U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Nowcasting VAT data in the retail trade sector using historical data and electronic payment data

Moyses Xavier Fontoura Neto

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:

PhD Maria Eduarda Silva

2023

Abstract

Timely information about the current state of the economy is essential, as it influences the population's input and output choices, and allows the government to react as quickly as possible to economic events that need intervention. Given the importance of timely information, EUROSTAT has regulated the Short-term business statistics (STS), setting deadlines, quality requirements, and other guidelines for their publication.

One of the STS published by the National Statistical Institute (INE) is the Retail Trade Turnover Index (RTTI), which is broken down into different economic activities classifications (CAEs). The estimation of the RTTI relies on important data provided by the Tax Authority - the e-Fatura data. However, sometimes the Tax Authority fails to deliver the data in time, which poses a threat to the compliance of the quality requirements of this early index.

To mitigate this risk, INE decided to build a framework to estimate in a timely manner (nowcast) this data for whenever it is not delivered in time again. To this end, in addition to historical e-Fatura data, Multibanco data is used as an auxiliary variable to nowcast the e-Fatura data. The models used were ARIMA, Linear regression, Dynamic regression, MIDAS regression and the mean of these four models' nowcasts.

After analysing the relationship between the response and the auxiliary variables, the nowcasting exericise was carried out. The results obtained showed that there is not a single model that can nowcast the e-Fatura data for all CAEs with the best accuracy. Although the models that used the Multibanco data as an auxiliary variable had the expectation to perform better than the classical ARIMA approach, the model that performed better for almost half of the CAEs was the ARIMA model, followed by the Linear regression, the mean, the Dynamic regression and the

MIDAS regression.

Keywords: Time series; Nowcasting; ARIMA; Linear regression; Dynamic regression; MIDAS; VAT data; Financial data; Short-term business statistics;

Resumo

Informações e dados em tempo útil sobre o atual estado da economia são essenciais, já que influenciam as escolhas de input e output da população, e permite que o governo reaja o mais rápido possível aos eventos económicos que precisam de intervenção. Considerando isto, o EUROSTAT regulamentou as Estatísticas de Conjuntura das Empresas (STS), definindo prazos, padrões de qualidade, e outras instruções relativas à publicação.

Uma das STS que são publicadas pelo Instituto Nacional de Estatística (INE) é o Índice de Volume de Negócios (RTTI), que dá informação sobre várias atividades económicas, seguindo o esquema CAE. A estimação do RTTI depende dos dados sobre o e-Fatura, que são enviados pela Autoridade Tributária (AT). No entanto, a AT, por vezes, não entrega estes dados nos prazos estabelecidos, gerando riscos em relação ao prazo de publicação e também à qualidade das estimativas.

Para amenizar estes riscos, o INE decidiu definir uma abordagem para estimar estes dados, em tempo útil, sempre que a AT não os consiga entregar a tempo. Além dos dados históricos do e-Fatura, também foram utilizados os dados do Multibanco como variável auxiliar na estimação. Os modelos utilizados foram ARIMA, Regressão linear, Regressão dinâmica, Regressão MIDAS, e a média da estimação destes quatro modelos.

Após a análise da relação entre os dados do e-Fatura e do Multibanco, foi feita a estimação dos modelos. Os resultados obtidos mostraram que não há apenas um modelo que consiga fazer a previsão com a melhor precisão possível para todas as CAEs. Embora a expectativa tenha sido que os modelos que utilizam os dados do Multibanco como variável auxiliar teriam melhor desempenho que o modelo ARIMA, o modelo que teve o melhor desempenho para quase metade

das CAEs foi o ARIMA, seguido pela Regressão linear, pela média, pela Regressão dinâmica e pela Regressão MIDAS.

Palavras-chave: Séries temporais; Nowcasting; Previsão; ARIMA; Regressão linear; Regressão dinâmica; MIDAS; e-Fatura; Dados financeiros; Estatísticas conjunturais

Contents

List of Figures

List of Tables

Chapter 1

Introduction

To achieve effectiveness and efficiency in decision making, it is crucial to have access to reliable, accurate and timely information. The lack of it can lead people, organisations, businesses and governments to make bad choices (Eslake, 2006). Nowadays, there are several indicators that signalise the current state of a country's economy, such as the quarterly *Gross Domestic Product* (GDP), and the *Short-term business statstics* in European case.

In industry, the information about the current state of the economy that is available for companies, as well as their expectations for the near future, influence their input and output choices (such as investment and employment), which highly impact their profitability.

Regarding governments, monetary policy is one of their two main means of influencing the direction and the pace of the economic activity, including employment, GDP and the general rate of which the prices evolve (inflation) (Friedman, 2000). Having timely data is essential to make efficient monetary policies, and the accuracy of these data is crucial for effectiveness (Bernanke, Boivin, & Eliasz, 2005).

Such needs have given rise to the concept of *Nowcasting* in Economics, a contraction of *forecasting* and *now*, which means the prediction of the very recent past, the present and the very near future (Banbura, Giannone, & Reichlin, 2010).

Typically, policy makers use incomplete information to make decisions, since key statistics and indicators are published with significant delay after the end of the reference period and are

frequently reviewed. Other economic agents also need to make decisions that are dependent on the unknown current situation of the economy, and publication lags create uncertainty about the recent past and the present.

In order to reduce these uncertainties, Nowcasting models have been developed in the last few decades and have been used extensively by Central Banks, National Statistical Institutes and other institutions (Richardson, van Florenstein Mulder, & Vehbi, 2021).

Many of these models have been used by National Statistical Institutes (NSIs) to estimate short-term statistics that keep track of the current state of the economy and its most relevant variables. In the European Union, these statistics are the *Short-term business statistics* mentioned above, and one of the most important indices of these statistics is the *Business turnover, employment, wage and hours worked index*.

In the past, the statistical operation to estimate the *Business turnover, employment, wage and hours worked index* relied solely on a questionnaire posed on companies to collect data about volume of sales, turnover, employment (number of employees and salaries), as well as the employees' hours worked.

Progressively, Statistics Portugal (INE) is replacing the questions in the survey by administrative data in order to reduce statistical burden on companies. In 2017, INE began to use administrative data about monthly salaries from the Social Security (DMR/SS) as source of data for the Employment, wage and hours worked indices. In retail, the DMR/SS became the only source of data to estimate these indices, whereas in Industry and Services it replaced only partially the survey.

Currently, in the retail sector, INE is using electronic invoice (*e-Fatura*), as well as a questionnaire posed on companies, to estimate the *Retail Trade Turnover Index*. The *e-Fatura* data is provided by the Tax Authority, and is sent to INE just a few days before the publication of these indices, which gives the Intitute a small window of time to work.

1.1 Motivation and problem description

Although the e-Fatura data is an important information used to estimate the Retail Trade Turnover, the Tax Authority has failed to deliver the data in time in a few occasions, which affects the Institute, as there is a deadline enforced by Eurostat for publishing the estimation of the index. Because of that, Statistics Portugal is interested in finding alternative data sources that could either replace or used to estimate the e-Fatura data whenever Tax Authority fails to deliver them in time again.

Considering the situation described above, this project focused on using alternative data, called *Multibanco*, in order to estimate the e-Fatura data and provide the Institute with a reliable and fast framework to estimate the e-Fatura data for a given reference month, reducing the risk of not complying with the deadlines and also reducing the inaccuracy and quality of these early estimates. This work was part of an internship at INE in order to obtain the European Master in Official Statistics (EMOS) certification.

1.2 Structure of the dissertation

This dissertation is organised in six Chapters. The first provides an introduction and the description of the problem; the second Chapter describes concepts that are relevant and used in this project. Chapter 3 describes the statistical models used in this work, which are applied following the methodological framework described in Chapter 4. Finally, the results of the application are shown in Chapter 5, which are used to draw the conclusions written in the sixth chapter, followed by an Appendix section that has relevant tables with the results of this project.

Chapter 2

Background

This chapter aims to shortly introduce important information about the data that was used in this project, as well as explain the relevant statistics for the European economy that depend on them.

2.1 Classification of Economic Activities

The *Portuguese Classification of Economic Activities*, also known as *CAE*, is the acronym used in Portugal to designate the numerous economic activities. In 1953, the first version of the CAE was published by Statistics Portugal as a translation of the *International Standard Industrial Classification of All Economic Activities (ISIC)*.

ISIC, as explained by United Nations (2008), is a system that provides a set of categories to classify the various economic activities, intended to be a standard classification that can be used for the data collection and publication of statistics of such activities. The system came out in 1949 with the purpose of providing an up-to-date framework for international comparison of national statistics, meeting the pressing needs for international comparability of such statistics.

Since then, the ISIC has been widely used internationally to classify data according to the kind of economic activity in the fields of population, employment, production, national income and other economic statistics. As a result, substantial comparability has been attained by countries that have adopted this system as a national standard or have rearranged their statistical data in

accordance with it. (United Nations, 2008)

After the ISIC release, changes in the economy took place and new types of economic activity emerged and became important, requiring the creation of new distinctions and the shifting of some groups' positions. In order to adapt the system to the then economic reality, the United Nations undertook the first revision of the ISIC and issued it in 1958 (United Nations, 1958), which was translated by Statistics Portugal and published in the country in 1961.

However, the carry-out of statistical work in the country revealed that the ISIC Rev.1 was not sufficient to meet the national needs at the time, which led to the publication of the first CAE adapted to the Portuguese economic reality in 1964. This CAE was developed using the ISIC Rev.1 as its foundation.

Five years later, in 1969, the second revision of the ISIC was issued, and its translation was published in 1970 after the approval of the United Nations Statistical Commission. As the ISIC Rev.2 also didn't meet the needs of the country, the National Council of Statistics of Portugal (CNE) named a comittee responsible for developing a new CAE based on the second revision of the ISIC, which was published under the name of CAE-Rev.1 in 1973.

In 1978, in order to meet the requirement of adjusting the national statistical system to the needs emerged from the process of Portugal joining the European Economic Community (EEC), the CNE created a comittee responsible for 2 projects: 1) reviewing the CAE-Rev.1 in accordance with the General Industrial Classification of Economic Activities within the European Communities (NACE) from 1970; and 2) creating the National Classification of Goods and Services (CNBS). These two projects were completed in 1985, but faced a disapproval due to the suspension of the CNE in 1986.

Later on, the CAE-Rev.2 was developed considering the NACE Rev.1 (1990) and was harmonised as much as possible with the CAE-Rev.1 (1973), being approved in 1991 by the EEC Commission and published in *Diário da República* (DR) in 1993. Its sucessor, the CAE-Rev.2.1, was approved in 2002 and published in DR in 2003.

Finally, the current version of the Portuguese CAE, the CAE-Rev.3, was created. It was developed in accordance with the ISIS Rev.4 (2008) and the NACE Rev.2 (2008), and then published in *Diário da República* in 2007 after being approved by EUROSTAT (Statistics Portugal,

2007)

2.2 Short-term business statistics

Also known as STS, the *Short-term business statistics* are the earliest statistics available that keep track of emerging economic trends in the European Union and third countries in a given reference period. They are index data and report information on a broad range of economic activities, covering the industry, construction, trade and services sectors, (EUROSTAT, 2022c).

STS are of great significance and provide essential information for businesses, academia and policy makers. In conjunction with other data, such as national accounts, the STS are used by the European Central Bank, the European Commission, companies, financial markets, national governments and national central banks to perform economic analysis, making decision-making and the monitoring of the economy easier.

Furthermore, as stated by (EUROSTAT, 2022b), almost half of the *Principal European Economic Indicators* (PEEIs) come from the STS. The PEEIs are key macroeconomic indicators that describe the labour market and the economic situation, as well price developments in the Euro Area and in the EU, which are of great importance for economic and monetary policy.

The EU Regulations No. 2019/2151 and 2020/1197 determine the scope of these short-term indicators, as well as their definition, reference period, form, degree of detail, deadlines, and the starting date of their time-series. Generally, the STS are published in the form of unadjusted, calendar adjusted, and calendar and seasonally adjusted data. The data used to compute them are mainly sourced by business surveys, but administrative data and other sources are also used.

Moreover, according to (EUROSTAT, 2022a), the comparability of STS data between Member States is ensured by methodological frameworks, such as NACE, discussed in the previous subsection, and by data harmonisation methods specified in the regulations cited above and in the methodological manuals for these indicators.

2.2.1 The turnover and employment indices

The *Turnover and employment indices* is an statistical operation carried out and funded by Statistics Portugal, under the scope of the STS, and is composed by four indicators that measure short-term changes in business turnover of goods and services, employment, wages, work input and production volumes over a given reference period (Statistics Portugal, 2019). These four short-term indicators are:

1. *Business turnover, employment, wage and hours worked index in industry*;
2. *Production, employment, wage and hours worked index in Construction and Public Work*;
3. *Business turnover, employment, wage and hours worked index in retail trade*; and
4. *Business turnover, employment, wage and hours worked index in services*.

These indices are estimated using direct and indirect sources, also, Statistics Portugal has been progressively replacing the surveys by administrative data in order to reduce the statistical burden on companies and other data subjects.

For instance, in Retail Trade, the *Monthly Salary Statement of the Social Security* (DMR/SS) has replaced a survey that was posed to firms to estimate the Employment and wages indices. In Industry and in Services, the Employment and wages indices has also started to use administrative data, replacing partly the data that was obtained only by survey. In Construction, it is still used direct sources instead of indirect/administrative ones, and the Hours worked index is still estimated using direct sources as well.

The target population of this statistical operation, based on the CAE Rev.3, is:

- Industry: Sections B, C, D, and E
- Construction and Public Work: Section F
- Retail trade: Division 47
- Services: Sections G (except for division 47), H, I, J, L (except for subclass 68322), M, and N.

The data collection is carried out by autocompletion of the surveys, either online or in paper, available from the first day of the month following the reference period onwards. In average, the times to complete the surveys are: 20 minutes in Industry, 14 minutes in Construction and

Public work, 6 minutes in Retail Trade, and 5 minutes in Services.

The deadline for publishing the index is different for each sector.

2.3 e-Fatura

With the goal of preventing tax avoidance, in July 2012, the Government created the *e-Fatura*, which is a system for invoice issuance that was implemented on 1 January 2013. In addition, other fiscal changes to prevent tax avoidance took effect in the country in the same year, along with the creation of some tax benefits to industries that typically issue invoices less often, such as hospitality, hair and beauty, catering, and repair of motor vehicles and motorcycles, (Autoridade Tributária, n.d.).

Some of these changes are relevant to mention, such as the Decree Law 197/2012 of 24 August, which made mandatory for companies and other entities to issue invoices, whether the ultimate consumer asks for them or not. Another relevant change was making mandatory for companies and other entities to send to the Tax Authority on a monthly basis the invoice documents issued by them up until the 25th day of the following month.

The changes mentioned above are highly relevant for this work, as they influence when the e-Fatura data (VAT data) for a given reference period is supposed to be available and complete, and also improve the coverage of the data.

2.4 Multibanco

The *Multibanco* network is a single system shared across all banks based in Portugal. It integrates ATMs and Point-of-Sales systems, and is designed to process electronic payments in the whole country. Created by the SIBS (Sociedade Interbancária de Serviços) group on the 2nd of September of 1985, the project's system could carry out only 3 types of operations by the time of its launch: cash withdrawal, account balance checking and Card PIN changing (Marques, 2014). Nowadays, it is possible to carry out over 90 operations, such as topping up mobile phones, transferring money, payment of private and public services, among others.

The dataset that the Multibanco network provides has weekly and monthly frequency, and consists of transactions carried out under the network, along with the CAE number associated to the commercial establishment that is on the receiving end of the transactions.

Chapter 3

Literature Review

This Chapter briefly describes the methods and models used in this project. It is important to note that Section 3.1 is heavily based on the books about time series forecasting written by R. J. Hyndman and Athanasopoulos (2018) and Montgomery, Jennings, and Kulahci (2015), and Section 3.2 is based on articles published by Ghysels, Sinko, and Valkanov (2007).

3.1 Time Series

Generally speaking, a time series is a set of observations on a given quantifiable variable indexed in time. Usually the observations are taken at regular intervals - although irregular time series observations do exist.

Time series data are common in many fields, including:

- Medicine: heartbeat rate per minute, brain wave activity per second, etc.
- Economics: quarterly GDP, monthly inflation rate, etc.
- Finance: monthly revenue, daily cash flow, etc.
- Meteorology: daily precipitation rate, daily average temperature, etc.

The analysis of a time series can be undertaken for many reasons and with many goals, such as predicting future values of the series (*forecasting*), identifying the underlying phenomenon represented by the series and understanding its nature, as well as finding the cause of unusual observations (*anomalies*) and recurring patterns.

By definition, a time series is a realisation limited in time of a stochastic process, which can be defined as the family of random variables $\boldsymbol{X} = \{X_t : t \in \mathcal{T}\}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for each $t \in \mathcal{T}$, there is a random variable $X_t : \Omega \to \mathbb{R}$, where $\mathcal{T}$ is an index set. If $\mathcal{T} \in \mathbb{N}$, then $\boldsymbol{X}$ is a stochastic process with discrete parameter, whereas if $\mathcal{T} \in [0, \infty[$, then the parameter is continuous.

3.1.1 Classical decomposition

One of the most simple ways to describe a time series is breaking down its components using the *Classical decomposition*, which came up in the 1920s. This method is relatively simple and was the starting point for the development of most alternative approaches.

In this method, a time series can be decomposed in the following components:

- Trend (T_t) - represents the underlying long-term direction of the time series, which can be an upward or a downward trend.
- Cycle (C_t) - represents the recurring pattern that do not take place at regular intervals and can be caused by many factors, such as natural cycles, economic cycles, etc.
- Seasonality (S_t) - represents the recurring pattern that takes place at regular intervals, such as daily, weekly or monthly intervals.
- Residual (E_t) - represents the variation in the time series that cannot be explained by the other components.

Usually, the components T_t and C_t are considered a single component called *Trend-cycle*, as separating them is a complex task. However, there are techniques which perform this separation, such as filtering, wavelet analysis and spectral analysis.

Supposing a time series with T periods of data available, where T is the most recent observation, and representing the variable of interest X observed at time t by X_t, where $t = 1, 2, ..., T$, the final decomposition can take either an additive (1) or a multiplicative form (2), as follows:

$$(1)\ X_t = T_t + C_t + S_t + E_t$$

$$(2)\ X_t = T_t \times C_t \times S_t \times E_t$$

3.1.2 Stationarity

One of the main concepts in time series analysis is *stationarity*. A stationary time series is easier to work with, since its behaviour does not change over time. This means that the underlying structure of the series is more predictable and consistent. It would be hard, or even impossible, to study and predict a time series if the behaviour of the process is constantly changing over time. This is the reason why the statistical analysis of time series data relies on the concept of stationarity.

A stationary time series has no trend or seasonality, as they would affect the value of the series at different times. Furthermore, the pattern of a stationary series is not predictable in the long-run, as R. J. Hyndman and Athanasopoulos (2018) explain.

Formally, a stochastic process $\boldsymbol{X} = \{X_t : t \in \mathcal{T}\}$ is called *strictly stationary* if its statistical properties do not change over time, i.e., if

$$\boldsymbol{P}(X_{t_1} \leq x_{t_1}, X_{t_2} \leq x_{t_2}, ..., X_{t_k} \leq x_{t_k}) =$$
$$= \boldsymbol{P}(X_{t_1+s} \leq x_{t_1+s}, X_{t_2+s} \leq x_{t_2+s}, ..., X_{t_k+s} \leq x_{t_k+s}), \forall t \in \mathbb{N}, s \in \mathbb{Z}.$$

However, this condition is rarely verified in real time series, which leads to the concept of *weak stationarity*. The conditions for $\boldsymbol{X}$ to be weakly stationary are the following:

1. $\boldsymbol{E}(X_t) = \mu, \forall t \in \mathbb{N}$
2. $\boldsymbol{V}(X_t) = \boldsymbol{E}(X_t - \mu)^2 = \sigma^2, \forall t \in \mathbb{N}$
3. $\boldsymbol{cov}(X_t, X_s) = \boldsymbol{\gamma_X}(t - s), \forall t \in \mathbb{N}, s \in \mathbb{Z}$
4. $\boldsymbol{\rho_X}(s) = \frac{\boldsymbol{\gamma_X}(s)}{\boldsymbol{\gamma_X}(0)}$

Where $\boldsymbol{\gamma_X}(s)$ is the *autocovariance function* and $\boldsymbol{\rho_X}(s)$ is the *autocorrelation function* (ACF) of X_t.

3.1.3 Autocorrelation

Considering that the correlation measures the linear relationship of two variables, the autocorrelation function measures the linear relationship between lagged observations of the same time series.

For instance, the autocorrelation of the stochastic process $\boldsymbol{X}$ at lag 1 is the correlation between X_t and X_{t-1}, and the autocorrelation at lag s is the correlation between X_t and X_{t-s} for all t.

The sample autocorrelation function at lag s is computed as

$$\hat{\boldsymbol{\rho}}_X(s) = \frac{\sum_{t=s+1}^{T}(X_t - \mu)(X_{t-s} - \mu)}{\sum_{t=1}^{T}(X_t - \mu)^2}$$

where T is the last observation.

3.1.4 White noise

A sequence of zero mean, independent and identically distributed random variables is called *white noise*. As such, $\boldsymbol{X}$ is a white noise if

1. $\boldsymbol{E}(X_t) = 0, \forall t \in \mathbb{N}$
2. $\boldsymbol{V}(X_t) = \boldsymbol{E}(X_t^2) = \sigma^2, \forall t \in \mathbb{N}$
3. $\boldsymbol{\rho_X}(0) = 1$ and $\boldsymbol{\rho_X}(s) = 0, s \neq 0, s \in \{\pm 1, \pm 2, \ldots\}$

Usually, a white noise is represented as $X_t \sim WN(0, \sigma^2)$

3.1.5 ARIMA models

The Autoregressive Integrated Moving Average (ARIMA) models, also known as Box-Jenkins models, are statistical models introduced in the 1970s that are used to model and analyse time series data. They are vastly used and are applied in many fields, such as finance and economics.

Autoregressive models

In an AutoRegressive (AR) model, the variable of interest X at time t is explained by using a linear combination of the past values of X. The most simple case of an AR model is the AR model of order 1, also referred as $AR(1)$, which is represented as follows

$$X_t = \phi_1 X_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim WN(0, \sigma^2)$$

In order for the $AR(1)$ model to be stationary, the parameter must meet the condition $|\phi_1| < 1$. Furthermore, it is possible to show that, when this condition is met, the most recent observation in the $AR(1)$ model can be written as the weighted average of the past residuals by applying recursive substitution

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$
$$X_t = \phi_1 \left(X_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t$$
$$X_t = \phi_1^2 X_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t$$
$$X_t = \phi_1^3 X_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t$$
$$\cdots$$
$$X_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j} < \infty$$

since ϕ_1^j decreases as j increases.

The general case of the model is the $AR(p)$, the autoregressive model of order p, and it is is defined as satisfying the following equation

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \text{ where } \varepsilon_t \sim WN(0,1)$$
$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) X_t = \varepsilon_t$$
$$\Phi(B) X_t = \varepsilon_t$$

where B is the *Backshift operator*, which simplifies operations such as

$$(1 - B^m) X_t = X_t - B^m X_t = X_t - X_{t-m}$$

The stationarity condition of this model is met when all of the roots of the corresponding $AR(p)$ characteristic polynomial

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$$

strictly lie outside the unit circle, i.e., if $|z_i| > 1, \forall i = 1, 2, ..., p$.

Moving Average models

As opposed to autoregressive models, the variable of interest in moving average (MA) models is described by the linear combination of the past residuals. The $MA(1)$ is written as follows

$$X_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t, \text{ where } \varepsilon_t \sim WN(0,1)$$

Similarly to the autoregressive of order 1 case, when $|\theta_1| < 1$, it is possible to write the most recent residual as the weighted average of the past observations in the $MA(1)$ process by applying recursive substitution

$$\varepsilon_t = X_t - \theta_1 \varepsilon_{t-1}$$
$$\varepsilon_t = X_t - \theta_1 (X_{t-1} - \theta_1 \varepsilon_{t-2})$$
$$\varepsilon_t = (-\theta_1)^2 \varepsilon_{t-2} - \theta_1 X_{t-1} + X_t$$
$$\varepsilon_t = (-\theta_1)^3 \varepsilon_{t-3} + (-\theta_1)^2 X_{t-2} - \theta_1 X_{t-1} + X_t$$
$$\cdots$$
$$\varepsilon_t = \sum_{j=0}^{\infty} (-\theta_1)^j X_{t-j} < \infty$$

since it is generated a converging geometric series.

It is possible to see that the $MA(1)$ model can be rewritten as an $AR(\infty)$ model when $|\theta_1| < 1$, as well as the $AR(1)$ model in the subsection below was rewritten as an $MA(\infty)$ when the condition $|\phi_1| < 1$ verifies.

In those cases, where $AR(p)$ and $MA(q)$ models can be rewritten as the infinite form of their counterparts, they are called *stationary* and *invertible*, respectively, and the conditions $|\phi_1| < 1$ and $|\theta_1| < 1$ are the stationarity and the invertibility conditions for the $AR(1)$ and the $MA(1)$ respectively.

The general case of the moving average model is the $MA(q)$, and it is written as

$$X_t = \theta_1 \varepsilon_{t-1} + \theta_1^2 \varepsilon_{t-2} + \cdots + \theta_1^q \varepsilon_{t-q} + \varepsilon_t, \text{ where } \varepsilon_t \sim WN(0,1)$$
$$X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) \varepsilon_t$$
$$X_t = \Theta(B) \varepsilon_t$$

Similarly to the stationarity condition of the $AR(p)$ model, the invertibility condition of this model is met when all of the roots of the corresponding $MA(q)$ characteristic polynomial

$$\Theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$$

strictly lie outside the unit circle, i.e., if $|z_i| > 1, \forall i = 1, 2, ..., q$.

(S)ARIMA models

The Autoregressive Integrated Moving Average (ARIMA) models combine autoregressive components with moving average ones, as well as adding a time lag differencing to the final model.

Differencing is an important concept when modelling ARIMA models, as it makes non-stationary time series become stationary. When differencing is applied, recurring patterns such as seasonality, as well as trends in the series can be filtered out, which improves forecasting accuracy.

When a given time series X_t is not stationary, but its first difference $(1 - B)X_t$ is, then the series is said to be integrated of order one and can be represented by $X_t \sim I(1)$. Similarly, when X_t needs to be differenced twice in order to be stationary, then $X_t \sim I(2)$, and when X_t is s stationary series, then $X_t \sim I(0)$, (Hanck, Arnold, Gerber, & Schmelzer, 2021).

The general case of the Box-Jenkins model is the $ARIMA(p, d, q)$, which can be represented by the following equation using the backshift operator

$$(1 - B)^d X_t = \phi_1 (1 - B)^d X_{t-1} + \cdots + \phi_p (1 - B)^d X_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

or by

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d X_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t$$

where $\varepsilon_t \sim WN(0, 1)$, and p, d and q are the orders of the autoregressive, the differencing and the moving average components, respectively.

The ARIMA models also support the modelling of seasonal patterns, adding seasonal components to the autoregressive and/or to the moving average part of the model. These models are, in the general case, the $SARIMA(p, d, q) \times (P, D, Q)_s$.

Considering monthly data, one possible version of these models is the $SARIMA(1,1,1) \times (1,1,1)_{12}$, represented by the following equation

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(1-B)(1-B^{12})X_t = (1+\theta_1 B)(1+\Theta_1 B^{12})\varepsilon_t$$

where Φ_1 and Θ_1 are the coefficients related to the autoregressive and moving average parts of the seasonal component, respectively.

It is important to note that, both for the seasonal and the non-seasonal components, the autoregressive and the moving average parts of the SARIMA models can be stationary or invertible, respectively, as long as the stationarity and the invertibility conditions in accordance to the order of each part (p and q) are verified.

Augmented Dickey-Fuller unit root test: URCA package

There are several methods to test a time series for stationarity (unit roots). One of the most popular is the Augmented Dickey-Fuller test (ADF test), developed by Said and Dickey (1984). The method used in function $ur.df(type = \text{"}none\text{"}, lags = p)$ of the R package *urca* (Pfaff, 2008) is based on the linear model without trend and intercept

$$\Delta X_t = \gamma X_{t-1} + \delta_1 \Delta X_{t-1} + \delta_2 \Delta X_{t-2} + \cdots + \delta_p \Delta X_{t-p} + \varepsilon_t$$

where $\Delta X_t = X_t - X_{t-1}$.

It is possible to run the test using any time series X_t that can be described by an $ARMA(p,q)$ structure, regardless of the orders. The hypotheses of the test are

$$H_0 : \gamma = 0 \Leftrightarrow H_0 : X_t \sim I(1)$$

$$H_1 : \gamma < 0 \Leftrightarrow H_1 : X_t \sim I(0)$$

and the test statistic is $ADF = \hat{\gamma}/SE(\hat{\gamma})$, which follows a non-standard t-student distribution and has special critical values that can be found in the paper published by Said and Dickey (1984).

AUTO ARIMA

When working with many different series, most of the times, it is not possible to assess the behaviour of each one of them and decide a specific model for them, considering time restrictions.

In that case, it comes in handy the use of functions that find suitable models for each of them, such as the *auto.arima* function from R the package called *forecast* (R. Hyndman et al., 2023), developed by R. J. Hyndman and Khandakar (2008).

The algorithm determines the order of the best ARIMA model, using information criteria (AIC or BIC) and appropriate unit roots tests. Considering seasonal data, the algorithm, before searching for the order of the AR and MA parts of the model, looks for an appropriate order of differencing by first finding out D using an extended Canova-Hansen test, which will return either $D = 0$ or $D = 1$. Following that, the algorithm searches for d, starting at $d = 0$, by running KPSS unit-root tests sequentially until it returns an insignificant test result and assigning the respective d order to the model.

After differencing the series, the algorithm finds the most suitable parameters p, q, P and Q for the $SARIMA(p, d, q) \times (P, D, Q)_s$ model by minimising the information criterion

$$AIC = -2log(L) + 2(p + q + P + Q + k)$$

where L is the maximised likelihood of the model fitted to the differenced series represented by $(1 - B^m)^D(1 - B)^d X_t$. It is important to note that, in case the series is not seasonal, the Canova-Hansen test is skipped, R. J. Hyndman and Khandakar (2008).

3.1.6 The Fractional Airline Model

When working with data sampled at frequencies that do not have an integer periodicity, such as weekly data, which has $365.25/7 \approx 52.18$ weeks in a year, it is necessary to use alternative time series decomposition methods in order do seasonally adjust the series, since the classical and other commonly used methods can not handle such data.

The Fractional Airline Model (FAM) is a decomposition method that can handle data with non-integer periodicity. The model is a fractional variant of the airline model $SARIMA(0, 1, 1) \times (0, 1, 1)_s$ that was adapted to fractional periodicities using the following first-order Taylor series expansion (Ollech & Bundesbank, 2023)

$$\tilde{\nabla}_s X_t \approx \ X_t - (1 - \alpha)B^{\lfloor s \rfloor} X_t - \alpha B^{\lfloor s \rfloor + 1} X_t$$

where $\tilde{\nabla}_s$ is the fractional differencing operator, and $s = \lfloor s \rfloor + \alpha$ is the fractional seasonal period, where $\alpha \in [0, 1[$.

As in Burman (1980), a canonical decomposition is performed, and the components are estimated by means of the Kalman smoothers, as explained by Evans, Monsell, and Sverchkov (n.d.).

Considering a weekly time series, the final model can be rewritten as

$$(1 - B)(1 - 0.82B^{52} - 0.18B^{53})X_t = (1 - \theta_1 B)(1 - 0.82\Theta_1 B^{52})(1 - 0.18\Theta_1 B^{53})\varepsilon_t$$

3.1.7 Hierarchical Time Series

One of the possible ways to organise time series data, is aggregating the data at levels that are based on features, such as geographic location, organisational structure, etc. These are called *hierarchical time series*, and as an example, the figure below illustrates how the hierarchy of the CAE 47 (retail trade) is organised.

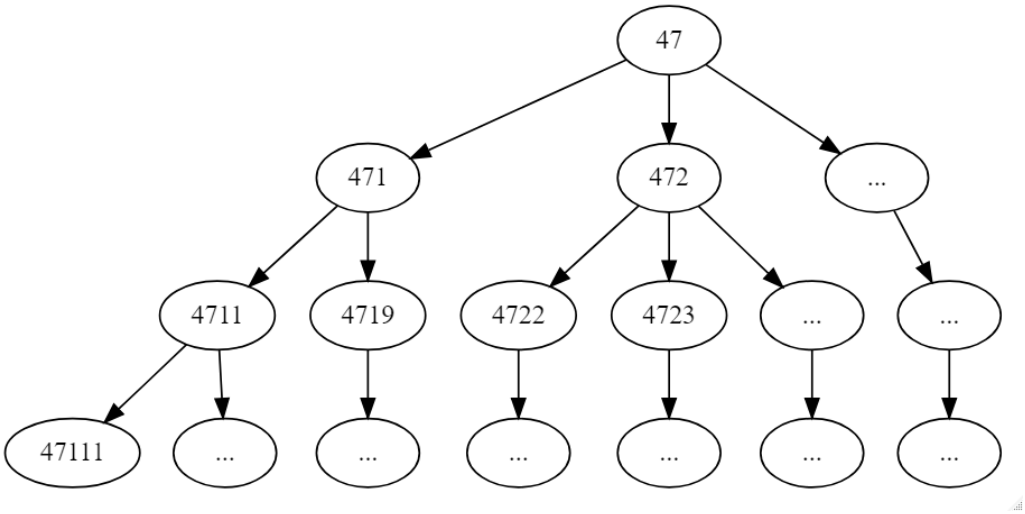

Figure 3.1: CAE 47 hierarchy

There are multiple applications in business and economics that require the forecasting of different time series that are related through a hierarchical structure. Usually, the forecasting

methods for these series use either *top-down* or *bottom-up* approaches, (R. J. Hyndman, Ahmed, Athanasopoulos, & Shang, 2011).

Top-down approach

The top-down approach to forecasting a hierarchical time series consists in forecasting the aggregated data at the highest level, then decomposing the forecast into the lower levels of the hierarchy.

There are several methods to decompose the forecast to lower hierarchies. The usual way is to compute the contribution of each lower level hierarchy to the aggregated data and disaggregate the forecast accordingly. These contributions can be estimated by using historical data or forecasted.

When using historical data to estimate the proportions, two possible methods excelled in a study conducted by Gross and Sohl (1990):

- Average historical proportions: $p_j = T^{-1} \sum_{t=1}^{T} X_{tj}/X_t$
- Proportions of the historical average: $p_j = \sum_{t=1}^{T} X_{tj}T^{-1} / \sum_{t=1}^{T} X_t T^{-1}$

where p_j is the proportion of the j-th bottom-level hierarchy and X_{tj} is the observation of the j-th bottom-level hierarchy at time t, with $t = 1, 2, ..., T$.

Although these two methods are widely used nowadays, the historical data approach do not take into account the changes over time of these proportions, which can lead to less accurate predictions at the bottom level hierarchies. An alternative approach was proposed by Athanasopoulos, Ahmed, and Hyndman (2009), which estimates proportions based on forecasts rather than with historical data.

Bottom-up approach

Alternatively, hierarchical time series can also be forecasted using the bottom-up approach. This method consists in building models for bottom-level hierarchies and summing the forecasts up to produce an overall forecast for the aggregated, top-level data.

Using the CAE hierarchy as an example to illustrate this approach, which can be seen in the Figure 3.1, it is possible to forecast the class 4711 by forecasting its subclasses (47111 and 47112)

and summing them

$$\hat{X}_{4711,t}\,(m) = \hat{X}_{47111,t}\,(m) + \hat{X}_{47112,t}\,(m)$$

where $\hat{X}_{j,t}(m)$ is the m-step ahead forecast of the hierarchy j at time t.

Although the main advantages of this approach is yielding more accurate predictions at lower-level hierarchies and the fact that there is no loss of information due to data aggregation, as stated by R. J. Hyndman and Athanasopoulos (2018), it may be more challenging to model those lower-level hierarchies as the data at their level tend to be noisy.

3.2 Regression models

Regression models are statistical techniques used to assess and describe the relationship between a response variable Y and one or more explanatory variables $X_1, X_2, ..., X_p$. The main goals of these models are to explain the relationship between these variables and predict values of the response variable using the explanatory variables values based on their relationship.

In this project, three types of regression will be used: Simple Linear Regression, Dynamic Regression and MIDAS Regression. In the next sections, these models will be explained further.

3.2.1 Linear regression

The general model of a multiple linear regression model can be written as

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon_t$$

where $\beta_0, \beta_1, ..., \beta_p$ are the coefficients that measure the marginal effect of each explanatory variable considering that all the other variables have already been taken into account, and ε_t is a residual variable which captures deviations from the straight line model or unexplained variations in the response variable.

The reliability of the model depends upon the following assumptions:

1. $\boldsymbol{E}(\varepsilon_t) = 0, \forall t$
2. $\boldsymbol{V}(\varepsilon_t) = \sigma^2, \forall t : 0 < \sigma^2 < \infty$

3. $\boldsymbol{cov}(\varepsilon_t, \varepsilon_{t+s}) = 0, \forall t, s : s \in \{\pm 1, \pm 2, \ldots\}$
4. $\boldsymbol{cov}(\varepsilon_t, X_{jt}) = 0, \forall j = 1, 2, ..., p$

When these assumptions are violated, the accuracy of the estimates are affected and the statistical inferences are not valid.

One of the possible ways to estimate the coefficients $\beta_0, \beta_1, ..., \beta_p$ is using the Ordinary Least Squares (OLS) estimation, which minimises the sum of the squared errors. In order words, the coefficient values are chosen by minimising the sum

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (Y_t - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_p X_p)^2$$

and then, obtaining the estimated values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$.

3.2.2 Dynamic regression

When working with time series, there are underlying time dynamics that the classic linear regression cannot capture. Besides, when modelling time series data with linear regression, it is possible that the residual term is serially correlated, which violates one of the hypothesis of the model.

In order to allow the residual term to be autocorrelated and make it possible to fit a linear regression using time series data, the following model was designed

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \eta_t$$

$$\eta_t \sim ARMA(p, d, q)$$

where the model η_t replaces the error term ε_t in the classical linear regression model Y_t, which results in 2 errors inside the model: the error η_t of the regression Y_t, and the error ε_t of the ARMA model η_t.

When the model is estimated, the minimisation should be on the error term ε_t, and not on the η_t, otherwise the associated statistical inference is not valid, the estimated coefficients $\hat{\beta}_p$ are not the best ones, and the AICc values are not good benchmarks for the models.

Furthermore, in order to use the ARMA errors, all of the variables in the model must be stationary, otherwise the estimators will not be consistent. The only exception for that is when the variables in the model are cointegrated, i.e., when there exists a linear combination of these variables that is stationary.

Cointegration

Although the relationship between two or more non-stationary time series may not be directly causal or clear and may deviate from one another in the short-term, they can have a stable equilibrium that can be observed in the long-run - such series are said to be cointegrated.

Formally, two non-stationary time series represented by the variables Y_t and X_t are cointegrated if there is a θ such that the linear combination $Y_t - \theta X_t$ is stationary, which means that Y_t and X_t have a common stochastic trend that can be cancelled out by the said linear combination, Hanck et al. (2021).

It is possible to test, for instance, the variables Y_t and X_t for cointegration by following the framework developed by Engle and Granger (1987), which consists of:

1. Estimating the regression $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$ using OLS
2. Pulling the residual series from the estimated model
3. Testing the residual series for stationarity using the ADF test

If the non-stationarity of the residual series is ruled out by the test for a given significance level, then the statistical evidence is that the two series are cointegrated.

It is important to highlight that, in this approach, the critical values for the ADF test are different than the ones presented by Said and Dickey (1984), as the distribution now depends on the number of series that are being tested. These special critical values were published by Engle and Granger (1987), and since the case studied in this work only uses two series, the relevant critical value, considering a significance level of $\alpha = 5\%$, is -3.41.

3.2.3 MIDAS regression

The Mixed data sampling (MIDAS), proposed by Ghysels, Santa-Clara, and Valkanov (2004), is a regression model that was developed to tackle a problem often encountered in the economics field, where the variable of interest has lower frequency than the relevant information at hand. Thus, the main feature of this model is the linkage of a lower-frequency dependent variable to higher-frequency independent variables, eventually sampled at different time intervals.

Supposing the dependent variable of interest Y sampled at a given fixed frequency, and the independent variable X^m sampled at a frequency m times faster, the simple MIDAS regression equation is represented by

$$Y_t = \beta_0 + \beta_1 b(B^{1/m}; \theta) X_t^m + \varepsilon_t^m$$

where $b(B^{1/m}; \theta)$ is a lagged polynomial of length K, $b(B^{1/m}; \theta) = \sum_{k=0}^{K} b(k; \theta) B^{k/m}$ and $B^{k/m} X_t = X_{t-k/m}$, with B the Backshift operator.

In a later study, Ghysels et al. (2007) addressed the issue of dealing with a very high number of parameters to estimate in the polynomial $b(B^{1/m}; \theta)$, caused by high-frequency explanatory variables when using a significant number of lags of X_t^m, which is not a challenge uncommonly faced. In order to tackle this problem, it was proposed to use the exponential almon lag polynomial, represented by

$$b(k; \theta) = \frac{e^{\theta_1 k + \cdots + \theta_Q k^Q}}{\sum_{k=1}^{K} e^{\theta_1 k + \cdots + \theta_Q k^Q}},$$

which has been shown to be useful by Armesto, Engemann, Owyang, et al. (2010) and Clements and Galvão (2008).

Chapter 4

Methodology

This chapter aims to describe the methodological framework used to address the research problem explained in the Chapter 1, which is finding the most accurate models to nowcast the e-Fatura data for a given reference month in order to comply with the deadline to the publication of the Retail Traide Turnover Index, imposed by Eurostat, in case the Tax Authority fails to deliver the e-Fatura data in time.

4.1 Data description

The Multibanco data and the e-Fatura data are provided by different entities and have different frequencies. The former is provided by SIBS at both weekly and monthly frequencies, whereas the *e-Fatura* data is monthly and provided by the Tax Authority. Also, the structure of both data follows the CAE framework.

As mentioned earlier, the CAE classes and subclasses are organised in hierarchies. For the retail trade activity, the CAE code starts with the 2-digit number 47, and it can be disaggregated up to 5 digits, yielding hierarchies of the form $47x$, $47xx$ and $47xxx$, as the Figure 3.1 already shows.

It is important to note that some of the data may overlap, as there may be transactions carried out using the Multibanco system that correspond to specific VAT entries, as well as the other way around. For instance, a single transaction carried out in the Multibanco system may correspond

to a couple of invoices or more in the VAT data, and a single invoice may correspond to one transaction or more in the Multibanco system.

Considering the nature of these data, it is easy to recognise the potential they have for being used for statistical purposes, as well as other types of administrative data, which is a topic that has been getting a lot of attention under the scope of the work of the National Statistical Institutes in the last few years.

4.1.1 e-Fatura data

As mentioned before, the frequency of the e-fatura data, also referenced as VAT data in this dissertation, is monthly and is sent out by the Tax Authority (AT) to Statistics Portugal (INE) between the third and the fourth week following the end of the reference month.

However, the raw dataset delivered every month needs to be processed, as there are problems that affect the quality of the data and may impair the statistical operation, such as missing values, outliers and negative values.

When it comes to the missing values, they can be either *fully missing* or *partially missing*. Fully missing values are the ones for which there are no invoice entries for an issuing entity in that month, whereas partially missing values are the ones for which the sum of the entries for an issuing entity in a given month is far below the expected.

With regards to negative values, they appear in the dataset every month. Usually, these negative entries are credit notes, booking and order cancellations, and some of them also aim to correct positive VAT entries that were registered incorrectly. At times, these negative values are of such great magnitude that they not only distort the total VAT of a given entity, but also distort the total VAT of the whole CAE subclass that the entity belongs to.

The dataset is processed by Statistics Portugal and then uploaded to the Institution's database. The final dataset that can be found in the database has six columns corresponding to the following six variables

1. YEAR: Year of the entry

2. MONTH: Month of the entry

3. **ISSUING_CAE:** CAE subclass of the entry
4. **N_ISSUERS:** Number of entities that issued invoices corresponding the VT_TOTAL of the entry
5. **N_ENTRIES:** Number of invoices issued by the entities of the entry
6. **VT_TOTAL:** Total value of the invoices issued of the entry, in euros

4.1.2 Multibanco data

The multibanco data is delivered monthly by SIBS in both weekly and monthly frequencies. The dataset has seven columns, corresponding to thge following seven variables

1. **YEAR:** Year of the entry
2. **MONTH:** Month of the entry
3. **WEEK:** Week of the year of the entry - weekly dataset only
4. **CAE_CODE:** CAE subclass of the entry
5. **CAE_DESCRIP:** Description of the activity of the CAE subclass
6. **TYPE_OPERATION:** Type of operation, which can be: *Purchases, Other operations, Payment of purchases/services* and *Special services*.
7. **TOTAL:** Total transactioned

The Multibanco data does not require pre-processing as it is delivered with high quality by SIBS and ready to use.

4.2 Methodology

As stated in the previous Chapters, this work will be undertaken in order to understand the behaviour of the e-Fatura data and find the most appropriated method for nowcasting its value for each CAE subclass for the most recent reference period.

Statistics Portugal currently uses the e-Fatura data as a complementary variable to estimate the *Business Turnover Index in Retail Trade*. However, at times, the Tax Authority does not deliver the data to the institution in time, which poses a threat to the compliance of the deadlines for releasing the Short-term business statistics reports imposed by EUROSTAT.

Thus, the current goal is to build a framework for nowcasting the e-Fatura data in case the Tax Authority fails to deliver the data in time again. The proposed approach is:

- Nowcast the e-Fatura data for the most recent reference period at the 4-digit level of the retail trade classification
  - Using the e-Fatura historical data
  - Using the Multibanco data as the predictor variable, sampled either monthly or weekly
- As an alternative, nowcast the e-Fatura at the 5-digit level (for CAEs that have 5-digit disagreggation) using the two options given above, and then aggregate the results back to the 4-digit level using the hierarchical bottom-up approach

In this project, the target hierarchy is the 4-digit one, and the nowcasting exercise will be focused on this hierarchy level. Although the 5-digit level will also be nowcasted, it will only be done with the purpose of aggregating the information back to the 4-digit level, in order to compare the performance of targeting the 4-digit directly versus using a bottom-up hierarchical approach.

It is important to note that, except for the MIDAS regression, which uses the weekly Multibanco data as explanatory variable to nowcast the e-Fatura data at monthly frequency, as it aims to link lower-frequency variables to higher-frequency ones, all of the other models used the Multibanco data sampled monthly, since they only support same-frequency variables.

4.2.1 Tools

The software used were Oracle SQL Developer to retrieve the datasets that will be used from the Institute's database, and the statistical programming language R with the following packages: *midasr* (Ghysels, Kvedaras, & Zemlys, 2016), *tsibble* (Wang, Cook, & Hyndman, 2020), *fable* (O'Hara-Wild, Hyndman, & Wang, 2023), *dplyr* (Wickham, François, Henry, Müller, & Vaughan, 2023), *urca* (Pfaff, 2008), *xts* (Ryan & Ulrich, 2023), *forecast* (R. Hyndman et al., 2023), *fpp3* (R. Hyndman, 2023) and *ggplot2* (Wickham, 2016).

The package *midasr* has the functions to apply the MIDAS regression, the packages *tsibble,*

dplyr and *fable* were used to manipulate and create time series objects. Regarding the package *urca*, it was used to carry-out the stationarity tests, and *forecast* was used to access the function $auto.arima()$. Finally, the plots were made using the package *ggplot2*.

Chapter 5

Results

In this Chapter, the results obtained from applying the methodology described above are shown and discussed. At first, the exploratory analysis of the e-Fatura and the Multibanco data are presented, although in a limited manner - the values will not be disclosed - due to confidentiality issues. Then, the results of nowcasting the e-Fatura data will be discussed.

The algorithms used to nowcast the e-Fatura data were ARIMA, Linear regression, Dynamic regression and MIDAS - an additional forecast will also be provided, as the combination of the forecasts made by all models using arithmetic mean. Furthermore, it is important to highlight that the ARIMA models were fitted using the function $auto.arima()$ from the R package *forecast*, which was briefly presented in Chapter 3.

Moreover, in Section 5.1 up to Section 5.5, the analysis and the results discussed regard only the aggregated 2-digit and the 4-digit level CAEs, as it would be redundant to also analyse the 5-digit disaggregation. However, the results obtained using the bottom-up approach from the 5-digit to the 4-digit level will be discussed in the Section 5.6 and compared to the results obtained directly from the 4-digit level.

5.1 Exploratory analysis

5.1.1 e-Fatura

The e-Fatura data, dating back to January 2016, and aggregated to the highest hierarchy, representing the whole Retail trade sector, is represented in Figure 5.1. The series clearly shows an upward trend and an underlying yearly seasonal pattern.

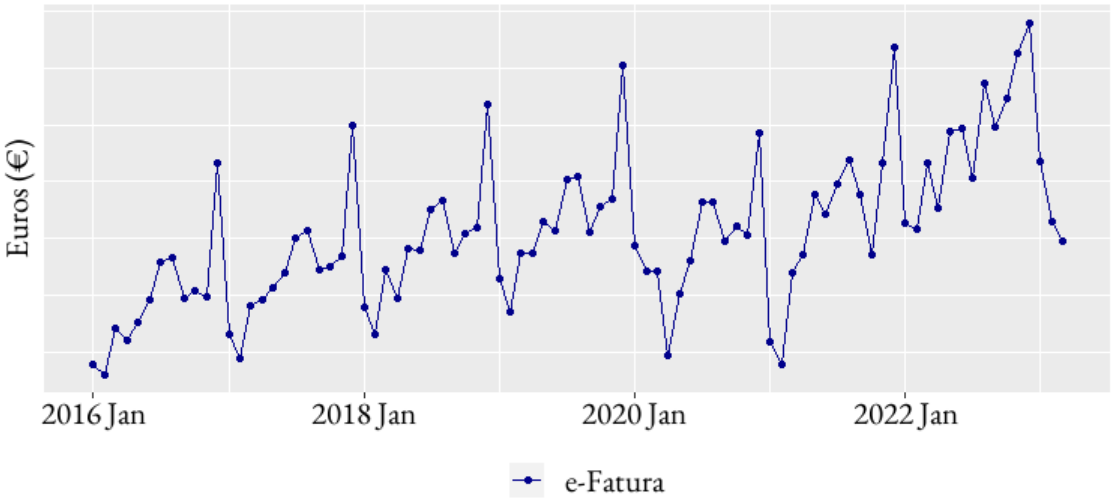

Figure 5.1: e-Fatura time series: CAE 47

In the Figure 5.1, it is possible to see that, until 2019, the observations had a more stable pattern across the years, and from 2020 onwards, they have become more noisy. Regardless of that, the seasonal behaviour remained very similar, as the seasonal plot in Figure 5.2 shows.

In general, as the Figure 5.2 shows, the month with the highest observed value across the years is December, followed by August and the month with the lowest is February - this is true for every observed year, except for 2021, when July had a higher observation than August, and 2023, when March had a lower observation than February.

When the series is broken down to the 4-digit hierarchy, it yields 37 different series that correspond to each of the 4-digit CAE categories. However, the average weight of the categories on the aggregated series is not balanced as Table 5.1 illustrates. For instance, considering monthly observations, the CAEs 4711 and 4730, on average, account for 38.41% and 11.16% and are the

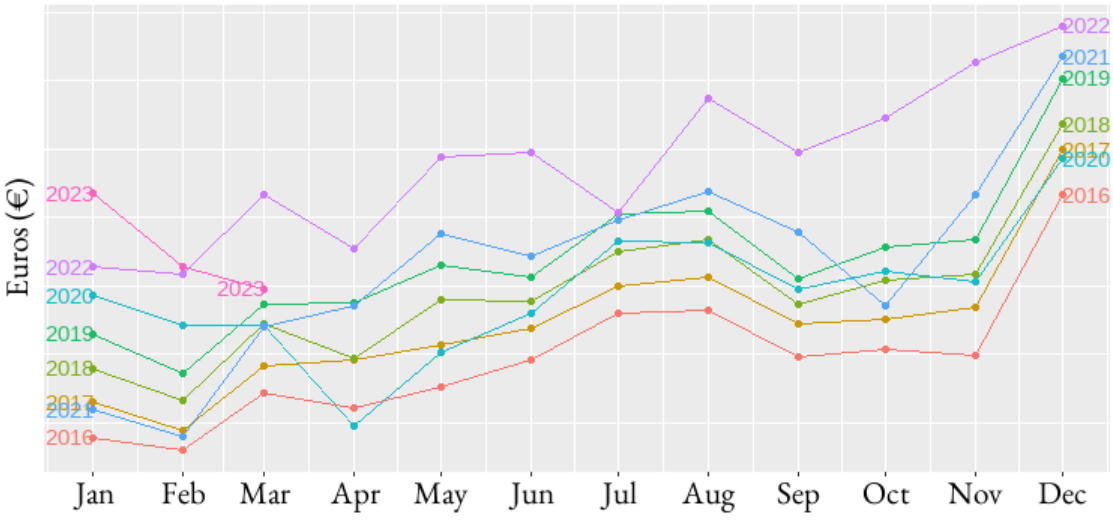

Figure 5.2: e-Fatura seasonal plot: CAE 47

categories with highest weights, whereas the CAEs 4763 and 4782 account for, on average, only 0.007% and 0.05% and are the ones with lowest weights. Out of the 37 CAEs, on average, 23 account for less than 1% each, 6 account between 1% and 3% each, and 6 account between 7% and 3% each. For the complete set of average weights see Table A.1.

CAE	4711	4730	4773	4771	4778	4752	...	4789	4782	4763
Avg. Prop. (%)	38.41	11.16	6.97	6.07	4.95	4.93	...	0.1	0.05	0.01

Table 5.1: 4-digit CAEs weight on aggregate: e-Fatura (preview table)

Among those 37 subseries, many of them have similar behaviour with one another and also with the aggregated series, however, a few of them seem to have a more unique behaviour as illustrated in Figure 5.3

This shows that some retail economic activities, such as supermarkets, when compared to the generality of the retail sector (aggregated series), have very similar fluctuations throughout the year, whereas other, like flower shops, do not.

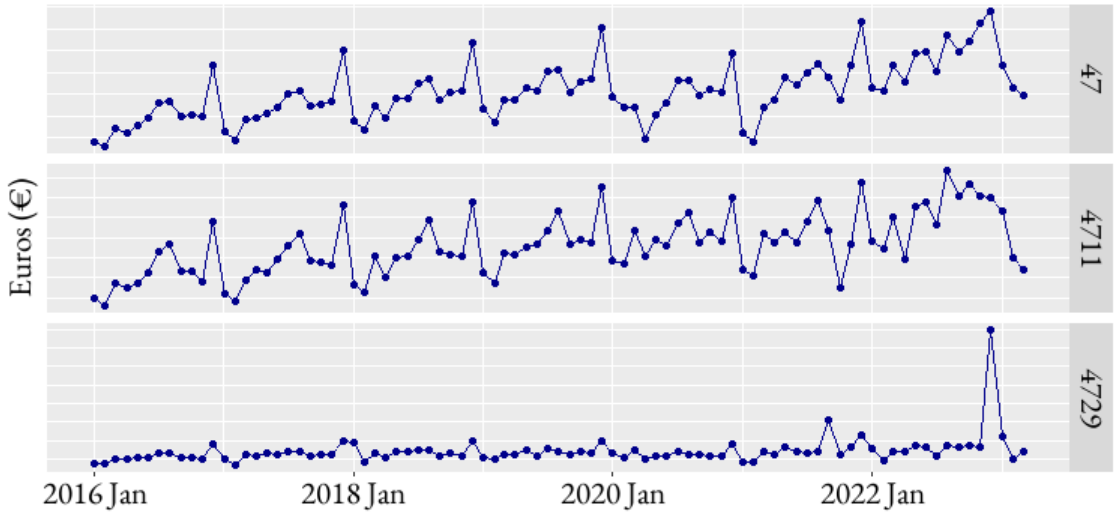

Figure 5.3: e-Fatura time series: CAEs 47, 4711 and 4729

5.1.2 Multibanco

The Multibanco series dates back to January 2020 and is organised in CAE hierarchies just like the e-Fatura series. The aggregated series is represented in Figure 5.4. This time series also shows an upward trend and yearly seasonality.

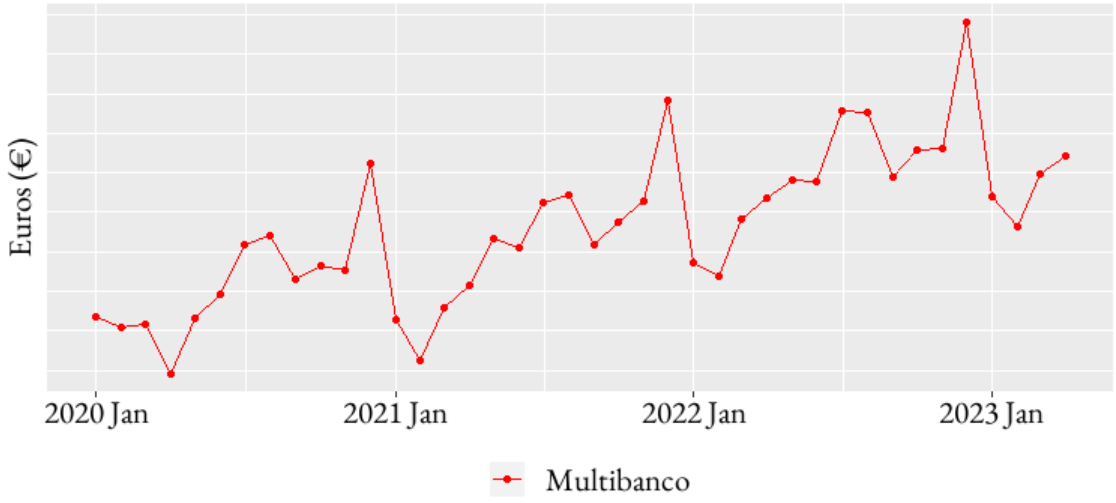

Figure 5.4: Multibanco time series: CAE 47

Moreover, the series seems to have a more stable and predictable behaviour across the years,

unlike the e-Fatura series, which is more noisy from 2020 onwards.

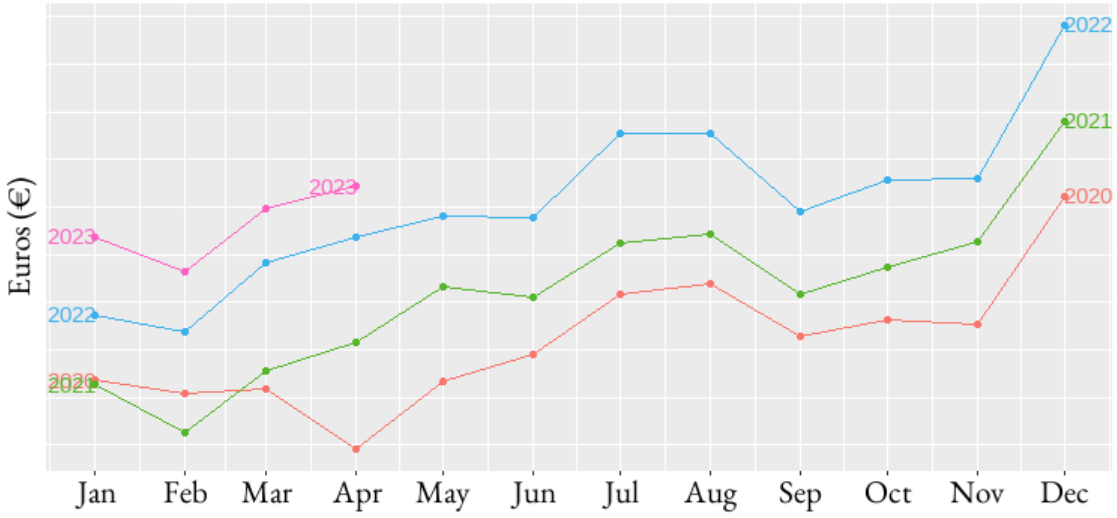

Figure 5.5: Multibanco seasonal plot: CAE 47

The seasonality of the series, shown in the Figure 5.5, is very similar to the one present in the e-Fatura data, with the highest and lowest values observed, in most of the years, in December and in February, respectively.

When the Multibanco series is broken down to its 37 4-digit CAEs, there is a great imbalance on the average weight that each of them accounts for the aggregated data as shown in the Table 5.2 (full set of weights in Table A.2), similarly to the imbalance in the e-Fatura data (Table A.1). The CAEs with the highest weights are the 4711 and the 4730, which account for 48.78% and 10.01%, and the ones with the lowest weights are the 4763 and the 4782, accounting for 0.01% and 0.03%, respectively. Regarding the overall distribution of the weights, 25 CAEs account for less than 1% each, 4 CAEs account between 1% and 3% each and 6 CAEs account between 7% and 3% each.

CAE	4711	4730	4771	4773	4752	4778	...	4789	4782	4763
Avg. Prop. (%)	48.78	10.01	6.15	4.85	4.83	4.12	...	0.04	0.03	0.01

Table 5.2: 4-digit CAEs weight on aggregate: Multibanco (preview table)

When we look at the 37 subseries time plots, similarly to what happens with the e-Fatura data, it is possible to see that many subseries behave similarly to the aggregated series, but some do not. However, the CAEs that behave differently and similarly to the aggregated CAE 47 in the Multibanco series are not all the same ones in the e-Fatura dataset, although many of them are. For instance, it is possible to see in the Figure 5.6 that the behaviour of the CAE 4789 is very similar to the overall retail sector, whereas the CAE 4753 behaves differently.

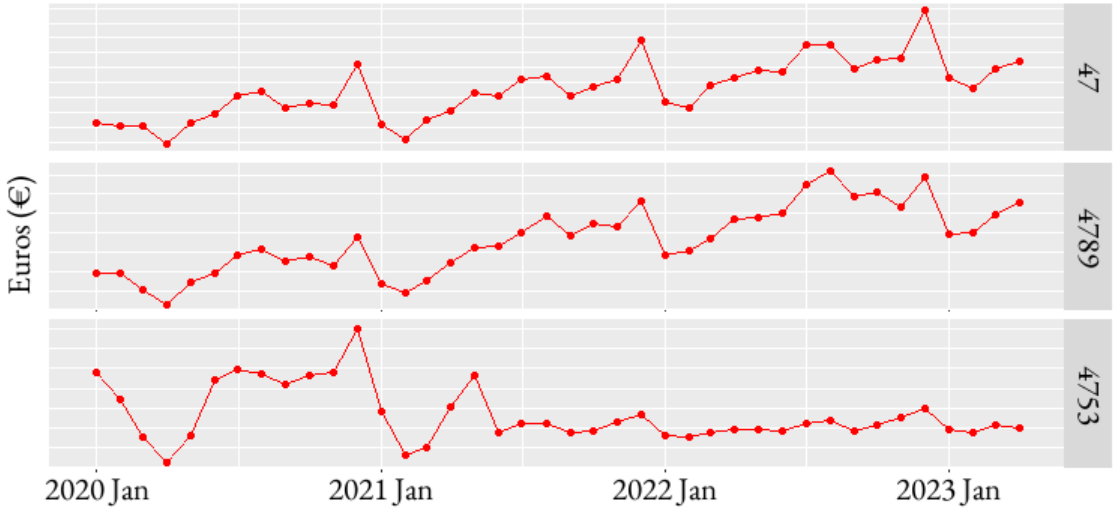

Figure 5.6: Multibanco time series: CAEs 47, 4753 and 4789

5.1.3 Relationship between the Multibanco and the e-Fatura series

Considering that the e-Fatura data are the transactions' invoices issued by the seller businesses, and that the Multibanco data are the bank card transactions with those establishments, one can intuitively guess that there is a relationship between these two series.

As the Figure 5.7 shows, regarding the 2-digit level, both series have an upward trend, and it seems that they have similar underlying seasonal structure, with the highest and lowest observations in almost every year being in December and February, respectively. Moreover, their monthly fluctuations across the years are very much alike.

The scatterplot in Figure 5.8, reveals a contemporaneous linear relationship between the two series at the 2-digit level, with a correlation coefficient $r = 0.911$. It is also relevant to check if

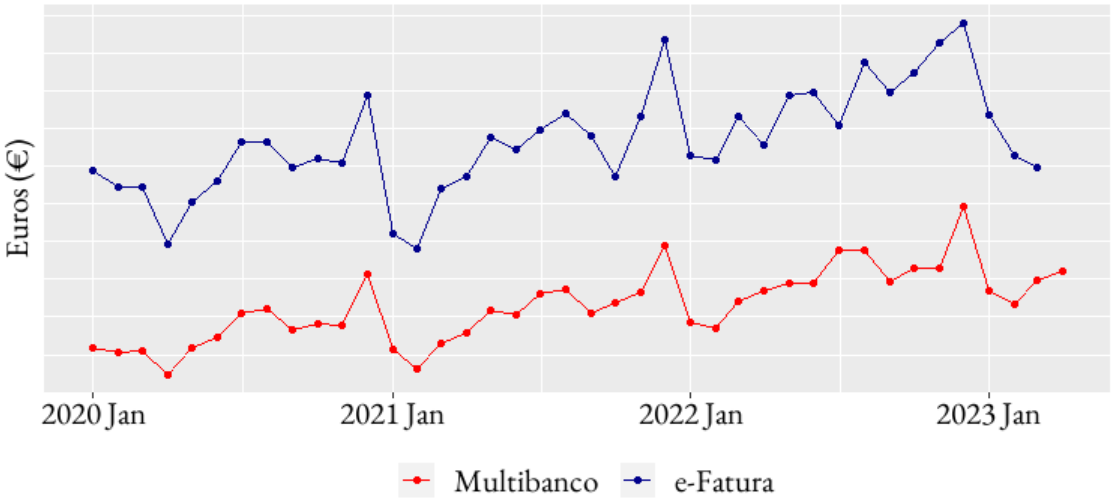

Figure 5.7: e-Fatura and Multibanco time series: CAE 47

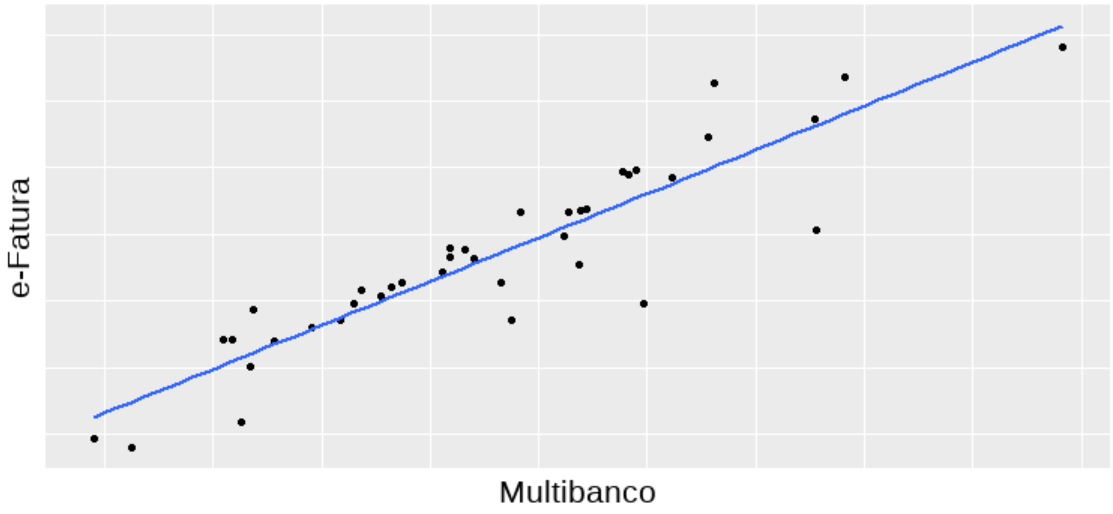

Figure 5.8: e-Fatura vs Multibanco at time t for CAE 47

these two series are cointegrated, by fitting the regression

$$eFatura_t = \beta_0 + \beta_1 Multibanco_t + \varepsilon_t$$

and checking the stationarity of the residuals.

After fitting the model, the results were

$$\hat{\beta}_0 = 1357653246, \quad SE(\hat{\beta}_0) = 262159200, \quad p - value(\hat{\beta}_0) < 0.001$$

$$\hat{\beta}_1 = 1.316, \quad SE(\hat{\beta}_1) = 0.09782086, \quad p - value(\hat{\beta}_1) < 0.001$$

$$\hat{R}^2 = 0.8303, \quad \hat{R}^2_{adj} = 0.8257, \quad SE = 292621002$$

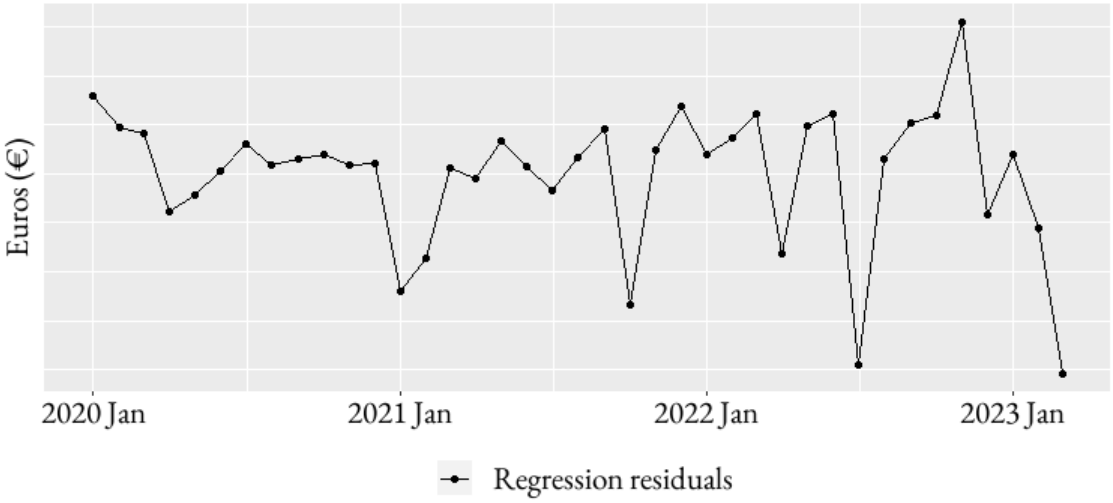

Figure 5.9: Regression residuals time series: CAE 47

After visually analysing the Figure 5.9 and running an ADF test using the $ur.df()$ function in the R package *urca*, with lag selection using AIC optimisation, the test statistic was -4.28, rejecting the null hypothesis of no cointegration considering a 5% significance level, which critical value is $CV = -3.41$. Given that, the statistical evidence is that, at the 2-digit level, these two series are cointegrated.

Even though the aggregated 2-digit series of both e-Fatura and Multibanco are cointegrated, this does not hold true for every CAE at the 4-digit level. The statistical evidence is that, out of 37 4-digit CAEs, only 9 of them are cointegrated, considering a significance level of 5%. The test statistic for each CAE can be found in the Table A.3.

The e-Fatura and the Multibanco time series of some 4-digit CAEs can be almost identical, however, there are a few ones that behave very differently, and the Figure 5.10 shows examples of these two possible situations, which poses a challenge to using the Multibanco as an auxiliary

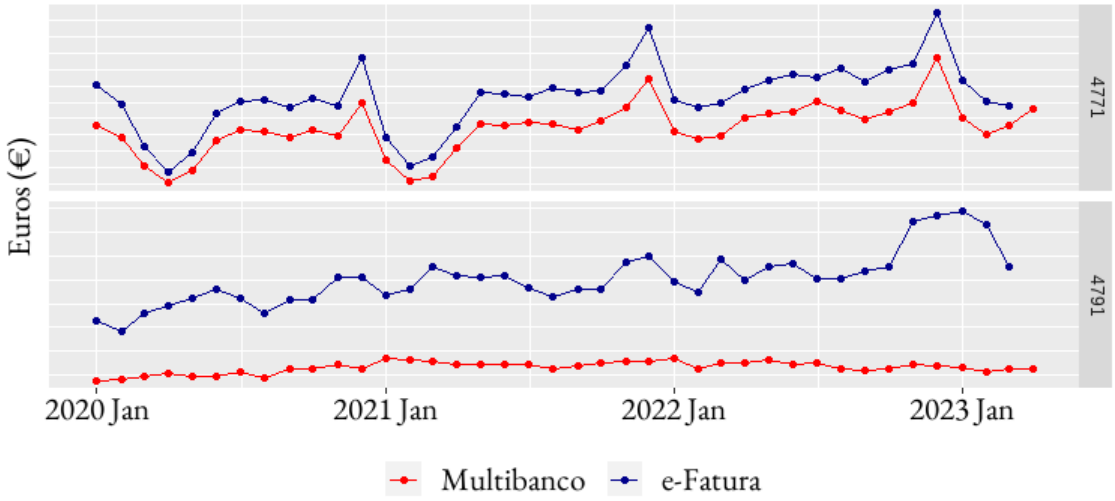

Figure 5.10: Regression residuals time series: CAE 47

variable to estimate the e-Fatura data in those CAEs.

Considering that not every 4-digit CAE has their e-Fatura and Multibanco series cointegrated and also that some of them behave differently in the short-term, it was decided to use ARIMA models to forecast the e-Fatura using historical data, and also to use two different linear regression models, which will be explained more in depth in the Sections 5.2 and 5.5.

5.2 Linear regression

The application of the Linear regression algorithm in this work was split in two models:

1. Simple linear regression

$$eFatura_t = \beta_0 + \beta_1 Multibanco_t + \varepsilon_t \tag{5.1}$$

2. Linear regression with seasonal dummies

$$eFatura_t = \beta_0 + \beta_1 Multibanco_t + \beta_2 FEB_t + \beta_3 MAR_t + \cdots + \beta_{12} DEC_t + \varepsilon_t \tag{5.2}$$

The reason for this approach is that, as shown in the exploratory analysis in the Subsection 5.1.3, some CAEs have their Multibanco subseries behaving almost identically to their counterpart e-Fatura subseries, including their seasonal fluctuation, whereas a few other CAEs seem to

have seasonal fluctuations that are out of sync. Thus, it was decided to use to these two regression models in order to assess their performance in each of these cases mentioned. The results of the models (5.1) and (5.1) can be found in the Tables A.4 and A.5 respectivelly.

Considering the results obtained from data at the 4-digit level only, as the Table A.6 shows, out of 37 CAEs, 14 were predicted with more accuracy when using the model (5.1), and 16 were better predicted using the model (5.2). For the remaining 7 CAEs, the performance of both models was very similar - the RMSE of the model (5.1) was neither 5% higher nor 5% lower than the RMSE of the model (5.2). For clarification purposes, the Table A.6 shows the ratio between the RMSE of both models, which can be useful for this direct comparison between the performance of these models.

Lastly, regarding the model (5.1), when we analyse the $\hat{R}^2_{adj}$ of all fitted models for every CAE, only 18 out of 37 scored high, having $\hat{R}^2_{adj} > 0.7$, which is shown in the Table A.7. This is a result of the fact that the behaviours of the e-Fatura and the Multibanco data are not consistently similar for every CAE.

5.3 Dynamic regression

Although the Dynamic regression models have a linear regression in their structure (see Subsection 3.2.2), it was decided not to train two different models for every CAE as it was done in the Section 5.2. This decision was made because this model has an ARIMA component, which captures the time-dependant dynamics in the data.

The models were trained automatically using the function $MODEL(ARIMA(Y \sim X))$ from the package *fpp3* (R. Hyndman, 2023), where Y and X are the dependent and the independent variables, respectively. The performance of this model across the CAEs, measured by the RMSE, are in the Table A.8.

5.4 MIDAS

The application of the MIDAS regression uses the weekly Multibanco data as the explanatory variable of the monthly e-Fatura, which has a major drawback - months do not have an integer number of weeks, and since the algorithm does not support non-integer frequencies, some information may be lost because some days of the month may not be accounted for, and it is also possible to allocate days from either preceding or following months, or both, depending on how the indexing of the higher-frequency data is organised.

Also, the MIDAS algorithm allocates the higher-frequency observations to each lower-frequency ones from backwards. For instance, in the case of this project, the algorithm first splits monthly subsets of the weeks, and then starts allocating them to their respective month in the regression from the most recent week of that month to the oldest one, according to the set number of lags k - so a regression with $k = 2$ lags will always allocate the last 2 weeks of each month to the regression.

Considering the process described above, and in order to mitigate the drawbacks explained in the first paragraph, the day set to index each observed week in the weekly Multibanco dataset was the last day of each week. For instance, if a week covers the days from 30 January 2019 to 5 January 2020, then it was indexed as 5 January 2020. This decision may result in allocating days from the preceding month, but it prevents from allocating days from the following month.

Moreover, since the weekly Multibanco and the e-Fatura data have different seasonality structures, before estimating the regression, both were seasonally adjusted using the Fractional airline model described in the Subsection 3.1.6 before training the models. Then, the seasonal effect is added back to the prediction by adding the estimated seasonal effect for the same period in the previous year. For instance, if the seasonally adjusted e-Fatura is estimated for April 2023, the seasonal effect of April 2022 is added back in order to reach the final prediction.

After following the steps described above, MIDAS models were trained for each CAE and their performance, measured using RMSE, can be found in the Table A.9. Although it was expected that this algorithm would have a better performance because of the granularity of the weekly data compared to the less granular monthly Multibanco data used in the other methods,

it actually performed much worse than the others except for the CAE 4742, which was the only CAE that this method performed better than the others (see Tables 5.3 and 5.4).

5.5 ARIMA

As shown in the Subsection 5.1.3, the behaviour of the e-Fatura and the Multibanco series is not similar for every 4-digit CAE, which makes the use of the Multibanco series as a predictor variable inappropriate for some CAEs. Given that, the e-Fatura series for each CAE was also predicted using ARIMA models and their performance are presented in the Table A.10.

Since there are many CAEs, and the models were analysed using 1-step ahead forecasts for 6 consecutive months for each CAE, the algorithm $auto.arima()$ from the R package *forecast* was used to fit the models, otherwise it would have taken a large amount of time to build those models individually for each month. The performance of the models estimated by the algorithm, measured using RMSE, can be found in the Table A.10

5.6 Comparison of results

After comparing the performance of all of the models across the CAEs, two tables were made: the Tables 5.3 and 5.4. The Table 5.3 shows the models that performed better for each CAE that do not have a 5-digit disagreggation (28 CAEs), whereas the Table 5.4 shows the models that performed better for the CAEs that have a 5-digit disaggregation (9 CAEs).

Considering that the Table 5.4 only has CAEs that have 5-digit disagreggation, the results were split into two rows, specifying which aggregation level the winner model was trained - being the "4-D LEVEL" the models trained using information at the 4-digit level only, whereas in the "BOTTOM-UP" row the results were obtained by nowcasting every 5-digit CAE and aggregating it back to the 4-digit level using the bottom-up approach.

Apart from the models that were used, it was decided to also combine the forecasts of all models by computing the arithmetic mean, in order to try to achieve a better nowcasting, since the evidence shown by Clemen (1989) is that combining forecasts from different models can lead

to more accurate forecasts. This combination is called *Mean* in this project and its results can be found in the Table A.11.

	RL (5.1)	RL (5.2)	MIDAS	ARIMA	DREG	MEAN
4-D LEVEL	3	6	1	13	1	4

Table 5.3: Best performing model across CAEs that do not have 5-digit disaggregation

	RL (5.1)	RL (5.2)	MIDAS	ARIMA	DREG	MEAN
4-D LEVEL	3	0	0	0	1	1
BOTTOM-UP	0	0	0	3	1	0

Table 5.4: Best performing model across CAEs that do have 5-digit disaggregation

As the Tables 5.3 and 5.4 show, the ARIMA models performed better in 16 CAEs, followed by the Linear regression models (5.1) and (5.2) models (in 6 CAEs each), the Mean, Dynamic regressions and MIDAS.

The fact that the ARIMA models performed better in almost half of the 4-digit CAEs is a reflection of the fact that the relationship between the e-Fatura and the Multibanco series is not the same across the CAEs, as the Section 5.1.3 points out. In some CAEs, the behaviour of both series is very similar, whereas in others, their fluctuations and seasonalities are out of sync and it can affect the performance of models that use the Multibanco series as predictor series, as explained in the Section 5.5.

When it comes to the Linear regression models, the models (5.1) and (5.2) have performed better for 6 CAEs each, which confirmed that it was indeed useful to build two different models, since the e-Fatura and the Multibanco series have their seasonality out of sync in some CAEs, as mentioned in the Subsection 5.1.3 and in the Section 5.2.

The Dynamic regression, Mean and the MIDAS regression did not perform as well as expected, performing better than the other models only for 9 CAEs when accounted together. The MIDAS case was the most disappointing one, because as stated in the Section 5.4, it was expected

that this model would have a better performance since it was using more granular data (weekly) in comparison to all the other models, which were using monthly data.

It is also important to note that, in 15 out of the 37 4-digit CAEs, the best performing model had their RMSE less than 5% smaller than the second best performing one, having the ratio $Best/2Best > 0.95$, which means that there are models performing relatively as good as other models for the same CAE. Moreover, in some CAEs, the best performing model had their RMSE more than 70% smaller ($Best/Worst < 0.30$) than the worst performing model, whereas in other CAEs the best performing model was less than 30% smaller than the worst ones ($Best/Worst > 0.70$). This disparity between the performance of the models across the 37 4-digit CAEs are shown in the Table A.12.

Chapter 6

Conclusion

6.1 Final remarks

This project, as explained in the previous chapters, aims to create a backup plan for Statistics Portugal for when the Tax Authority fails to deliver the e-Fatura data in time to nowcast the Business Turnover in Retail Trade index, reducing the risk of the institution not complying with the deadlines imposed by EUROSTAT and the risk of compromising the overall quality of these early estimates.

At first, an exploratory analysis was done for both time series (e-Fatura and Multibanco), and then, the relationship between them was analysed as well. Then, the nowcasting exercise was done using 4 different algorithms - Linear regression, Dynamic regression, MIDAS regression and ARIMA - for two different CAE aggregation levels (4-digit and 5-digit levels) in order to assess the aggregation level and the model that are the most efficient for estimating each 4-digit CAE category.

From the exploratory analysis, it was possible to see that the behaviour (level, fluctuations, etc.) of the CAEs can be similar or very different from one another, which holds true for both the e-Fatura and the Multibanco data. Furthermore, the relationship between the e-Fatura and the Multibanco series of each CAE depends on the specific CAE that is being analysed - in some CAEs, they are quite linear and behave quite similarly, whereas in other CAEs the relationship is

not clear, as both series behave differently.

Overall, as the results show, there is not a single model that can yield the most accurate nowcasts for all of the 37 4-digit CAEs, and there is not a specific CAE aggregation that yields the most accurate forecasts either. This is expected, since the relationship between both data for each CAE is different, and the behaviour of each CAE in each series is different too. The ARIMA models is the most accurate one for 16 CAEs, followed by the linear regression (with and without seasonal dummies) for 12 CAEs, the Mean for 5 CAEs, the Dynamic regression for 3 and the MIDAS regression only for 1 CAE.

6.2 Limitations and future work

The main limitation of this project is that, from 2020 onwards, the e-Fatura data are estimates made by Statistics Portugal. This happens because the data until 2019 was delivered by the Tax Authority at once, and then from 2020 onwards, they were delivered monthly, and these monthly delivers have gone through many pre-processing operations in order to handle the problems with the data that, which were explained in the Subsection 4.1.1.

The situation mentioned above significantly affects the potential of these models and framework, since what is being actually forecasted is the estimation of the Statistics Portugal, and not the true values like the ones available until 2019.

Furthermore, given the potential that this framework has shown, other models, apart from the ones that have been used in this project, should be tested in order to try and find models that could better describe and predict some (or all) CAEs' behaviour.

Bibliography

Armesto, M. T., Engemann, K. M., Owyang, M. T., et al. (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, *92*(6), 521–36.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, *25*(1), 146–166.

Autoridade Tributária. (n.d.). *Sobre o e-fatura.* https://info.portaldasfinancas.gov.pt/pt/faturas/pages/sobre_efatura.aspx. (Accessed: 08-01-2023)

Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics*, *120*(1), 387–422.

Burman, J. P. (1980). Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society: Series A (General)*, *143*(3), 321–337.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, *5*(4), 559–583.

Clements, M. P., & Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business & Economic Statistics*, *26*(4), 546–554.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Eslake, S. (2006). The importance of accurate, reliable and timely data. *Victoria, sn*.

EUROSTAT. (2022a). *Data - short-term business statistics.* https://ec.europa.eu/eurostat/web/

short-term-business-statistics/data. (Accessed: 12/12/2022)

EUROSTAT. (2022b). *Methodology - short-term business statistics.* https://ec.europa.eu/eurostat/web/short-term-business-statistics/methodology. (Accessed: 12/12/2022)

EUROSTAT. (2022c). *Short-term business statistics.* https://ec.europa.eu/eurostat/cache/metadata/en/sts_esms.htm. (Accessed: 12/12/2022)

Evans, T. D., Monsell, B. C., & Sverchkov, M. (n.d.). Review of available programs for seasonal adjustment of weekly data december 2021.

Friedman, B. M. (2000). *Monetary policy.* National Bureau of Economic Research Cambridge, Mass., USA.

Ghysels, E., Kvedaras, V., & Zemlys, V. (2016). Mixed frequency data sampling regression models: The R package midasr. *Journal of Statistical Software*, *72*(4), 1–35. doi: 10.18637/jss.v072.i04

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The midas touch: Mixed data sampling regression models.

Ghysels, E., Sinko, A., & Valkanov, R. (2007). Midas regressions: Further results and new directions. *Econometric reviews*, *26*(1), 53–90.

Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of forecasting*, *9*(3), 233–254.

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2021). *Introduction to econometrics with r.* Universität Duisburg-Essen.

Hyndman, R. (2023). fpp3: Data for "forecasting: Principles and practice" (3rd edition) [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=fpp3 (R package version 0.5)

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., … Yasmeen, F. (2023). forecast: Forecasting functions for time series and linear models [Computer software manual]. Retrieved from https://pkg.robjhyndman.com/forecast/ (R package version 8.21)

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis*, *55*(9),

2579–2589.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of statistical software*, *27*, 1–22.

Marques, M. S. M. A. (2014). *O impacto da rede multibanco na rentabilidade bancária em portugal* (Unpublished doctoral dissertation). FEUC.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

O'Hara-Wild, M., Hyndman, R., & Wang, E. (2023). fable: Forecasting models for tidy time series [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=fable (R package version 0.3.3)

Ollech, D., & Bundesbank, D. (2023). Economic analysis using higher-frequency time series: challenges for seasonal adjustment. *Empirical Economics*, *64*(3), 1375–1398.

Pfaff, B. (2008). *Analysis of integrated and cointegrated time series with r* (Second ed.). New York: Springer. Retrieved from https://www.pfaffikus.de (ISBN 0-387-27960-1)

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting gdp using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941–948.

Ryan, J. A., & Ulrich, J. M. (2023). xts: extensible time series [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=xts (R package version 0.13.1)

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607.

Statistics Portugal. (2007). *Classificação portuguesa de actividades económicas, revisão 3*. Author.

Statistics Portugal. (2019). Indices de volume de negocios e de emprego base 2015, versao 3.1.

United Nations. (1958). *International standard industrial classification of all economic activities (isic), revision 1*. Author.

United Nations. (2008). *International standard industrial classification of all economic activities (isic), revision 4*. Author.

Wang, E., Cook, D., & Hyndman, R. J. (2020). A new tidy data structure to support exploration

and modeling of temporal data. *Journal of Computational and Graphical Statistics*, *29*(3), 466-478. Retrieved from https://doi.org/10.1080/10618600.2019.1695624 doi: 10.1080/10618600.2019.1695624

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 1.1.2)

Appendix A

Appendix

CAE	4711	4719	4721	4722	4723	4724	4725	4726
Avg. Prop. (%)	38.41	4.76	0.71	2.41	0.53	0.2	0.26	0.45
CAE	4729	4730	4741	4742	4743	4751	4752	4753
Avg. Prop. (%)	0.76	11.16	1.55	0.57	0.17	0.32	4.93	0.16
CAE	4754	4759	4761	4762	4763	4764	4765	4771
Avg. Prop. (%)	1.36	3.12	0.3	1.57	0.01	1.44	0.16	6.07
CAE	4772	4773	4774	4775	4776	4777	4778	4779
Avg. Prop. (%)	0.96	6.97	0.62	0.9	0.78	1.26	4.95	0.2
CAE	4781	4782	4789	4791	4799			
Avg. Prop. (%)	0.23	0.05	0.1	0.79	0.83			

Table A.1: 4-digit CAEs weight on aggregate: e-Fatura

CAE	4711	4719	4721	4722	4723	4724	4725	4726
Avg. Prop. (%)	48.78	3.01	0.41	1.74	0.27	0.1	0.13	0.14
CAE	4729	4730	4741	4742	4743	4751	4752	4753
Avg. Prop. (%)	0.63	10.01	0.56	0.27	0.07	0.34	4.83	0.1
CAE	4754	4759	4761	4762	4763	4764	4765	4771
Avg. Prop. (%)	3.81	2.26	0.35	0.8	0.01	1.81	0.18	6.15
CAE	4772	4773	4774	4775	4776	4777	4778	4779
Avg. Prop. (%)	0.79	4.85	0.22	0.78	0.58	1.09	4.12	0.09
CAE	4781	4782	4789	4791	4799			
Avg. Prop. (%)	0.06	0.03	0.04	0.51	0.05			

Table A.2: 4-digit CAEs weight on aggregate: Multibanco

CAE	47	4711	4719	4721	4722	4723	4724	4725
TEST STAT	-4.18	-2.73	-1.10	-2.98	-2.47	-2.74	-2.87	-2.89
CAE	4726	4729	4730	4741	4742	4743	4751	4752
TEST STAT	-0.99	-3.86	-1.70	-2.63	-0.66	-2.56	-2.82	-3.75
CAE	4753	4754	4759	4761	4762	4763	4764	4765
TEST STAT	-3.67	-0.52	-3.02	-1.82	-1.96	-2.85	-2.11	-1.39
CAE	4771	4772	4773	4774	4775	4776	4777	4778
TEST STAT	-3.71	-1.84	-2.39	-0.37	-3.13	-3.74	-2.38	-2.98
CAE	4779	4781	4782	4789	4791	4799		
TEST STAT	-4.03	-3.91	-2.42	-3.66	-1.32	-4.47		

Table A.3: Cointegration tests: unit root test statistic on linear regression residuals

CAE	4711	4719	4721	4722	4723	4724
LR (5.1) 4D	253743307	69252254	3414689	11047629	1944071	845811
LR (5.1) BU	256234126	73538211	-	-	-	-
CAE	4725	4726	4729	4730	4741	4742
LR (5.1) 4D	1418709	3605527	24095060	190278337	12901375	8538421
LR (5.1) BU	-	-	24618605	-	-	-
CAE	4743	4751	4752	4753	4754	4759
LR (5.1) 4D	1480865	1651579	10694675	1050907	13390214	12179971
LR (5.1) BU	-	-	11891620	-	-	24479377
CAE	4761	4762	4763	4764	4765	4771
LR (5.1) 4D	5680791	6644838	42718	2523530	2944572	23436183
LR (5.1) BU	-	-	-	-	-	23195714
CAE	4772	4773	4774	4775	4776	4777
LR (5.1) 4D	4694536	26742658	6018515	9678690	1994214	7653884
LR (5.1) BU	4873893	-	-	-	2185921	-
CAE	4778	4779	4781	4782	4789	4791
LR (5.1) 4D	21595225	1480895	627630	555371	1459447	23075001
LR (5.1) BU	20994868	-	-	-	-	-
CAE	4799					
LR (5.1) 4D	10408119					
LR (5.1) BU	-					

Table A.4: Linear regression model 5.1 RMSE across the CAEs: 4-digit level (4D) and bottom-up (BU) approaches

CAE	4711	4719	4721	4722	4723	4724
LR (5.2) 4D	313667186	50088287	5139126	10387622	2695771	928758
LR (5.2) BU	314620239	52229439	-	-	-	-
CAE	4725	4726	4729	4730	4741	4742
LR (5.2) 4D	1259694	3140698	23445021	188627948	11307560	5890936
LR (5.2) BU	-	-	24129605	-	-	-
CAE	4743	4751	4752	4753	4754	4759
LR (5.2) 4D	756096	1826587	14978452	669602	11745515	13122740
LR (5.2) BU	-	-	15421652	-	-	26314617
CAE	4761	4762	4763	4764	4765	4771
LR (5.2) 4D	4115292	2117563	70410	8277326	2311236	27386731
LR (5.2) BU	-	-	-	-	-	27370256
CAE	4772	4773	4774	4775	4776	4777
LR (5.2) 4D	5205354	27221953	2233817	10922786	3019361	6272145
LR (5.2) BU	5710013	-	-	-	2972726	-
CAE	4778	4779	4781	4782	4789	4791
LR (5.2) 4D	21613412	3129398	894366	567070	1441214	25563761
LR (5.2) BU	23306405	-	-	-	-	-
CAE	4799					
LR (5.2) 4D	10272470					
LR (5.2) BU	-					

Table A.5: Linear regression model 5.2 RMSE across the CAEs: 4-digit level (4D) and bottom-up approaches (BU)

CAE	4711	4719	4721	4722	4723	4724	4725	4726
RATIO	0.809	1.383	0.664	1.064	0.721	0.911	1.126	1.148
CAE	4729	4730	4741	4742	4743	4751	4752	4753
RATIO	1.028	1.009	1.141	1.449	1.959	0.904	0.714	1.569
CAE	4754	4759	4761	4762	4763	4764	4765	4771
RATIO	1.140	0.928	1.380	3.138	0.607	0.305	1.274	0.856
CAE	4772	4773	4774	4775	4776	4777	4778	4779
RATIO	0.902	0.982	2.694	0.886	0.660	1.220	0.999	0.473
CAE	4781	4782	4789	4791	4799			
RATIO	0.702	0.979	1.013	0.903	1.013			

Table A.6: Ratio between the RMSE of the linear regression models (5.1) and (5.2)

CAE	4711	4719	4721	4722	4723	4724	4725	4726
$\hat{R}^2_{adj}$	0.426	0.171	0.555	0.372	0.473	0.845	0.963	0.718
CAE	4729	4730	4741	4742	4743	4751	4752	4753
$\hat{R}^2_{adj}$	0.380	0.630	0.112	0.471	0.731	0.901	0.795	0.860
CAE	4754	4759	4761	4762	4763	4764	4765	4771
$\hat{R}^2_{adj}$	0.617	0.863	0.037	0.594	0.695	0.941	0.814	0.975
CAE	4772	4773	4774	4775	4776	4777	4778	4779
$\hat{R}^2_{adj}$	0.944	0.541	0.364	0.875	0.695	0.964	0.836	0.050
CAE	4781	4782	4789	4791	4799			
$\hat{R}^2_{adj}$	0.825	0.810	0.636	0.170	0.145			

Table A.7: $\hat{R}^2_{adj}$ of the linear regression model (5.1) across 4-digit CAEs

CAE	4711	4719	4721	4722	4723	4724
DREG 4D	329590098	29207107	3345817	10833177	2029123	854878
DREG BU	327000423	29629457	-	-	-	-
CAE	4725	4726	4729	4730	4741	4742
DREG 4D	1350914	3022599	20628924	192871530	9380724	5315334
DREG BU	-	-	22068820	-	-	-
CAE	4743	4751	4752	4753	4754	4759
DREG 4D	719954	1283742	14952442	805851	6586415	11628386
DREG BU	-	-	13579163	-	-	14824225
CAE	4761	4762	4763	4764	4765	4771
DREG 4D	4405436	9606491	40835	8997750	4075250	27719359
DREG BU	-	-	-	-	-	34312972
CAE	4772	4773	4774	4775	4776	4777
DREG 4D	6057216	19749366	2804601	9822362	2721681	7168448
DREG BU	5538469	-	-	-	2326540	-
CAE	4778	4779	4781	4782	4789	4791
DREG 4D	12989258	1542161	716658	442512	1459447	12854912
DREG BU	12821624	-	-	-	-	-
CAE	4799					
DREG 4D	10727157					
DREG BU	-					

Table A.8: Dynamic regression model (DREG) RMSE across the CAEs: 4-digit level (4D) and bottom-up (BU) approaches

CAE	4711	4719	4721	4722	4723	4724
MIDAS 4D	305933361	83810093	7171463	23142786	4748253	2146984
MIDAS BU	320029473	85533632	-	-	-	-
CAE	4725	4726	4729	4730	4741	4742
MIDAS 4D	6113181	3457693	27541244	222234659	15248486	4560820
MIDAS BU	-	-	27493655	-	-	-
CAE	4743	4751	4752	4753	4754	4759
MIDAS 4D	2487183	5298994	25005788	2100173	21046969	33257779
MIDAS BU	-	-	23512684	-	-	42746288
CAE	4761	4762	4763	4764	4765	4771
MIDAS 4D	7125837	11355580	115305	23235039	7504526	109594253
MIDAS BU	-	-	-	-	-	102927151
CAE	4772	4773	4774	4775	4776	4777
MIDAS 4D	19116916	36920549	8467750	21861633	5479455	37041986
MIDAS BU	17768315	-	-	-	5205315	-
CAE	4778	4779	4781	4782	4789	4791
MIDAS 4D	46089088	3273025	3460428	920943	1498662	22681228
MIDAS BU	44474684	-	-	-	-	-
CAE	4799					
MIDAS 4D	13048708					
MIDAS BU	-					

Table A.9: MIDAS regression RMSE across the CAEs: 4-digit level (4D) and bottom-up (BU) approaches

CAE	4711	4719	4721	4722	4723	4724
ARIMA 4D	314537002	31821219	3240614	8580240	1464344	708286
ARIMA BU	313159982	28491369	-	-	-	-
CAE	4725	4726	4729	4730	4741	4742
ARIMA 4D	1095344	2409328	24637376	255754743	7653306	6233497
ARIMA BU	-	-	24816519	-	-	-
CAE	4743	4751	4752	4753	4754	4759
ARIMA 4D	644662	757117	11403686	758769	4722365	13596455
ARIMA BU	-	-	10452744	-	-	10470902
CAE	4761	4762	4763	4764	4765	4771
ARIMA 4D	3058228	2727825	36413	8617398	2296786	18250797
ARIMA BU	-	-	-	-	-	18146113
CAE	4772	4773	4774	4775	4776	4777
ARIMA 4D	7389664	16609268	2171178	9056894	3483519	7954036
ARIMA BU	6413752	16609268	-	-	4156207	-
CAE	4778	4779	4781	4782	4789	4791
ARIMA 4D	13514558	2560695	830518	310811	1982091	14959994
ARIMA BU	15130063	-	-	-	-	-
CAE	4799					
ARIMA 4D	11197702					
ARIMA BU	-					

Table A.10: ARIMA model RMSE across the CAEs: 4-digit level (4D) and bottom-up (BU) approaches

CAE	4711	4719	4721	4722	4723	4724
MEAN 4D	255281145	43726066	4245735	8851733	1684329	699091
MEAN BU	261760842	45143786	-	-	-	-
CAE	4725	4726	4729	4730	4741	4742
MEAN 4D	1188569	2749766	23950588	198377133	8461726	5419634
MEAN BU	-	-	24557956	-	-	-
CAE	4743	4751	4752	4753	4754	4759
MEAN 4D	947733	1819756	8918184	814158	10439761	10769994
MEAN BU	-	-	10242326	-	-	21860325
CAE	4761	4762	4763	4764	4765	4771
MEAN 4D	4397630	5562780	53026	5298847	1956551	25484028
MEAN BU	-	-	-	-	-	25099954
CAE	4772	4773	4774	4775	4776	4777
MEAN 4D	7735291	15896079	3513866	7387304	2674887	9355646
MEAN BU	6766139	-	-	-	2898908	-
CAE	4778	4779	4781	4782	4789	4791
MEAN 4D	19444883	1659705	1118005	458765	1443046	16498017
MEAN BU	19751656	-	-	-	-	-
CAE	4799					
MEAN 4D	10280242					
MEAN BU	-					

Table A.11: Mean method nowcast RMSE across the CAEs: 4-digit level (4D) and bottom-up (BU) approaches

CAE	4711	4719	4721	4722	4723	4724
Best	LR (5.1) 4D	ARIMA BU	ARIMA 4D	ARIMA 4D	ARIMA 4D	MEAN 4D
Best/2Best	0.9940	0.9755	0.9686	0.9693	0.8694	0.9870
Best/Worst	0.7699	0.3331	0.4519	0.3708	0.3084	0.3256
CAE	4725	4726	4729	4730	4741	4742
Best	ARIMA 4D	ARIMA 4D	DREG 4D	LR (5.2) 4D	ARIMA 4D	MIDAS 4D
Best/2Best	0.9216	0.8762	0.9348	0.9913	0.9045	0.8580
Best/Worst	0.1792	0.6682	0.7490	0.7375	0.5019	0.5342
CAE	4743	4751	4752	4753	4754	4759
Best	ARIMA 4D	ARIMA 4D	MEAN 4D	LR (5.2) 4D	ARIMA 4D	ARIMA BU
Best/2Best	0.8954	0.5898	0.8707	0.8825	0.7170	0.9722
Best/Worst	0.2592	0.1429	0.3566	0.3188	0.2244	0.2450
CAE	4761	4762	4763	4764	4765	4771
Best	ARIMA 4D	LR (5.2) 4D	ARIMA 4D	LR (5.1) 4D	MEAN 4D	ARIMA BU
Best/2Best	0.7431	0.7763	0.8917	0.4762	0.8519	0.9943
Best/Worst	0.4292	0.1865	0.3158	0.1086	0.2607	0.1656
CAE	4772	4773	4774	4775	4776	4777
Best	LM 4D	MEAN 4D	ARIMA 4D	MEAN 4D	LR (5.1) 4D	LR (5.2) 4D
Best/2Best	0.9632	0.9571	0.9720	0.8157	0.9123	0.8750
Best/Worst	0.2456	0.4305	0.2564	0.3379	0.3639	0.1693
CAE	4778	4779	4781	4782	4789	4791
Best	DREG BU	LR (5.1) 4D	LR (5.1) 4D	ARIMA 4D	LR (5.2) 4D	DREG 4D
Best/2Best	0.9871	0.9603	0.8758	0.7024	0.9987	0.8593
Best/Worst	0.2782	0.4525	0.1814	0.3375	0.7271	0.5029

CAE	4799
Best	LR (5.2) 4D
Best/2Best	0.9992
Best/Worst	0.7872

Table A.12: Best performing model (Best), ratio between the best performing and second best performing models' RMSE (Best/2Best), and ratio between the best and worst performing models' RMSE across all CAEs (Best/Worst)