conseguido pelo modelo VAR (Modelo 3), sendo que para além do modelo não captar a tendência de recessiva da variável dependente nos primeiros trimestres de 2020 resultantes da pandemia de Covid-19 (à semelhança do Modelo 5), ao contrário da estimação feita pelo modelo NNARX, o ajustamento para o restante conjunto de teste também não é eficaz. Assim, conclui-se que os modelos VAR e NNARX não têm capacidade de prever corretamente as variações abruptas sofridas no PIB.

Por último, o Modelo 1, estimado com recurso à metodologia ARIMAX, apresenta resultados previstos muito precisos para o conjunto de teste, como se pode perceber para análise gráfica. Ainda, não só o modelo conseguiu prever a diminuição abrupta na variável dependente, como o valor previsto é muito próximo do valor real observado. Também, o critério RSME apresenta o valor mais baixo das previsões de todos os modelos, nomeadamente 392.53.

Desta forma, conclui-se que a melhor metodologia para o estudo em questão é a utilização do Modelo 1, nomeadamente o Modelo Autorregressivo Integrado de Médias Móveis com inclusão de variáveis explicativas exógenas, sendo elas os dados provenientes da rede multibanco (*Compras* e *Levantamentos*). Tendo isto em conta, foi estimado um modelo final incluindo todas as observações desde 2001Q2 a 2020Q1 de forma a fazer uma estimativa rápida para 2020Q2 com base nos valores observados para as variáveis exógenas *Compras* e *Levantamentos*.

Para além do modelo estimado apresentar muita precisão nos valores estimados em comparação aos valores reais (por exemplo, para o trimestre 2018Q1 a diferença foi de apenas 2.6 milhões de euros), o valor previsto para 2020Q2 pelo modelo trata-se também de um valor muito próximo do observado na série *PIB*. Ainda, deve-se considerar que o trimestre de 2020Q2 trata-se de um *outlier* severo das séries, o que comumente prejudica a eficácia dos modelos. Estes resultados permitem então confirmar que a metodologia proposta é robusta até na presença da estimação de um *outlier*, como é exemplo o trimestre 2020Q2.

Apesar de os dados do e-Fatura não permitirem concluir que a sua inclusão no processo de estimação rápida é relevante, pelo contrário, os dados de multibanco, e em especial compras em PV, sugerem que devem ser incluídos na estimação das estatísticas preliminares do PIB trimestral.

Não só este estudo motiva à utilização dos dados de multibanco no cálculo das estimativas rápidas do PIB como incentiva a que sejam procuradas e estudadas as relações entre outras fontes de *Big Data* com outros indicadores macroeconómicos relevantes de forma a responder ao crescente incentivo dos agentes económicos. Este exercício pode não ser só direcionado às entidades políticas e estatísticas, mas a qualquer agente económico como jornais ou investidores financeiros, uma vez que os dados da rede multibanco são públicos e publicados no *website* do Multibanco, normalmente, 6 dias após o fim do trimestre.

Num trabalho futuro, o presente estudo poderá ser desenvolvido por várias vias, tais como: novo estudo à série de e-Fatura com mais observações; fazer a estimação aproveitando a frequência mensal; comparação de previsões com ou sem inclusão de indicadores tradicionais; identificação de novas bases de dados que permitam a estimação rápida de outros indicadores macroeconómicos.

Bibliografia

Abounoori, A. A., Mohammadali, H., Gandali Alikhani, N., & Naderi, E. (2012). *Comparative study of static and dynamic neural network models for nonlinear time series forecasting*. (MPRA Paper, no. 46466). University Library of Munich. Germany.

Aprigliano, V., Ardizzi, G., & Monteforte, L. (2017). *Using the payment system data to forecast the Italian GDP*. (Temi di Discussione, no. 1098). Bank of Italy.

Banco de Portugal. (2020). Principais indicadores. *Boletim Estatístico do Banco de Portugal* (Agosto 2020), 19 - 22. Retirado de https://www.bportugal.pt/sites/default/files/anexos/beago20.pdf acedido em 15.09.2020.

Barcellan, R. (2010, Novembro). *Flash Estimates of GDP*. Documento apresentado na Third International Seminar on Early Warning and Business Cycle Indicators, Moscow, Russian Federation.

Bell, D. (1976). The Coming of the Post-Industrial Society. *The Educational Forum, 40*(4), 574-579. doi:10.1080/00131727609336501

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control* (5ª ed.). Hoboken, New Jersey: John Wiley & Sons.

Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*(1), 93-145.

Cabral, S., & Duarte, C. (2016). Previsão contemporânea das exportações portuguesas de turismo. *Revista de Estudos Económicos do Banco de Portugal, II(4),* 81 - 107.

Cardoso, F., & Rua, A. (2011). As contas nacionais trimestrais em tempo real: uma análise das revisões na última década. *Boletim Económico do Banco de Portugal* (Outono 2011), 147-164.

Carlsen, M., & Storgaard, P. E. (2010). *Dankort payments as a timely indicator of retail sales in Denmark*. (Working Papers, no. 66). Danmarks Nationalbank.

Castle, J. L., Fawcett, N. W., & Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review, 210*(1), 71-89. doi:10.1177/0027950109354412

Chamberlin, G. (2009). Flash estimates of European labour costs. *Economic & Labour Market Review, 3*(11), 32-39. doi:10.1057/elmr.2009.190

Charpin, F., Mathieu, C., & Mazzi, G. (2008, Setembro). *Construction of coincident indicators for the euro area.* Documento apresentado na 5th EUROSTAT Colloquium on Modern Tools For Business Cycle Analysis, Luxembourg.

Charpin, F., Mazzi, G. L., & Moauro, F. (2016). *Early estimates of GDP growth rates by error correction models*. (Statistical Working Papers). Eurostat. Luxembourg: Publications Office of the European Union.

Chatfield, C. (2000). *Time-series forecasting*: CRC Press.

Chatfield, C., & Xing, H. (2019). *The analysis of time series: an introduction with R* (7ª ed.): CRC press.

Christiano, L. J. (2012). Christopher A. Sims and vector autoregressions. *The Scandinavian Journal of Economics, 114*(4), 1082-1104. doi:10.1111/j.1467-9442.2012.01737.x

Dias, F., Pinheiro, M., & Rua, A. (2016). Previsão do PIB através de uma abordagem bottom-up num contexto rico em informação. *Revista de Estudos Económicos do Banco de Portugal, II(3),* 1-20.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association, 74*(366), 427-431.

Duarte, C., Rodrigues, P. M. M., & Rua, A. (2017). A mixed frequency approach to the forecasting of private consumption with ATM/POS data. *International Journal of Forecasting, 33*(1), 61-75. doi:10.1016/j.ijforecast.2016.08.003

Esteves, P. (2011). Are ATM/POS Data Relevant When Nowcasting Private Consumption? *The Empirical Economic Letters, 10*(9).

Esteves, P., & Rua, A. (2012). Previsões económicas de curto prazo para portugal: Uma síntese metodológica. *Boletim Económico do Banco de Portugal* (Outono 2012), 143 - 155.

Eurostat. (2012). Rapid estimates glossary. Retirado de https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Category:Rapid_estimates_glossary, acedido em 10.02.2020.

Eurostat. (2016a). *Euro area and European Union GDP flash estimates at 30 days*. (Statistical Working Papers). Luxembourg: Publications Office of the European Union.

Eurostat. (2016b). *Overview of GDP flash estimation methods*. (Statistical Working Papers). Luxembourg: Publications Office of the European Union.

Eurostat. (2019). Preliminary GDP flash estimate in 30 days for Europe. Retirado de https://ec.europa.eu/eurostat/statistics-explained/index.php/Preliminary_GDP_flash_estimate_in_30_days_for_Europe#How_is_the_preliminary_flash_produced.3F, acedido em 23.05.2020.

Eurostat. (2020). Release Calendar for Euro Indicators. Retirado de https://ec.europa.eu/eurostat/news/release-calendar, acedido em 20.09.2020.

Galbraith, J., & Tkacz, G. (2007). *Electronic Transactions as High-Frequency Indicators of Economic Activity*. (Working Paper 2007-58). Bank of Canada.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics, 55*(4), 665-676. doi:10.1016/j.jmoneco.2008.05.010

Gil, M., Pérez, J. J., Sanchez Fuentes, A. J., & Urtasun, A. (2018). *Nowcasting private consumption: traditional indicators, uncertainty measures, credit cards and some internet data*. (Documentos de Trabajo, no. 1842). Banco de España.

Golinelli, R., & Parigi, G. (2008). Real-time squared: A real-time data set for real-time GDP forecasting. *International Journal of Forecasting, 24*(3), 368-385. doi:10.1016/j.ijforecast.2008.05.001

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society, 37*(3), 424-438. doi:10.2307/1912791

Grudkowska, S., Buono, D., Palate, J., & Ciebiera, W. (2013). *Advanced Tools for Time Series Analysis and Seasonal Adjustment in the New JDemetra+*. Documento apresentado na JSM2013 Joint Statistical Meetings of American Association, Montreal.

Haluška, J. (2006). The use of business survey results for GDP flash estimates. *Biatec, 14*(4).

Hill, T., O'Connor, M., & Remus, W. (1996). Neural network models for time series forecasts. *Management science, 42*(7), 1082-1092. doi:10.1287/mnsc.42.7.1082

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*(2ª ed.). Retirado de https://otexts.com/fpp2/?fbclid=IwAR1tR3GxcE3CqK7ERP_1c_ZflufOAjjOv-FvDsA1KBSWD_qgyj1zcsi8Nvo, acedido em 20.06.2020.

Instituto Nacional de Estatística. (2018). *Como se calcula o PIB*. Retirado de https://www.ine.pt/ngt_server/attachfileu.jsp?look_parentBoui=340538499&att_display=n&att_download=y, acedido em 03.12.2019.

Instituto Nacional de Estatística. (2020a). *Contas Nacionais Trimestrais - Estimativa Rápida 45 dias*. Retirado de https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=415333594&DESTAQUEStema=00&DESTAQUESmodo=2, acedido em 25.09.2020.

Instituto Nacional de Estatística. (2020b). *Contas Nacionais Trimestrais a 60 dias*. Retirado de https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=414810619&DESTAQUEStema=00&DESTAQUESmodo=2, acedido em 25.09.2020.

Instituto Nacional de Estatística. (2020c). *Contas Nacionais Trimestrais Estimativa Rápida a 30 dias*. Retirado de https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=445252249&DESTAQUESmodo=2, acedido em 26.09.2020.

Karbasi, A. R., Laskukalayeh, S. S., & Fahimifard, S. M. (2009, Agosto). *Comparison of NNARX, ANN and ARIMA Techniques to Poultry Retail Price Forecasting*. Documento apresentado na International Association of Agricultural Economists Conference, Beijing, China.

Kongcharoen, C., & Kruangpradit, T. (2013, Junho). *Autoregressive integrated moving average with explanatory variable (ARIMAX) model for Thailand export*. Documento apresentado na 33rd International Symposium on Forecasting, Seoul, South Korea.

León, C., & Ortega, F. (2018). Nowcasting economic activity with electronic payments data: A predictive modeling approach. *Revista de economía del Rosario, 21(2),* 381-407.

Lourtie, P. (2011). Portugal no contexto da crise do euro. *Relações Internacionais (Revista R:I), nº32*, 061-105.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*: Springer Science & Business Media.

Mazzi, G. L., Moauro, F., Montana, G., & Cannata, R. R. (2016). *Towards a real-time system for the analysis of the euro area cyclical situation*. (Statistical Working Papers). Eurostat.

Modugno, M., Soybilgen, B., & Yazgan, E. (2016). Nowcasting Turkish GDP and news decomposition. *International Journal of Forecasting, 32*(4), 1369-1384. doi:10.1016/j.ijforecast.2016.07.001

Radovan, J. (2017). *Short-term forecasting of Slovenian GDP using monthly information*. (Working Papers, no. 1/2017). Bank of Slovenia.

Raju, S., & Balakrishnan, M. (2019). Nowcasting economic activity in India using payment systems data. *Journal of Payments Strategy & Systems, 13*(1), 72-81.

Rua, A. (2004). Um novo indicador coincidente para a economia portuguesa. *Boletim Económico do Banco de Portugal* (Junho 2004), 21-29.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society, 48*(1), 1-48. doi:10.2307/1912017

Verbaan, R., Bolt, W., & van der Cruijsen, C. (2017). *Using debit card payments data for nowcasting Dutch household consumption*. (DNB Working Papers, no. 571). De Nederlandsche Bank.

Wooldridge, J. M. (2012). Basic regression analysis with time series data. Em *Introductory Econometrics: A Modern Approach* (5ª ed.). Cincinnati, Ohio: South-Western College Publishing.

Zhang, P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159-175. doi:10.1016/S0925-2312(01)00702-0

Zhang, P., Patuwo, E., & Hu, M. (1998). Forecasting With Artificial Neural Networks: The State of the Art. *International Journal of Forecasting, 14*, 35-62. doi:10.1016/S0169-2070(97)00044-7

Zivot E., & Wang J. (2003). Vector Autoregressive Models for Multivariate Time Series. Em *Modeling Financial Time Series with S-Plus*® (pp. 369-413). New York: Springer.

Anexos

Anexo 1 – Estimativas de teste em t+30 para a taxa de crescimento do PIB trimestral e respetivas revisões para a União Europeia (EU)

	GDP growth rates					Revisions to t+30 estimate at			
	t+30	t+45	t+65	t+100	t+3m+65d	t+45	t+65	t+100	t+3m+65d
2012Q1	-0.03	-0.05	-0.02	0.00	-0.02	-0.01	0.01	0.03	0.01
2012Q2	-0.10	-0.16	-0.14	-0.15	-0.15	-0.06	-0.04	-0.05	-0.05
2012Q3	0.25	0.13	0.12	0.11	0.11	-0.12	-0.13	-0.14	-0.14
2012Q4	-0.48	-0.51	-0.49	-0.49	-0.50	-0.04	-0.01	-0.01	-0.02
2013Q1	-0.08	-0.11	-0.06	-0.13	-0.05	-0.03	0.02	-0.04	0.03
2013Q2	0.35	0.32	0.36	0.32	0.36	-0.03	0.01	-0.03	0.01
2013Q3	0.26	0.24	0.23	0.26	0.31	-0.02	-0.03	0.00	0.05
2013Q4	0.37	0.39	0.42	0.39	0.39	0.03	0.05	0.02	0.03
2014Q1	0.42	0.32	0.30	0.33	0.33	-0.09	-0.11	-0.09	-0.08
2014Q2	0.22	0.18	0.16	:	0.23	-0.04	-0.06	:	0.01
2014Q3	0.27	0.30	0.30	:	0.32	0.03	0.02	:	0.05
2014Q4	0.30	0.38	0.41	:	0.43	0.08	0.11	:	0.14
2015Q1	0.34	0.42	0.44	:	0.54	0.08	0.10	:	0.20
2015Q2	0.38	0.41	0.45	:	0.46	0.03	0.06	:	0.08
2015Q3	0.31	0.37	0.37	:	:	0.06	0.06	:	:
2015Q4	0.31	0.35	:	:	:	0.03	:	:	:
Average revision						-0.01	0.00	-0.04	0.02
Average absolute revision						0.05	0.05	0.05	0.06
Root of the mean squared errors (RMSQE)						0.06	0.07	0.06	0.08

Fonte: "*Euro area and European Union GDP flash estimates at 30 days*" (Eurostat, 2016a)

Anexo 2 – Estimativas de teste em t+30 para a taxa de crescimento do PIB trimestral e respetivas revisões para a Zona Euro (EA)

	GDP growth rates					Revisions to t+30 estimate at			
	t+30	t+45	t+65	t+100	t+3m+65 d	t+45	t+65	t+100	t+3m+65d
2012Q1	-0.06	0.02	-0.01	0.01	-0.02	0.08	0.05	0.07	0.04
2012Q2	-0.07	-0.18	-0.17	-0.18	-0.17	-0.11	-0.09	-0.10	-0.10
2012Q3	0.07	-0.05	-0.05	-0.07	-0.07	-0.13	-0.12	-0.14	-0.14
2012Q4	-0.51	-0.59	-0.59	-0.59	-0.59	-0.08	-0.09	-0.08	-0.08
2013Q1	-0.21	-0.22	-0.21	-0.27	-0.16	-0.01	0.00	-0.06	0.05
2013Q2	0.31	0.27	0.29	0.26	0.30	-0.04	-0.01	-0.04	-0.01
2013Q3	0.11	0.09	0.08	0.13	0.15	-0.02	-0.03	0.02	0.04
2013Q4	0.26	0.28	0.28	0.23	0.26	0.02	0.02	-0.03	0.00
2014Q1	0.28	0.19	0.18	0.21	0.22	-0.10	-0.10	-0.08	-0.06
2014Q2	0.07	0.05	0.03	:	0.07	-0.03	-0.04	:	0.00
2014Q3	0.13	0.16	0.16	:	0.18	0.03	0.03	:	0.06
2014Q4	0.22	0.34	0.33	:	0.36	0.12	0.11	:	0.14
2015Q1	0.29	0.40	0.37	:	0.52	0.11	0.08	:	0.23
2015Q2	0.27	0.31	0.36	:	0.39	0.04	0.09	:	0.12
2015Q3	0.26	0.30	0.29	:	:	0.04	0.03	:	:
2015Q4	0.24	0.27	:	:	:	0.03		:	:
Average revision						0.00	0.00	-0.05	0.02
Average absolute revision						0.06	0.06	0.07	0.08
Root of the mean squared errors (RMSQE)						0.07	0.07	0.08	0.10

Fonte: "*Euro area and European Union GDP flash estimates at 30 days*" (Eurostat, 2016a)

Anexo 2 – Códigos R

```
################################## CARREGAR PACKAGES ##################################
#######################################################################################
library(ggplot2)
library(tidyverse)
library(vars)
library(caret)
library(xtable)
library(tseries)
library(forecast)
library(astsa)
library(dplyr)
library(MTS)
library(dynlm)
library(kableExtra)
library(tsDyn)

################################## CARREGAR DADOS #####################################
#######################################################################################

# Dados Originais
df1 <- read.csv2("C:/Users/pedro/Desktop/Dados/Dados.1.csv")
colnames(df1) <- c('Ano', 'Trimestre', 'PIB', 'Compras', 'Levantamentos', 'Fatura')

# Dados com sazonalidade corrigida pelo X-13 ARIMA
df2 <- read.csv2("C:/Users/pedro/Desktop/Dados/Dados.2.csv")
colnames(df2) <- c('Ano', 'Trimestre', 'PIB', 'Compras', 'Levantamentos', 'Fatura')

############################### CRIAR SERIES TEMPORAIS ################################
#######################################################################################

# Objeto de series temporais (ts) - Dados Originais
tserie_1 <- ts(df1[,3:6], start=c(2000, 4), frequency=4)

# Objeto de series temporais (ts) - Dados com sazonalidade corrigida pelo X-13 ARIMA
tserie_2 <- ts(df2[,3:6], start=c(2000, 4), frequency=4)

################################## SAZONALIDADE #######################################
#######################################################################################

#### Dados Originais ####

# Grafico com todas as series originais
plot.ts(tserie_1, main = "")

# Determinacao da Componente Sazonal
seasonal = lapply(df1[,3:6], function(x, vars) {
  x <- ts(x, start=c(2000, 4), frequency=4)
  temp <- decompose(x)
  return(temp$seasonal)
})

# Representacao Grafica da Componente Sazonal
plot.ts(do.call(cbind, seasonal), main=" ")

#### Dados com sazonalidade corrigida pelo X-13 ARIMA ####

# Grafico com todas as series temporais
plot.ts(tserie_2, main = "")

# Determinacao da Componente Sazonal
seasonal = lapply(df2[,3:6], function(x, vars) {
  x <- ts(x, start=c(2010, 1), frequency=4)
  temp <- decompose(x)
  return(temp$seasonal)
})

# Representacao Grafica da Componente Sazonal
```

```
plot.ts(do.call(cbind, seasonal), main=" ")

##################################### SUMARIO ########################################
######################################################################################

summary(df2)

######################################################################################
################################ ANALISE MULTIVARIADA ################################
######################################################################################

# Grafico com todas as series
autoplot(tserie_2)

# Grafico das Series Emparelhadas
pairs(tserie_2, cex=0.75, pch=16)


########################### Graficos das Series aos Pares ############################
######################################################################################

n <- nrow(df2)

# Grafico PIB VS Compras
par(mar = c(5, 5, 3, 5))
plot(df2[, 3], type ="l", ylab = "PIB",
     main = "PIB vs Compras", xlab = "Tempo",
     col = "blue")
par(new = TRUE)
plot(df2[, 4], type = "l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col = "red", lty = 2)
axis(side = 4)
mtext("Compras", side = 4, line = 3)
legend("topleft", c("PIB", "Compras"),
       col = c("blue", "red"), lty = c(1, 2))

# Grafico PIB VS Levantamentos
par(mar = c(5, 5, 3, 5))
plot(df2[, 3], type ="l", ylab = "PIB",
     main = "PIB vs Levantamentos", xlab = "Tempo",
     col = "blue")
par(new = TRUE)
plot(df2[,5], type = "l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col = "red", lty = 2)
axis(side = 4)
mtext("Levantamentos", side = 4, line = 3)
legend("topleft", c("PIB", "Levantamentos"),
       col = c("blue", "red"), lty = c(1, 2))

# Grafico PIB VS Fatura
par(mar = c(5, 5, 3, 5))
plot(df2[62:n, 3], type ="l", ylab = "PIB",
     main = "PIB vs Fatura", xlab = "Tempo",
     col = "blue")
par(new = TRUE)
plot(df2[62:n,6], type = "l", xaxt = "n", yaxt = "n",
     ylab = "", xlab = "", col = "red", lty = 2)
axis(side = 4)
mtext("Fatura", side = 4, line = 3)
legend("topleft", c("PIB", "Fatura"),
       col = c("blue", "red"), lty = c(1, 2))

################################## Estacionariedade ##################################
######################################################################################

# Série c/ Todas as séries s/Último trimestre
tserie_5 <- ts(df2[1:78,3:6], start=c(2000, 4), end = c(2020,1), frequency=4)
```

```
# Série s/ Fatura e s/Último trimestre
tserie_6 <- ts(df2[1:78,3:5], start=c(2000, 4), end = c(2020,1), frequency=4)

# Série c/ Todas as séries c/Último trimestre
tserie_7 <- ts(df2[1:79,3:6], start=c(2000, 4), end = c(2020,2), frequency=4)

# Testes de Dickey-Fuller Aumentados
apply(tserie_5, 2, adf.test)
apply(tserie_6, 2, adf.test)

# Primeira diferença
tserie_5.1 <- diffM(tserie_5)
tserie_6.1 <- diffM(tserie_6)
tserie_7.1 <- diffM(tserie_7)

# Caixa de Bigodes
df3 <- data.frame(tserie_7.1)
df_long <- gather(df3, condition, measurement, PIB:Fatura, factor_key=TRUE)
ggplot(df_long, aes(x=condition, y=measurement)) +
  geom_boxplot()

# Teste de Dickey-Fuller
apply(tserie_5.1, 2, adf.test)
apply(tserie_6.1, 2, adf.test)

# Segunda diferença
# Série c/ Todas as séries s/Último trimestre
tserie_5.2 <- ts(diffM(tserie_5.1), start = c(2001,2), end = c(2020,1), frequency = 4)

# Série s/ Fatura e s/Último trimestre
tserie_6.2 <- ts(diffM(tserie_6.1), start = c(2001,2), end = c(2020,1), frequency = 4)

# Série c/ Todas as séries c/Último trimestre
tserie_7.2 <- ts(diffM(tserie_7.1), start = c(2001,2), end = c(2020,2), frequency = 4)

################################## POS TRANSFORMACAO ###################################
########################################################################################

# Grafico com todas as series transformadas
autoplot(tserie_5.2)

# Calculo das correlações
cor_test1 <- cor(tserie_6.2[1:76, 1:3])
round(cor_test1, 2)
cor_test2 <- cor(tserie_5.2[62:76, 1:4])

######################################## ARIMAX ########################################
########################################################################################

###################################### MODELO 1 ########################################

#### Ajustamento do Modelo 1 ####
# Série de Treino
train1 <- window(tserie_7.2[, 1:3], start=c(2001, 2), end=c(2018, 4))

(fit1 <- auto.arima(train1[,1], xreg= train1[,2:3]))

# Sumário Modelo 1
summary(fit1)

# Grafico dos Valores Ajustados
plot(train1[,'PIB'], ylab='dPIB')
points(train1[,'PIB'], ylab='dPIB')
lines(fitted(fit1), col='red')
points(fitted(fit1), col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))
```

```
#### Residuos do Modelo  1####

ggtsdisplay(residuals(fit1), main="Erros Modelo 1")

# Teste de Box-Ljung
Box.test(residuals(fit1), fitdf=5, lag=10, type="Ljung")

#### Previsão do Modelo 1 ####

# Série de Teste
test1 <- window(tserie_7.2[, 1:3], start=c(2019, 1), end=c(2020, 2))
bind_1 <- cbind(comp=test1[,'Compras'], levan=test1[,'Levantamentos'])

fcast_1 <- forecast(fit1, level=95, xreg=bind_1)

# Gráfico dos valores previstos
plot(tserie_7.2[, 'PIB'], ylab='dPIB', xlim=c(2010, 2021),
     ylim=range(fcast_1$lower, fcast_1$upper, tserie_7.2[, 'PIB']))
lines(test1[, 'PIB'], col = 'darkgrey')
points(test1[, 'PIB'], col = 'darkgrey')
lines(fcast_1$upper, col = 'red', lty = 2)
points(fcast_1$mean, col = 'blue')
lines(fcast_1$mean, col = 'blue')
lines(fcast_1$lower, col = 'red', lty = 2)
legend('bottomleft', c('Dados de Treino', 'Dados de Validação', 'Pontos Estimados', '95%
IC'),
       col=c(1, 'grey', 'blue', 'red'), lty=c(1,1,1,2), cex=0.75)

##################################### MODELO 2 #####################################

#### Ajustamento do Modelo 2 ####

# Série de Treino
train2 <- window(tserie_7.2[, 1:4], start=c(2016, 3), end=c(2019,4))

(fit2 <- auto.arima(train2[,1], xreg= train2[,2:4]))

# Sumário Modelo 2
summary(fit2)

# Grafico dos Valores Ajustados
plot(train2[,'PIB'], ylab='dPIB')
points(train2[,'PIB'], ylab='dPIB')
lines(fitted(fit2), col='red')
points(fitted(fit2), col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))

#### Residuos do Modelo 2 ####

ggtsdisplay(residuals(fit2), main="Erros Modelo 2")

# Teste de Box-Ljung
Box.test(residuals(fit2), fitdf=5, lag=10, type="Ljung")

#### Previsão do Modelo 2 ####

# Série de Teste
test2 <- window(tserie_7.2[, 1:4], start=c(2020, 1), end=c(2020, 2))

bind_2 <- cbind(comp=test2[,'Compras'], levan=test2[,'Levantamentos'],
fat=test2[,'Fatura'])

fcast_2 <- forecast(fit2, level=95, xreg=bind_2)

# Gráfico da Previsão do Modelo 2
plot(tserie_7.2[, 'PIB'], ylab='dPIB', xlim=c(2010, 2021),
     ylim=range(fcast_2$lower, fcast_2$upper, tserie_7.2[, 'PIB']))
```

```
lines(test2[, 'PIB'], col = 'darkgrey')
points(test2[, 'PIB'], col = 'darkgrey')
lines(fcast_2$upper, col = 'red', lty = 2)
points(fcast_2$mean, col = 'blue')
lines(fcast_2$mean, col = 'blue')
lines(fcast_2$lower, col = 'red', lty = 2)
legend('bottomleft', c('Dados de Treino', 'Dados de Validação', 'Pontos Estimados', '95%
IC'),
       col=c(1, 'grey', 'blue', 'red'), lty=c(1,1,1,2), cex=0.75)


######################################### VAR ##########################################
#########################################################################################

###################################### MODELO 3 ########################################

#### Ajustamento Modelo 3 ####

# Seleção do desfasamento
VARselect(na.omit(train1[,c(1:3)]),
          type="none"
)[["selection"]]

# Determinacao do modelo VAR baseado em VARselect
var1 <-  vars::VAR(na.omit(train1[,c(1:3)]), p=4, type="none")
#Sumário
summary(var1)

# Granger causality PIB
causality(var1, cause="PIB")

# Granger causality Compras
causality(var1, cause="Compras")

# Granger causality Levantamentos
causality(var1, cause="Levantamentos")

#### Estimar novamente sem Levantamentos

# Seleção do desfasamento
VARselect(na.omit(train1[,c(1:2)]),
          type="none"
)[["selection"]]

# Determinacao do modelo VAR baseado em VARselect
var1 <-  vars::VAR(na.omit(train1[,c(1:2)]), p=4, type="none")

# Serie Temporal dos Valores Ajustados
ts_fitted_1 <- ts(fitted(var1), start=c(2002, 2), end = c(2018 ,4), frequency=4)
ts_fit_1 <- ts(ts_fitted_1[, 1], start=c(2002, 2), end = c(2018 ,4), frequency=4)

#Sumário
summary(var1)

# Grafico dos Valores Ajustados
plot(train1[,'PIB'], ylab='PIB')
points(train1[,'PIB'], ylab='PIB')
lines(ts_fit_1, col='red')
points(ts_fit_1, col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))

#### Resíduos Modelo 3 ####

ggtsdisplay(residuals(var1)[,1], main="Erros Modelo 3")
Box.test(residuals(var1)[,1], fitdf=5, lag=10, type="Ljung")

#### Previsão Modelo 3 ####
```

```
# Série de teste
trend <- time(test1)

fcast_3 <- predict(var1, n.ahead=nrow(test1), dumvar = matrix(trend))

names(fcast_3$fcst) <- colnames(test1)

# Gráfico da Previsão do Modelo 3
plot(tserie_7.2[, 'PIB'], ylab='dPIB', xlim = c(2010, 2020),
     ylim=range(fcast_3$fcst$'PIB'[, 1:2], tserie_7.2[, 'PIB']))
lines(test1[, 'PIB'], col='darkgrey')
points(test1[, 'PIB'], col='darkgrey')
fcast_3$fcst$'PIB' <- ts(fcast_3$fcst$'PIB', start=c(2019, 1), frequency=4)
lines(fcast_3$fcst$'PIB'[,2], col='red', lty=2)
lines(fcast_3$fcst$'PIB'[,1], col='blue')
points(fcast_3$fcst$'PIB'[,1], col='blue')
lines(fcast_3$fcst$'PIB'[,3], col='red', lty=2)
legend('bottomleft', c('Dados de Treino', 'Dados de Validação', 'Pontos Estimados', '95%
IC'),
       col=c(1, 'grey', 'blue', 'red'), lty=c(1,1,1,2), cex=0.75)

###################################### MODELO 4 ######################################

# Seleção do desfasamento
VARselect(na.omit(train2[,c(1:4)]),
          type="none"
)[["selection"]]

# Determinacao do modelo VAR baseado em VARselect
var2 <-  vars::VAR(na.omit(train2[,c(1:4)]), p=1, type="none")

# Sumário
summary(var2)

# Serie Temporal dos Valores Ajustados
ts_fitted_2 <- ts(fitted(var2), start=c(2016, 4), end = c(2019 ,4), frequency=4)
ts_fit_2 <- ts(ts_fitted_2[, 1], start=c(2016, 4), end = c(2019 ,4), frequency=4)

# Grafico dos Valores Ajustados
plot(train2[, 'PIB'], ylab='PIB', ylim=range(-2000:1000))
points(train2[, 'PIB'], ylab='PIB')
lines(ts_fit_2, col='red')
points(ts_fit_2, col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))

# Granger causality PIB
causality(var2, cause="PIB")

# Granger causality Compras
causality(var2, cause="Compras")

# Granger causality Levantamentos
causality(var2, cause="Levantamentos")

# Granger causality Fatura
causality(var2, cause="Fatura")

#### Resíduos Modelo 4 ####

ggtsdisplay(residuals(var2)[,1], main="Erros Modelo 2")

# Teste Box-Ljung
Box.test(residuals(var2)[,1], fitdf=5, lag=10, type="Ljung")

#### Previsão Modelo 4 ####
```

```
# Série de teste
trend2 <- time(test2)

fcast_4 <- predict(var2, n.ahead=nrow(test2), dumvar = matrix(trend2))

names(fcast_4$fcst) <- colnames(test2)

# Gráfico da Previsão do Modelo 4
plot(tserie_7.2[, 'PIB'], ylab='dPIB', xlim = c(2010, 2020),
     ylim=range(fcast_4$fcst$'PIB'[, 1:4], tserie_7.2[, 'PIB']))
lines(test2[, 'PIB'], col='darkgrey')
points(test2[, 'PIB'], col='darkgrey')
fcast_4$fcst$'PIB' <- ts(fcast_4$fcst$'PIB', start=c(2020, 1), frequency=4)
lines(fcast_4$fcst$'PIB'[,2], col='red', lty=2)
lines(fcast_4$fcst$'PIB'[,1], col='blue')
points(fcast_4$fcst$'PIB'[,1], col='blue')
lines(fcast_4$fcst$'PIB'[,3], col='red', lty=2)
legend('bottomleft', c('Dados de Treino', 'Dados de Validação', 'Pontos Estimados', '95%
IC'),
       col=c(1, 'grey', 'blue', 'red'), lty=c(1,1,1,2), cex=0.75)

##################################### NNARX ##########################################
#######################################################################################

###################################### MODELO 5 ######################################

#### Ajustamento do Modelo 5 ####

bind_3 <- cbind(comp=train1[,'Compras'], levan=train1[,'Levantamentos'])
train_trend1 = time(train1); test_trend1 = time(test1)

# Ajustar o Modelo NNAR
attach(as.data.frame(train1))

ts_reg_3 <- nnetar(`PIB`, xreg=bind_3)

detach(as.data.frame(train1))

# Grafico dos Valores Ajustados
plot(train1[,'PIB'], ylab='PIB')
points(train1[,'PIB'], ylab='PIB')
lines(ts(fitted(ts_reg_3), start=c(2001, 2), frequency=4), col='red')
points(ts(fitted(ts_reg_3), start=c(2001, 2), frequency=4), col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))

#### Resíduos Modelo 4 ####

ggtsdisplay(residuals(ts_reg_3), main="Erros Modelo 5")

# Teste Box-Ljung
Box.test(residuals(ts_reg_3), fitdf=5, lag=10, type="Ljung")

#### Previsão Modelo 5 ####

bind_4 <- cbind(comp=test1[,'Compras'], levan=test1[,'Levantamentos'])

# Previsao para Modelo Compras + Levantamentos
fcast_5 <- forecast(ts_reg_3, level=95, xreg=bind_4, PI=T)

fcast_5$x <- ts(fcast_5$x, start=c(2001, 2), frequency=4)
fcast_5$mean = ts(fcast_5$mean, start=c(2019, 1), frequency=4)
fcast_5$upper = ts(fcast_5$upper, start=c(2019, 1), frequency=4)
fcast_5$lower = ts(fcast_5$lower, start=c(2019, 1), frequency=4)

# Gráfico da Previsão
plot(tserie_7.2[,'PIB'], xlim = c(2000, 2020), ylab='PIB',
     ylim=range(fcast_5$lower, fcast_5$upper, tserie_7.2[, 'PIB']))
```

```
lines(test1[,'PIB'], col = 'darkgrey')
points(test1[,'PIB'], col = 'darkgrey')
lines(fcast_5$upper, col = 'red', lty = 2)
lines(fcast_5$mean, col = 'blue')
points(fcast_5$mean, col = 'blue')
lines(fcast_5$lower, col = 'red', lty = 2)
legend('bottomleft', c('Dados de Treino', 'Dados de Validação', 'Pontos Estimados', '95%
IC'),
       col=c(1, 'grey', 'blue', 'red'), lty=c(1,1,1,2), cex=0.75)



##################################### MODELO 6 #######################################

#### Ajustamento do Modelo 6 ####

bind_5 <- cbind(comp=train2[,'Compras'], levan=train2[,'Levantamentos'],
fat=train2[,'Fatura'])

# Ajustar o Modelo NNAR
attach(as.data.frame(train2))

ts_reg_4 <- nnetar(`PIB`, xreg=bind_5)

detach(as.data.frame(train2))

# Grafico dos Valores Ajustados
plot(train2[,'PIB'], ylab='PIB')
points(train2[,'PIB'], ylab='PIB')
lines(ts(fitted(ts_reg_4), start=c(2016, 3), frequency=4), col='red')
points(ts(fitted(ts_reg_4), start=c(2016, 3), frequency=4), col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))

#### Resíduos Modelo 6 ####

ggtsdisplay(residuals(ts_reg_4), main="Erros Modelo 6")

# Teste Box-Ljung
Box.test(residuals(ts_reg_4), fitdf=5, lag=10, type="Ljung")

#### Previsão Modelo 6 ####

bind_6 <- cbind(comp=test2[,'Compras'], levan=test2[,'Levantamentos'],
fat=test2[,'Fatura'])

# Previsao para Modelo 6
fcast_6 <- forecast(ts_reg_4, level=95, xreg=bind_6, PI=T)

fcast_6$x <- ts(fcast_6$x, start=c(2001, 2), frequency=4)
fcast_6$mean = ts(fcast_6$mean, start=c(2020, 1), frequency=4)
fcast_6$upper = ts(fcast_6$upper, start=c(2020, 1), frequency=4)
fcast_6$lower = ts(fcast_6$lower, start=c(2020, 1), frequency=4)

# Gráfico da Previsão
plot(tserie_7.2[,'PIB'], xlim = c(2016, 2021), ylab='PIB',
     ylim=range(fcast_6$lower, fcast_6$upper, tserie_7.2[, 'PIB']))
lines(test2[,'PIB'], col = 'darkgrey')
points(test2[,'PIB'], col = 'darkgrey')
lines(fcast_6$upper, col = 'red', lty = 2)
lines(fcast_6$mean, col = 'blue')
points(fcast_6$mean, col = 'blue')
lines(fcast_6$lower, col = 'red', lty = 2)
legend('bottomleft', c('Dados de Treino', 'Dados de Validação', 'Pontos Estimados', '95%
IC'),
       col=c(1, 'grey', 'blue', 'red'), lty=c(1,1,1,2), cex=0.75)
```

```
###################################### RMSE ########################################
###################################################################################

# Criacao de Tabela de Resultados
results <- matrix(0, nrow=1, ncol=5,
                  dimnames=list(c('RMSE'),
                                c('fcast_1', 'fcast_3', 'fcast_4', 'fcast_5',
'fcast_6')))

# RMSE Modelo 1
results[ ,"fcast_1"] <- sqrt(c(mean((test1[,"PIB"] - fcast_1$mean)^2)))

# RMSE Modelo 3
results[, "fcast_3"] <- sqrt(c(mean((test1[,'PIB'] - fcast_3$fcst$'PIB'[,'fcst'])^2)))

# RMSE Modelo 4
results[, "fcast_4"] <- sqrt(c(mean((test2[,'PIB'] - fcast_4$fcst$'PIB'[,'fcst'])^2)))

# RMSE Modelo 5
results[, 'fcast_5'] <- sqrt(c(mean((fcast_5$mean - test1[,'PIB'])^2)))

# RMSE Modelo 6
results[, 'fcast_6'] <- sqrt(c(mean((fcast_6$mean - test2[,'PIB'])^2)))

################################## MODELO FINAL ####################################
###################################################################################

# Compras + Levantamentos
(fit1 <- auto.arima(tserie_5.2[,1], xreg= tserie_5.2[,2:3]))
(mod1 <- Arima(tserie_5[1:78,1], order = c(1,2,2), xreg=tserie_5[1:78,2:3]))
(MOD1 <- sarima(tserie_5[1:78,1], p = 1, d=2, q =2,  xreg=tserie_5[1:78,2:3])) ### muito
bom este
MOD1 $ttable

fitted_1 <- ts(fitted(mod1), start = c(2000,4), frequency = 4)

# Grafico dos Valores Ajustados
plot(tserie_5[,'PIB'], ylab='PIB')
points(tserie_5[,'PIB'], ylab='PIB')
lines(fitted_1, col='red')
points(fitted_1, col='red')
legend('bottomleft', c('Dados Originais', 'Valores Ajustados'), lty=c(1,1),
col=c(1,'red'))

#Previsão
fcast1 <- forecast(mod1, xreg=tserie_2[c(79, 70:76),2:3])

autoplot(fcast1) + xlab("Ano") +
  ylab("Valores do PIBl") +
  ggtitle("Previsões com regressãoo e erros ARIMA(1, 2, 2)")
```