Using the Statistical Machine Learning Models ARIMA and SARIMA to Measure the Impact of Covid-19 on Official Provincial Sales of Cigarettes in Spain

Andoni Andueza[1], Miguel Ángel Del Arco-Osuna[1], Bernat Fornés[1], Rubén González-Crespo[2], Juan Manuel Martín-Álvarez[1]*

[1] Faculty of Economics and Business, Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)
[2] School of Engineering and Technology, Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)

Received 9 November 2022 | Accepted 16 February 2023 | Early Access 20 February 2023

Abstract

From a public health perspective, tobacco use is addictive by nature and triggers several cancers, cardiovascular and respiratory diseases, reproductive disorders, and many other adverse health effects leading to many deaths. In this context, the need to eradicate tobacco-related health problems and the increasingly complex environments of tobacco research require sophisticated analytical methods to handle large amounts of data and perform highly specialized tasks. In this study, time series models are used: autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) to forecast the impact of COVID-19 on sales of cigarette in Spanish provinces. To find the optimal solution, initial combinations of model parameters automatically selected the ARIMA model, followed by finding the optimized model parameters based on the best fit between the predictions and the test data. The analytical tools Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were used to assess the reliability of the models. The evaluation metrics that are used as criteria to select the best model are: mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), mean percentage error (MPE), mean error (ME) and mean absolute standardized error (MASE). The results show that the national average impact is slight. However, in border provinces with France or with a high influx of tourists, a strong impact of COVID-19 on tobacco sales has been observed. In addition, the least impact has been observed in border provinces with Gibraltar. Policymakers need to make the right decisions about the tobacco price differentials that are observed between neighboring European countries when there is constant and abundant cross-border human transit. To keep smoking under control, all countries must make harmonized decisions.

Keywords

ARIMA, Cigarette Sales, COVID-19, Machine Learning, SARIMA, Statistical Modeling, Time-series Forecast.

DOI: 10.9781/ijimai.2023.02.010

I. Introduction

FUNDAMENTALLY, there are two strategic reasons why the development of tobacco usage and behavior in any nation through time is a pertinent subject. First, smoking is addictive by nature, and it causes many different cancers, cardiovascular and respiratory conditions, reproductive problems, and a host of other harmful health impacts that result in thousands of deaths every year. As a result, the health system is burdened with significant costs related to the harm caused by tobacco use – on average, health spending accounts for 11.5% of the country's GDP [1]. Second, high-income nations' budgets are significantly impacted by the special taxes collected on tobacco; in Spain, tobacco is the product that provides the most to tax collection.

* Corresponding author.
E-mail addresses: juanmanuel.martin@unir.net

Additionally, a recent study that concentrated on the Spanish market demonstrates that some provinces do not have accurate official sales data that may be used to evaluate smoking control measures [2]. Furthermore, the empirical literature on regional heterogeneity in tobacco sales in Spain, concludes that areas of Spain bordering countries with high price differentials, such as Gibraltar and France, generate clusters of low and high per capita tobacco consumption, respectively [3]. In this regard, the border and tourist provinces in that study [2] are those in which sales are most impacted, supporting the prevalence of illegal commerce and substantial cross-border transactions. Thus, the findings of this study demonstrate the efficacy of shared policies adopted by the governments of neighboring nations that preserve a little price difference between them. In addition, the Spanish context is characterized by the strong impact that economic recessions have on cigarette sales [4]-[6]. Finally, a recent study suggests that in certain regions the demand for tobacco is not inelastic with respect to the price in the long term [7], which can generate large effects on provincial sales.

This scenario calls for advanced analytical tools to handle vast volumes of data and carry out highly specialized activities to eradicate tobacco-related health issues and the increasingly complicated environments of tobacco research. Due to this, some research has already used machine learning (hereafter ML) methods to analyze data pertaining to the tobacco market [8]. The definition of machine learning (ML) historically has been described as "a branch of research that offers computers the ability to learn without being explicitly programmed" to forecast future data or make decisions in uncertain situations [9]. The main goal of ML is to employ "brute force" instead of human supervision while analyzing data. Because ML requires far less human supervision than computer guidance, it can be considered as a natural extension of conventional statistical methodologies [10]. Unsupervised learning and supervised learning are categories found within machine learning. The two sets of ML approaches each have distinctive qualities that may be of interest to researchers studying tobacco. They are each geared toward resolving a certain difficulty. The focus of supervised learning is prediction. To predict the values of one or more output or response variables for a specific set of input or predictor variables, a model must be trained and validated [11]. In this sense, supervised learning techniques are used when the goal is to create a high-precision predictive model for future data. For example, supervised learning is useful for any tobacco market research that calls for extremely precise forecasts, such the creation of a public health surveillance program that predicts the likelihood of adolescent smoking beginning automatically [8]. Unsupervised learning, on the other hand, does not require an output variable because its goal is to ascertain the underlying probability distribution of the data (also known as density estimation) [8]. Examining tobacco-related social media discussions and identifying probable nicotine dependency subtypes by examining patient brain MRI data are two examples of unsupervised learning in tobacco research [8].

As stated, ML is a very powerful analytical tool for tobacco market researchers, the approaches can be broadly divided into supervised and unsupervised learning. However, in addition to this classification of techniques, studies that apply ML to tobacco market analysis can also be classified by the data (input) used. In this sense, we can find studies that analyze content on social networks, clinical report texts or administrative data [8]. In fact, several published papers that analyze the tobacco market focus on administrative data of the analysis [12]-[13]. Many of these studies apply supervised learning techniques to predict a binary phenomenon related to smoking cessation, including the intention to quit [14], adherence to smoking cessation therapies [15] and craving smoking highs or lows during a quit attempt [16]. However, few studies have applied supervised learning techniques with the aim of predicting continuous variables using, for example, regression or random forest [8], [17]-[19].

Although ML has been applied to the analysis of tobacco-related topics, to our knowledge, ML has never been applied to study the relationship between COVID-19 and tobacco. The COVID-19 pandemic has posed a unique opportunity to combat tobacco use [20]. Tobacco use and site bans, border closures, and lockdowns have had both positive and negative impacts on tobacco control. A recent study concludes that cigarette consumption decreased during the COVID-19 lockdown in 2020 [21]. However, other papers conclude the opposite. Specifically, one of the recent works concludes that the pandemic generated a 13% increase in tobacco sales [22]. Another paper indicates that this increase is because nicotine users use tobacco as their main mechanism to cope with stress and anxiety [23]. In addition, a paper indicates that the COVID-19 pandemic is related to higher tobacco sales and suggests research into whether smoking habits have changed since the pandemic lockdowns [24]. Regarding the use of time series analysis to analyze changes in cigarette sales, only one study has been found that addresses this problem and concludes that the sales observed during the pandemic are higher than expected [25]. Following on from this, in relation to the increase in tobacco sales, another study suggests that the intention to quit smoking has seen a post COVID-19 pandemic decrease [26]. Finally, other works that analyze smoking and COVID-19 suggest that tobacco sales should have been prohibited during the pandemic given the great opportunity that COVID-19 presented to eradicate smoking [27]-[29].

In Spain, although there are no works in which ML is applied to the tobacco market to explain the impact of COVID-19 on tobacco sales, there are papers that have analyzed the influence of COVID-19 on different aspects related to tobacco from another perspective. Some literature indicates that during the COVID-19 lockdown in Spain, tobacco consumption decreased [30]. In this same line of lower prevalence, another paper indicates that the success rate for quitting smoking went from 25% to 35% [31]. Another work, which focuses on analyzing smokers' perception of their exposure to the virus, suggests that many smokers may have changed their smoking patterns and it is possible that those who reduced their tobacco use outnumbered those who increased their consumption [32]. Another study that analyzes the impact of COVID-19 on tobacco consumption suggests that no significant effect of the pandemic on tobacco consumption is observed in Spain [33]. Finally, there is a group of works that indicate that the impact that COVID-19 has had on tobacco consumption depends on personal demographic issues and that not all people acted the same [34]. In addition, this block includes works that warn of the urgent need for tobacco consumers to give up smoking due to the damage to the health of consumers caused by this harmful product [35], [36].

To the best of our knowledge, no study has yet been done on the regional effects that COVID-19 has had on the Spanish tobacco market. In this study, we attempt to predict what the provincial tobacco market would have looked like in the absence of the COVID-19 pandemic. Then, we quantify the impact of the pandemic on cigarette sales as the difference between the forecast and the actual data. The data used in the current study comes from the Commission for the Trade of Tobacco and covers the period from January 2005 to December 2021 in terms of cigarette sales. The remainder of the document is structured as follows: Section II provides a description of the data and statistical models employed, together with information about the mathematics that underlies them, analytical tools, and evaluation measures. Section III discusses the computational architecture of the model parameter selection process. Section IV uses time series analysis to explore in depth the provincial impact of COVID-19 on the tobacco market. The conclusions reached from this investigation are provided in Section V.

II. Methods

To accomplish the goal outlined in this work, we generated an estimate of cigarette sales for the 48 Spanish provinces from January 2020 to December 2021 using the ML ARIMA and SARIMA statistical models. The ARIMA and SARIMA models as the best model over the uncorrelated ones and the models based on neural networks, because although these have a similar accuracy, the computational cost is much higher [37]. The suggested models have been optimized by choosing the most suitable parameters for each province. To ensure that the time series is the same length across all provinces, we used January 1, 2005, as the start date for each province. A minimum sample size of 30 observations is reportedly needed to provide a statistically significant forecast of time series data [38]. Given that each province's model was trained using data from January 2005 to December 2017 (168 observations), the sample size for estimating cigarette sales is significantly larger than the threshold set.

A. Data

A panel of monthly data from the Spanish provinces from January 2005 to December 2021 was used to build our empirical research. The Commission for the Trade of Tobacco's website's statistics section provided the cigarette sales data in euros and units. The National Institute of Statistics of Spain has been used to collect data on the population over the age of 18 to estimate provincial sales per capita.

The Islas Canarias, Ceuta and Melilla have been excluded from the analysis. As for the Islas Canarias, neither the tobacco market is regulated under a monopoly, nor is the price set by the Spanish government. That is, there is free trade, and the Spanish government does not intervene in the price. In addition, the restrictive regulations on consumption also have special features. In this sense, if that region is included in the study, the paper would present two important limitations. On the one hand, the behavior of Islas Canarias could be totally different as the population could more easily access tobacco consumption, given the free sale. On the other hand, the fact that the market is not regulated under a monopoly in these regions (singular), makes the data not homogeneous and reliable. As for Ceuta and Melilla, the data published by the Commission for the Trade of Tobacco is not homogeneous. Although sales of Ceuta and Melilla have been separated for a few years, until then the aggregate data was published, although they are two independent autonomous cities. Therefore, we do not have consistent data to analyze what happened in these autonomous cities.

B. Statistical Models and Description

Time series are collections of numerical values that each have a periodic component. Time series can be divided into two groups: stationary time series and non-stationary time series, depending on how the numerical values of the time series behave. Non-stationary time series have patterns that prevent the mean and/or variance from being constant, whereas stationary time series do not exhibit patterns in their mean and/or variance with respect to time. Seasonality or trend may be to blame for these trends. Calculating the difference between two succeeding observations can make non-stationary time series stationary. The trend and seasonality are eliminated from the time series using the differencing approach. First and second order differentiation are the two differentiation procedures that are most frequently employed; their calculation processes are described in equations (1) and (2):

$$y_t^{`} = y_t - y_{t-1} \tag{1}$$

$$y_t^{``} = y_t - 2y_{t-1} + y_{t-2} \tag{2}$$

where y_t are non-stationary time series data, $y_t^{`}$ is the time series after first order differentiation, $y_t^{``}$ is the time series after second order differentiation, y_{t-1} is the time series observation in period t-1, y_{t-2} is the time series observation in period t-2. Only when the time series is non-stationary after first-order differentiation is second-order differentiation required. There is also the option of seasonal distinction. In this instance, the distance between an observation and the identical observation from the prior year is used to calculate the difference (or period). Equation (3) provides a definition for the first degree of seasonal differentiation .

$$y_s^{`} = y_t - y_{t-m} \tag{3}$$

where $y_s^{`}$ is the time series after the first-order seasonal differentiation, y_{t-m} is the observation of the period t-m, m is the number of periods that exist between an observation and the same in the previous period. In this work, the time series were subjected to differentiation to eliminate seasonality and the resulting dataset is the one used to make the estimates. In addition, it must be taken into account that the estimation of the parameters of the ARIMA and SARIMA models is carried out assuming 4 basic assumptions: (i) the time series do not contain atypical points, (ii) the time series are composed of a single variable that is the one that, with its past values, helps to make the predictions; (iii) the time series are stationary, (iv) the model parameters and errors are constant throughout the time period.

Box and Jenkin created the ARIMA (p, d, q) model in 1976 [39], which can be used to predict stationary time series without seasonality. Three terms—p, d, and q—define this ARIMA model. The order of the moving average (MA) term is q, the order of the autoregression (AR) term is p, and the order of differentiation required to keep the time series stationary is d. The regression of the variable against itself to forecast its future behavior is known as autoregression. It involves comparing the value observed at a certain point to the values from earlier times. MA is a regression-like model that forecasts a variable in a later stage using the forecasting errors from an earlier time stage. The generalized equations for the p-th order AR model and the q-th order MA model are given below (Eqs. (4) and (5), respectively).

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p + y_{t-p} + \varepsilon_t \tag{4}$$

$$y_t = C + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{5}$$

The AR model (Eq. (4)), the integration (I), and the MA model (Eq. (5)) are all combined to create ARIMA models in this study. To create the forecast, integration (I) uses differentiation in reverse. The mathematical formulation of the generalized ARIMA model is Eq (6).

$$y_t = C + \phi_1 y + \phi_p y_{t-p} + \cdots + \phi_n y_{t-n} + \theta_1 \varepsilon_{t-1} + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{6}$$

Where C is the independent term, ϕ_i ($i = 1, 2 \ldots p$) are the autoregressive model parameters, θ_i ($i = 1, 2 \ldots p$) are the moving average model parameters, y_t is the current time series, $y_{t-1}, y_{t-2}, \ldots y_{t-p}$ are past values and ε_t is random error of period t and is given by the following equation:

$$\varepsilon_t = y_t - y_{t-1} \tag{7}$$

To account for the seasonality of the time series, the seasonal ARIMA (SARIMA) model combines the non-seasonal ARIMA (p, d, and q) with additional seasonal terms (P, D, and Q). The seasonal AR term, seasonal moving average term, and seasonal differencing term are represented, respectively, by the P, Q, and D terms. The general SARIMA model is mathematically represented as follows:

$$\Phi_p(B^m)\phi_p(B)(1 - B^m)^D(1 - B)^d y_t = \Theta_Q(B^m)\theta_q(B)w_t \tag{8}$$

Where y_t is the non-stationary time series, w_t is the Gaussian white noise process, $\phi(B)$ is a non-seasonal autoregressive polynomial and $\theta(B)$ is a non-seasonal moving average polynomial, D is the seasonal differencing (the term is equal to 1 or 2, etc.). However, the value of D = 1 is sufficient to impose stationarity on the data, $\Phi(B^m)$ is a seasonal autoregressive polynomial, and $\Theta(B^m)$ is a seasonal moving average polynomial. Where B is defined as the backtracking operator which is expressed as follows:

$$B^k y_t = y_{t-k} \tag{9}$$

The expressions for the moving average model -Eq. (11)-, non-seasonal autoregressive model -Eq. 10-, seasonal AR model -Eq. 12-, and seasonal MA model -Eq. 13- are provided below.

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^{Pm} \tag{10}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{11}$$

$$\Phi_p(B^m) = 1 - \Phi_1 B^m - \Phi_2 B^{2m} + \cdots + \Phi_p B^{Pm} \tag{12}$$

$$\Theta_Q(B^m) = 1 + \Theta_1 B^m + \Theta_2 B^{2m} + \cdots + \Theta_Q B^{Qs} \tag{13}$$

Indicators are used to judge the accuracy of the time series analysis once the parameters of the ARIMA and SARIMA models have been estimated and the predictions have been produced. These indicators include the partial autocorrelation function (PACF), the Akaike information criterion (AIC), the autocorrelation function (ACF), and the Bayesian information criteria (BIC). These metrics show how the time series' observations relate to one another. While PACF correlates the time series with its own lagged values spaced by specific time units, ACF provides the correlation of the time series data with its prior time series data. The AIC and BIC penalized likelihood criterion's values are related; the lower they are, the more probable it is that the model will be accepted as a genuine model. Additionally, this study's evaluation criteria include mean error (ME), root mean square error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), and scaled mean absolute error (MASE).

In a time series, autocorrelation is the relationship between the most recent observation and lagging observations. The ACF describes the linear relationship between the observation at time t and the observation at a previous time, and the autocorrelation plot is the time series' representation of autocorrelation vs delays (t-k). To illustrate, the ACF for the time series y_t is given by:

$$ACF(y_t, y_{t-k}) = \frac{Covariance\ (y_t, y_{t-k})}{variance\ (y_t)} \tag{14}$$

where k is the delay and is defined as the difference between y_t and y_{t-k}. On the other hand, in partial autocorrelation, the intermediate observations are considered when calculating the correlation between two observations at different times. For example, consider that a time series y_t, the PACF between two observations y_t and y_{t-2} (assuming k = 2) can be written as shown in the equation (15).

$$PACF(y_t, y_{t-2}) = \frac{Covariance\ (y_t, y_{t-2}|y_{t-1})}{\sqrt{variance\ (y_t|y_{t-1})}\sqrt{variance\ (y_{t-2}|y_{t-1})}} \tag{15}$$

Testing the created models is necessary to see how well they function in terms of elucidating the relationships between the variables. We have evaluated a model's ability to explain relationships using the information criteria. AIC and BIC are two widely used measures that assess the quality of models by rewarding those that have fewer mistakes and penalizing those that have too many parameters. The following is how AIC is mathematically represented:

$$AIC = -2logL(\hat{\theta}) + 2K \tag{16}$$

Where K is the total number of model parameters and $logL(\hat{\theta})$ is the likelihood function. BIC is a different model selection criterion in a similar vein. Compared to AIC, BIC imposes a lower penalty on the quantity of parameters. The model with the highest probability value is represented by the lower value in both the AIC and BIC settings. As a result, it aids time series analysts in selecting the optimal model from among the limited number of generated alternative models. The following is how BIC is mathematically represented:

$$BIC = -2logL(\hat{\theta}) + KlogN \tag{17}$$

Where N is the number of observations.

MAE, RMSE, MAPE, MPE, ME and MASE are often used to assess the accuracy of the ML models [40]-[41], which are given by the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{18}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{20}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \tag{21}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i \tag{22}$$

$$MASE = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|} \tag{23}$$

Where $\hat{y}_i$ is the prediction made by the model and y_t is the actual value.

III. Computational Framework for Model Development

The scripts were created using the R programming language, which was set up in the RStudio environment, to accomplish the goal mentioned in this study [42]-[43]. The tidyverse and prediction libraries have also been used to clean the data, estimates, and graphic representations [44]–[46]. The appendix contains the R script that was utilized throughout key stages of the data analysis. Although there has already been a data cleansing phase, the actions used by the ML algorithm to accomplish the specified aim are detailed in this section. The algorithm initially determines whether each time series exhibits non-stationarity (if it had been done manually, this would have been checked using ACF and PACF plots). The time series is not stationary if the autocorrelation only slightly decreases as the number of delays increases. Next, the technique applies differences before executing ARIMA or SARIMA modeling if there is evidence that the time series is not stationary. Depending on which option best fits the time series, the algorithm selects either ARIMA or SARIMA. Given the substantial seasonal component present in the time series of tobacco sales, the method used SARIMA in the case study in this paper for all the series. The SARIMA models require an average processing time of 7 seconds to complete each simulation on the local computer.

The manual selection of the best parameter (p, d, q) $(P, D, Q)_m$ of the ARIMA and SARIMA models using ACF and PACF graphs can take a long time, since the models have been estimated for 48 provinces and 3 different variables (euros, packs and per capita packs). To select the appropriate combination of model parameter values, we perform a grid search using the forecast library, as indicated in the previous paragraph. This library uses AIC as an evaluation metric to choose the best model among several ARIMA and SARIMA models. Given that all the time series used begin in January 2005, end in December 2021 and tobacco sales show a strong seasonality, the parameter m took a value of 12 in all cases (Tables I, II and III).

The time series data of the 48 Spanish provinces was divided into two parts: the selected training dataset goes from January 2005 to December 2017 and the validation dataset goes from January 2018 to December 2019. Utilizing the training dataset, the model is constructed, and the validation dataset is used to estimate the model's performance. The following assessment metrics were used to assess the model: MAE, RMSE, MAPE, MPE, ME, and MASE. The model was used to forecast tobacco sales values from January 2020 to December 2021 (the period in which actual sales are altered due to lockdowns, restrictions in the hotel industry, and closure of borders for the 48 Spanish provinces), after the best model had been determined by training on the training dataset. Finally, to estimate the impact that COVID-19 has had on tobacco sales in Spain, the estimates made by the SARIMA models are compared with the actual sales observed from January 2020 to December 2021.

TABLE I. Selected SARIMA Models for Forecasting Euros

Province	SARIMA (p,d,q) (P,D,Q,m)	AIC	BIC	MAPE
Alava	(1,1,4)(2,0,0,12)	4,63E+03	4,66E+03	7,60E+00
Albacete	(2,1,2)(1,1,0,12)	4,25E+03	4,26E+03	7,63E+00
Alicante	(1,0,2)(2,1,1,12)	4,77E+03	4,79E+03	1,05E+01
Almería	(4,1,0)(2,1,2,12)	4,33E+03	4,35E+03	9,14E+00
Asturias	(2,1,1)(2,1,2,12)	4,48E+03	4,51E+03	6,97E+00
Ávila	(2,1,2)(1,1,0,12)	4,03E+03	4,05E+03	1,48E+01
Badajoz	(2,1,2)(2,1,0,12)	4,39E+03	4,41E+03	7,44E+00
Balears	(3,0,1)(2,1,0,12)	4,68E+03	4,70E+03	1,22E+01
Barcelona	(5,1,1)(2,0,0,12)	5,35E+03	5,38E+03	6,78E+00
Burgos	(0,0,0)(2,0,0,12)	4,78E+03	4,79E+03	9,73E+00
Cáceres	(2,1,2)(1,1,0,12)	4,28E+03	4,30E+03	9,32E+00
Cádiz	(2,1,2)(2,1,2,12)	4,44E+03	4,47E+03	1,19E+01
Cantabria	(2,1,2)(1,1,0,12)	4,36E+03	4,38E+03	8,78E+00
Castellón	(2,1,2)(2,1,0,12)	4,36E+03	4,39E+03	1,07E+01
Ciudad Real	(2,1,1)(2,1,0,12)	4,31E+03	4,33E+03	7,42E+00
Córdoba	(5,1,3)(2,0,0,12)	4,75E+03	4,79E+03	6,68E+00
Coruña (A)	(0,1,4)(2,0,0,12)	4,89E+03	4,91E+03	7,49E+00
Cuenca	(2,1,2)(2,1,0,12)	4,17E+03	4,19E+03	1,00E+01
Girona	(2,1,2)(2,1,0,12)	4,67E+03	4,69E+03	2,25E+01
Granada	(2,1,2)(2,1,0,12)	4,44E+03	4,46E+03	7,50E+00
Guadalajara	(2,1,2)(1,1,0,12)	4,09E+03	4,11E+03	7,84E+00
Guipúzcoa	(2,1,2)(1,1,0,12)	4,56E+03	4,58E+03	1,07E+01
Huelva	(2,1,2)(2,1,2,12)	4,30E+03	4,33E+03	1,04E+01
Huesca	(2,1,2)(2,1,0,12)	4,13E+03	4,15E+03	1,08E+01
Jaén	(2,1,2)(2,0,0,12)	4,78E+03	4,80E+03	5,86E+00
León	(2,1,2)(1,1,0,12)	4,33E+03	4,35E+03	9,12E+00
Lleida	(1,1,2)(1,0,0,12)	4,84E+03	4,86E+03	1,06E+01
Lugo	(2,1,2)(1,1,0,12)	4,23E+03	4,24E+03	8,42E+00
Madrid	(5,1,0)(2,0,0,12)	5,36E+03	5,38E+03	5,78E+00
Málaga	(2,1,1)(2,1,2,12)	4,62E+03	4,64E+03	1,20E+01
Murcia	(2,1,2)(2,1,2,12)	4,53E+03	4,56E+03	6,90E+00
Navarra	(2,1,2)(2,1,0,12)	4,54E+03	4,56E+03	1,08E+01
Ourense	(2,1,2)(1,1,0,12)	4,14E+03	4,16E+03	7,92E+00
Palencia	(2,1,2)(0,0,2,12)	4,51E+03	4,53E+03	9,06E+00
Pontevedra	(3,1,2)(2,1,2,12)	4,40E+03	4,43E+03	8,91E+00
Rioja (La)	(2,1,1)(2,0,0,12)	4,62E+03	4,64E+03	7,74E+00
Salamanca	(0,1,1)(0,0,2,12)	4,73E+03	4,74E+03	1,03E+01
Segovia	(2,1,2)(1,1,0,12)	4,03E+03	4,05E+03	1,02E+01
Sevilla	(2,1,2)(2,1,0,12)	4,60E+03	4,62E+03	6,71E+00
Soria	(2,1,2)(1,1,0,12)	3,94E+03	3,96E+03	9,94E+00
Tarragona	(1,0,0)(2,1,0,12)	4,57E+03	4,58E+03	1,37E+01
Teruel	(2,1,2)(1,1,0,12)	4,01E+03	4,03E+03	1,15E+01
Toledo	(2,1,2)(2,1,2,12)	4,37E+03	4,40E+03	7,16E+00
Valencia	(4,1,1)(2,0,0,12)	5,12E+03	5,15E+03	6,31E+00
Valladolid	(2,1,1)(2,0,0,12)	4,76E+03	4,78E+03	6,89E+00
Vizcaya	(4,1,3)(2,0,0,12)	4,88E+03	4,92E+03	5,48E+00
Zamora	(2,1,2)(2,1,0,12)	4,07E+03	4,09E+03	1,08E+01
Zaragoza	(2,1,2)(2,0,0,12)	4,89E+03	4,91E+03	6,29E+00

TABLE II. Selected SARIMA Models for Forecasting Packs

Province	SARIMA (p,d,q) (P,D,Q,m)	AIC	BIC	MAPE
Alava	(2,1,2)(2,0,1,12)	4,31E+03	4,33E+03	7,56E+00
Albacete	(2,1,2)(2,1,2,12)	3,92E+03	3,94E+03	7,62E+00
Alicante	(1,0,2)(2,1,1,12)	4,77E+03	4,79E+03	1,05E+01
Almería	(3,1,2)(2,1,1,12)	4,01E+03	4,04E+03	9,13E+00
Asturias	(2,1,2)(2,0,0,12)	4,54E+03	4,57E+03	6,94E+00
Ávila	(2,1,2)(2,1,0,12)	3,71E+03	3,73E+03	1,48E+01
Badajoz	(5,1,1)(2,1,1,12)	4,05E+03	4,08E+03	7,45E+00
Balears	(1,1,4)(2,1,0,12)	4,38E+03	4,40E+03	1,23E+01
Barcelona	(2,1,2)(2,0,0,12)	4,99E+03	5,02E+03	6,75E+00
Burgos	(4,1,1)(2,0,0,12)	4,39E+03	4,41E+03	9,71E+00
Cáceres	(3,1,1)(1,1,2,12)	3,94E+03	3,97E+03	9,31E+00
Cádiz	(2,1,3)(2,1,0,12)	4,15E+03	4,18E+03	1,19E+01
Cantabria	(4,1,1)(2,1,2,12)	4,02E+03	4,05E+03	8,74E+00
Castellón	(2,1,0)(2,1,1,12)	4,05E+03	4,07E+03	1,06E+01
Ciudad Real	(2,1,2)(2,1,1,12)	3,98E+03	4,00E+03	7,45E+00
Córdoba	(2,1,2)(2,0,0,12)	4,43E+03	4,45E+03	6,71E+00
Coruña (A)	(2,1,2)(2,0,0,12)	4,53E+03	4,55E+03	7,47E+00
Cuenca	(1,1,4)(2,1,2,12)	3,86E+03	3,89E+03	1,00E+01
Girona	(2,1,2)(1,1,0,12)	4,38E+03	4,40E+03	2,22E+01
Granada	(3,1,3)(2,1,2,12)	4,12E+03	4,15E+03	7,49E+00
Guadalajara	(3,1,2)(2,1,2,12)	3,75E+03	3,78E+03	7,84E+00
Guipúzcoa	(2,1,1)(1,1,2,12)	4,23E+03	4,25E+03	1,06E+01
Huelva	(2,1,1)(2,1,2,12)	4,01E+03	4,03E+03	1,03E+01
Huesca	(4,1,1)(2,1,1,12)	3,81E+03	3,83E+03	1,07E+01
Jaén	(2,1,2)(2,0,0,12)	4,43E+03	4,46E+03	5,82E+00
León	(4,1,3)(1,1,2,12)	3,96E+03	3,99E+03	9,10E+00
Lleida	(2,1,1)(2,0,0,12)	4,49E+03	4,51E+03	9,46E+00
Lugo	(4,1,1)(2,0,0,12)	4,22E+03	4,25E+03	8,41E+00
Madrid	(3,1,3)(2,0,0,12)	5,03E+03	5,06E+03	5,73E+00
Málaga	(2,1,2)(2,1,0,12)	4,32E+03	4,34E+03	1,19E+01
Murcia	(3,1,1)(2,1,1,12)	4,22E+03	4,25E+03	6,87E+00
Navarra	(2,1,2)(2,1,0,12)	4,23E+03	4,25E+03	1,07E+01
Ourense	(2,1,2)(1,1,1,12)	3,81E+03	3,83E+03	7,91E+00
Palencia	(2,1,2)(2,0,0,12)	4,13E+03	4,16E+03	9,01E+00
Pontevedra	(4,1,0)(2,1,2,12)	4,09E+03	4,12E+03	8,84E+00
Rioja (La)	(3,1,4)(2,0,0,12)	4,26E+03	4,29E+03	7,71E+00
Salamanca	(0,1,2)(2,0,0,12)	4,36E+03	4,37E+03	1,03E+01
Segovia	(2,1,2)(2,1,0,12)	3,71E+03	3,73E+03	1,01E+01
Sevilla	(2,1,0)(2,1,2,12)	4,31E+03	4,33E+03	6,80E+00
Soria	(3,0,1)(2,1,0,12)	3,63E+03	3,65E+03	9,90E+00
Tarragona	(3,1,1)(2,1,0,12)	4,22E+03	4,24E+03	1,36E+01
Teruel	(3,1,3)(2,1,0,12)	3,71E+03	3,73E+03	1,14E+01
Toledo	(2,1,2)(2,1,2,12)	4,06E+03	4,09E+03	7,17E+00
Valencia	(2,1,0)(2,0,0,12)	4,78E+03	4,79E+03	6,24E+00
Valladolid	(1,1,2)(2,0,0,12)	4,42E+03	4,44E+03	6,86E+00
Vizcaya	(2,1,2)(2,0,0,12)	4,54E+03	4,56E+03	5,45E+00
Zamora	(1,1,3)(1,1,2,12)	3,74E+03	3,77E+03	1,08E+01
Zaragoza	(2,1,2)(2,0,0,12)	4,53E+03	4,55E+03	6,26E+00

TABLE III. Selected SARIMA Models for Forecasting Per Capita Packs

Province	SARIMA (p,d,q) (P,D,Q,m)	AIC	BIC	MAPE
Alava	(2,1,2)(2,0,1,12)	4,30E+02	4,57E+02	7,67E+00
Albacete	(2,1,2)(2,1,2,12)	2,85E+02	3,11E+02	7,61E+00
Alicante	(2,1,0)(2,1,0,12)	3,72E+02	3,87E+02	1,03E+01
Almería	(3,1,2)(2,1,1,12)	2,47E+02	2,73E+02	9,19E+00
Asturias	(2,1,2)(2,0,0,12)	2,71E+02	2,95E+02	6,94E+00
Ávila	(5,1,1)(2,1,0,12)	2,94E+02	3,20E+02	1,48E+01
Badajoz	(5,1,1)(2,1,1,12)	2,61E+02	2,91E+02	7,44E+00
Balears	(2,1,0)(2,1,0,12)	4,96E+02	5,11E+02	1,36E+01
Barcelona	(2,1,2)(2,0,0,12)	2,42E+02	2,66E+02	6,82E+00
Burgos	(2,1,1)(2,0,0,12)	4,60E+02	4,78E+02	9,71E+00
Cáceres	(2,1,2)(2,1,1,12)	2,90E+02	3,13E+02	9,31E+00
Cádiz	(2,1,2)(2,1,0,12)	2,08E+02	2,29E+02	1,18E+01
Cantabria	(4,1,1)(2,1,2,12)	2,68E+02	2,98E+02	8,74E+00
Castellón	(3,1,2)(2,1,1,12)	3,03E+02	3,30E+02	1,06E+01
Ciudad Real	(2,1,2)(2,1,1,12)	2,69E+02	2,92E+02	7,45E+00
Córdoba	(2,1,2)(2,0,0,12)	2,77E+02	3,02E+02	6,67E+00
Coruña (A)	(2,1,2)(2,0,0,12)	2,45E+02	2,67E+02	7,46E+00
Cuenca	(1,1,2)(2,1,1,12)	3,91E+02	4,12E+02	1,00E+01
Girona	(2,1,2)(1,1,1,12)	5,82E+02	6,02E+02	2,21E+01
Granada	(4,1,0)(2,1,2,12)	2,57E+02	2,84E+02	7,43E+00
Guadalajara	(5,1,0)(2,1,1,12)	2,69E+02	2,96E+02	7,72E+00
Guipúzcoa	(2,1,1)(1,1,1,12)	4,26E+02	4,44E+02	1,06E+01
Huelva	(2,1,1)(2,1,0,12)	3,15E+02	3,33E+02	1,03E+01
Huesca	(4,1,1)(2,1,1,12)	3,23E+02	3,50E+02	1,07E+01
Jaén	(2,1,2)(2,0,0,12)	3,39E+02	3,63E+02	5,82E+00
León	(4,1,3)(1,1,2,12)	2,39E+02	2,72E+02	9,10E+00
Lleida	(2,1,1)(2,0,0,12)	5,22E+02	5,40E+02	1,07E+01
Lugo	(4,1,1)(2,0,0,12)	2,88E+02	3,12E+02	8,41E+00
Madrid	(2,1,4)(2,0,0,12)	2,31E+02	2,61E+02	5,88E+00
Málaga	(2,1,2)(2,1,0,12)	3,06E+02	3,27E+02	1,20E+01
Murcia	(5,1,0)(2,1,1,12)	2,28E+02	2,55E+02	6,99E+00
Navarra	(2,1,2)(2,1,0,12)	4,66E+02	4,87E+02	1,07E+01
Ourense	(2,1,2)(1,1,1,12)	1,95E+02	2,16E+02	7,91E+00
Palencia	(2,1,1)(2,0,0,12)	4,29E+02	4,47E+02	9,00E+00
Pontevedra	(4,1,0)(2,1,2,12)	2,03E+02	2,29E+02	8,84E+00
Rioja (La)	(2,1,2)(2,0,0,12)	3,88E+02	4,12E+02	7,77E+00
Salamanca	(0,1,2)(2,0,0,12)	4,37E+02	4,52E+02	1,03E+01
Segovia	(5,1,0)(2,1,2,12)	3,16E+02	3,46E+02	1,01E+01
Sevilla	(2,1,0)(2,1,0,12)	2,41E+02	2,55E+02	6,70E+00
Soria	(1,0,0)(1,1,0,12)	3,83E+02	3,95E+02	9,90E+00
Tarragona	(3,1,1)(2,1,0,12)	4,00E+02	4,21E+02	1,36E+01
Teruel	(5,0,2)(2,1,0,12)	3,40E+02	3,73E+02	1,14E+01
Toledo	(4,1,0)(2,1,2,12)	2,80E+02	3,07E+02	7,09E+00
Valencia	(2,1,2)(2,0,0,12)	2,68E+02	2,92E+02	6,33E+00
Valladolid	(2,1,2)(2,0,0,12)	3,76E+02	3,98E+02	6,88E+00
Vizcaya	(5,1,4)(2,0,0,12)	2,36E+02	2,73E+02	5,53E+00
Zamora	(1,1,3)(1,1,1,12)	2,84E+02	3,05E+02	1,08E+01
Zaragoza	(2,1,2)(2,0,0,12)	3,05E+02	3,26E+02	6,35E+00

IV. Results and Discussions

Table IV shows the results of the comparison between the actual sales observed after COVID-19 and the estimates made by the model (from January 2020 to December 2021). In this sense, the results of the gaps detected in terms of sales in euros, in packs and in per capita packs are shown. Positive gaps indicate that observed sales exceed the estimates made by the model, while negative gaps indicate that actual sales after COVID-19 are lower than the estimates made by the estimated SARIMA models. In the table, the minimum, maximum and average of the calculated provincial gaps can be observed. In addition, Fig. 1 graphically shows the dynamics of the time series together with the forecast made using the variable per capita packs.

If we focus on the calculated average gap, in some provinces the impact of COVID-19 on tobacco sales has been almost nil. Specifically, in Almería, Ávila, Cantabria, Coruña (A), Valladolid and Zaragoza, the impact of COVID-19 on per capita packs is less than 1% in absolute value. Given this situation, in Fig. 1 it can be seen how in these provinces, in which tobacco sales were not affected by COVID-19, the forecast lines and actual post-COVID-19 sales overlap. However, in other provinces the impact of COVID-19 has caused a significant negative effect on sales per capita packs that reaches, on average, up to -25.72%. The provinces in which this situation is observed are Alicante/Alacant, Baleares (Illes), Girona, Guipúzcoa, Lleida and Málaga, in which the average impact of COVID-19 on monthly tobacco sales has been -18, 95%, -25.72%, -22.71%, -14.98%, -16.66% and -11.41%, respectively. In the case of these provinces, Fig. 1 shows that the forecast line exceeds the Post COVID-19 sales line from January 2020 to December 2021. In all cases, the provinces in which these effects are observed are from areas with a high influx of tourists and border areas with France. These results are in line with previous literature indicating that tobacco sales in Spain are highly conditioned by sales to tourists and residents of France [2],[47].

Regarding the minimum value of the provincial gaps calculated in sales per capita packs, Table IV shows that the greatest impact, in absolute value, of COVID-19 on tobacco sales was observed in Alicante/Alacant, Balears (Illes), Girona, Guipúzcoa, Lleida and Navarra, in which the minimum value of the impact of COVID-19 on monthly tobacco sales was -39.11%, -58.20%, -66.74%, -54, 48%, -46.69% and -51.38%, respectively. In all cases, this minimum value was detected in the months of February and/or March 2020, months in which the borders of Spain were closed due to the COVID-19 pandemic. In other words, the greatest impact in absolute value of COVID-19 on tobacco sales is also observed in provinces bordering France and provinces with a high influx of tourists. On the other hand, the provinces in which the minimum impact has been smaller in absolute value are Cádiz and Sevilla, where said impact has been -11.37% and -5.41%, respectively. These results are also in line with previous literature that indicates that sales in Cádiz and Sevilla are affected by the proximity of these provinces to Gibraltar, an area with which there is a significant price differential [2].

Our results indicate that the restrictions implemented by governments due to COVID-19 have had a significant effect on provincial tobacco sales in Spain. In this sense, we find that the provinces in which sales are most affected are the border and tourist provinces, which seems to indicate that, regardless of the limitation of leisure, the restriction that has most affected sales is the closure of borders. The results suggest that in tourist and border areas with France, COVID-19 has caused a negative effect on tobacco sales that in most cases had not yet been reversed by December 2021.

Fig. 1 (A). 2020 and 2021 forecast of the per capita packs based on the best SARIMA models selected.

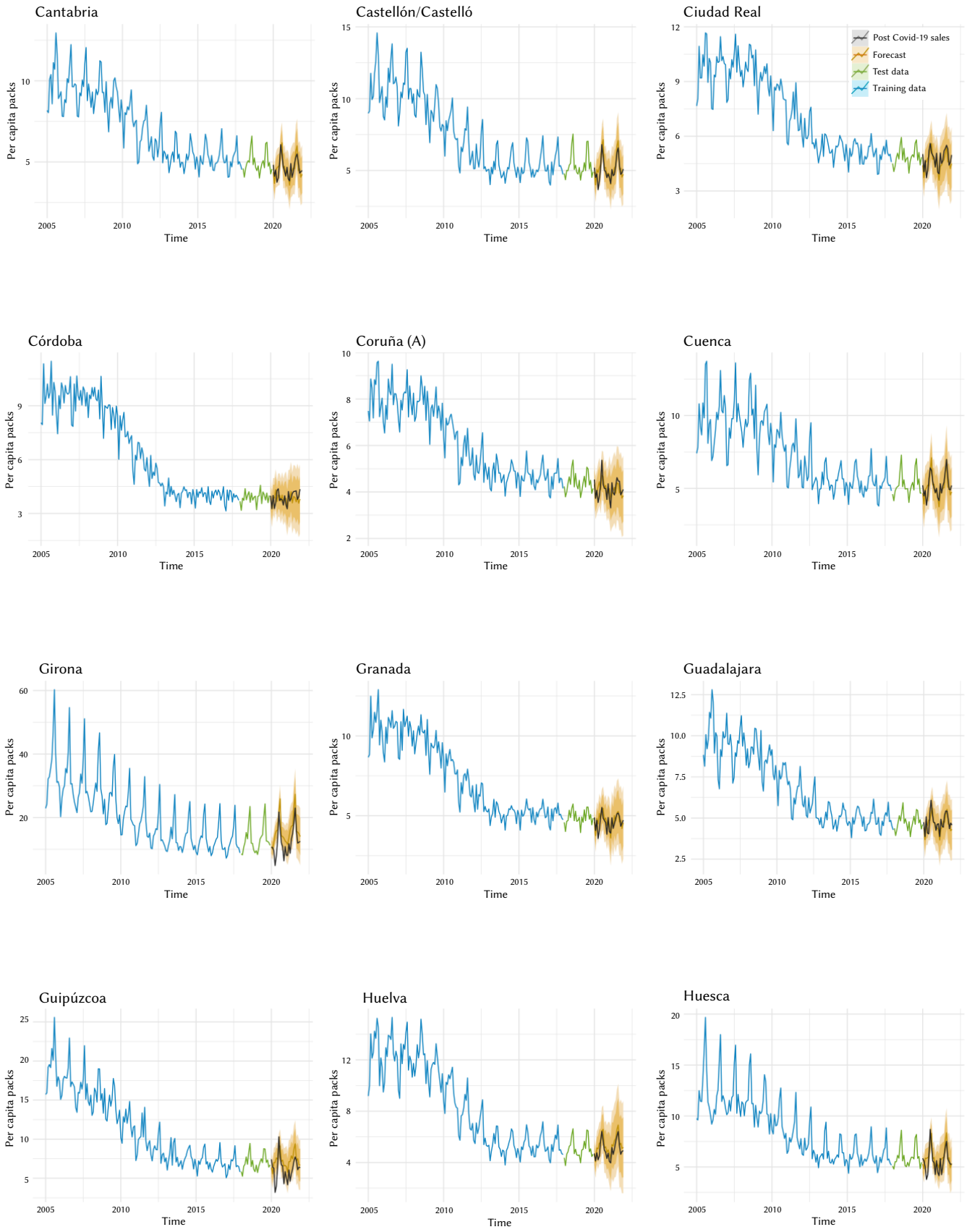

Fig. 1 (B). 2020 and 2021 forecast of the per capita packs based on the best SARIMA models selected.

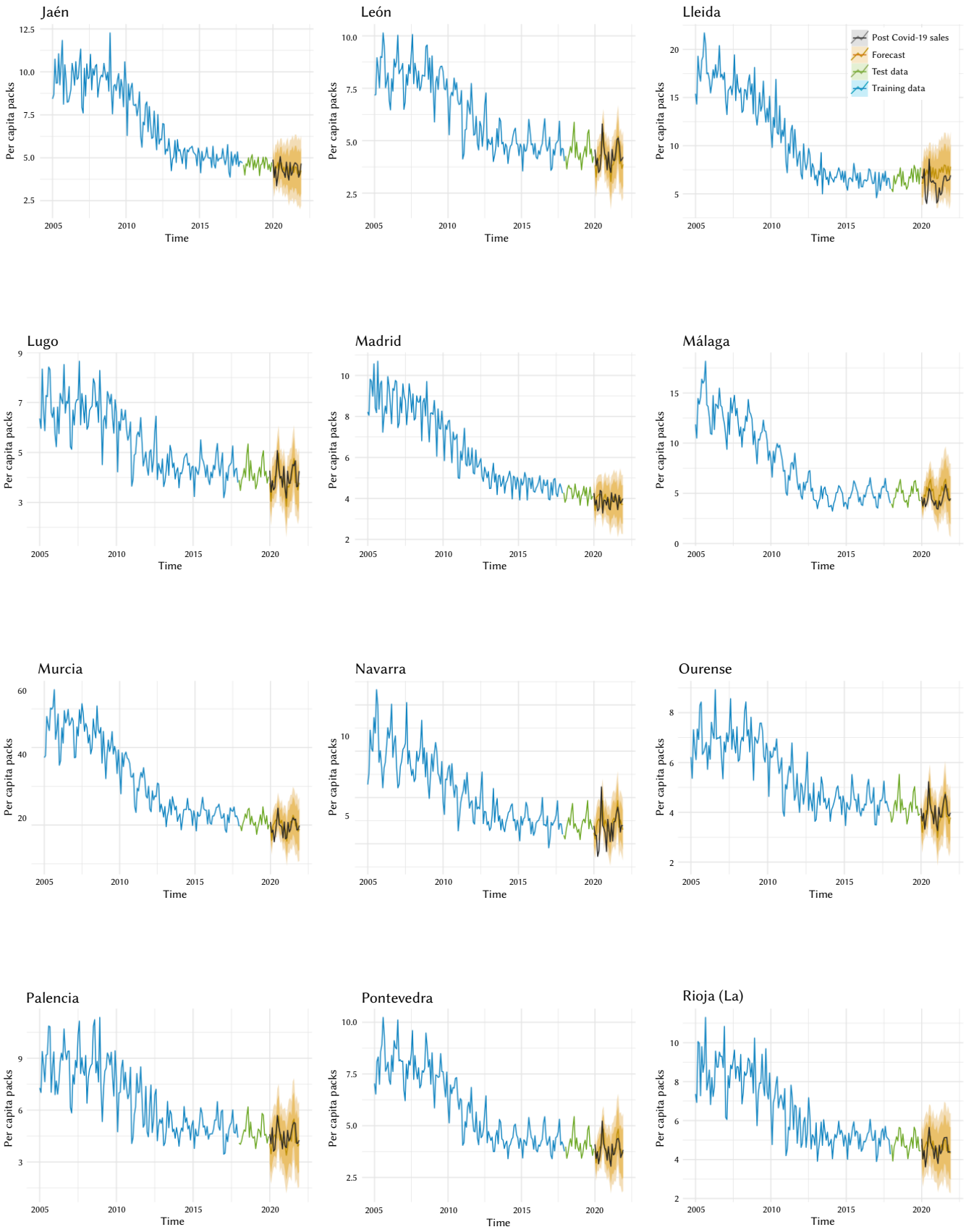

Fig. 1 (C). 2020 and 2021 forecast of the per capita packs based on the best SARIMA models selected.

Fig. 1 (D). 2020 and 2021 forecast of the per capita packs based on the best SARIMA models selected.

TABLE IV. Provincial Impact of COVID-19 on the Spanish Tobacco Market

Province	Gap in Euros (%)			Gap in Per capita packs (%)		
	Min	Max	Mean	Min	Max	Mean
Alava	-18,01	22,02	0,26	-19,00	27,09	2,87
Albacete	-21,68	12,96	-3,04	-5,83	59,23	26,89
Alicante	-38,13	11,16	-14,88	-39,11	8,52	-18,95
Almería	-23,63	10,38	-1,47	-21,65	14,01	0,22
Asturias	-18,21	10,08	-1,15	-15,70	19,49	5,17
Ávila	-26,73	18,05	-0,55	-25,08	15,36	0,29
Badajoz	-21,31	11,16	-0,03	-21,42	13,57	1,41
Balears	-58,08	8,50	-26,50	-58,20	5,41	-25,72
Barcelona	-20,96	5,34	-5,50	-18,91	9,57	-2,35
Burgos	-35,41	16,68	-5,01	-33,38	21,68	-2,26
Cáceres	-28,66	14,05	-2,10	-26,76	16,16	2,08
Cádiz	-8,68	22,96	5,74	-11,37	17,13	2,15
Cantabria	-20,81	12,65	-1,76	-19,17	13,81	-0,09
Castellón	-26,30	11,48	-3,67	-24,73	12,79	-3,13
Ciudad Real	-20,81	13,59	0,86	-19,36	16,78	3,36
Córdoba	-15,72	11,83	0,89	-16,35	17,01	1,72
Coruña (A)	-14,06	10,25	-1,17	-15,98	13,76	0,40
Cuenca	-22,24	15,22	2,42	-23,45	14,14	2,13
Girona	-65,26	9,61	-17,54	-66,74	4,76	-22,71
Granada	-21,43	9,42	-3,10	-20,43	8,63	-2,02
Guadalajara	-20,37	14,15	1,05	-17,27	14,58	3,00
Guipúzcoa	-53,43	22,25	-11,65	-54,48	20,10	-14,98
Huelva	-18,14	16,42	0,37	-18,38	2,95	-7,14
Huesca	-33,40	15,58	-5,25	-35,35	15,17	-6,60
Jaén	-25,28	6,99	-4,10	-23,43	12,57	-0,99
León	-22,22	15,01	-1,02	-16,86	22,07	4,79
Lleida	-47,06	10,75	-16,42	-46,69	11,46	-16,66
Lugo	-16,04	20,94	-0,76	-11,83	23,14	3,83
Madrid	-20,06	5,36	-4,66	-17,69	11,76	-1,33
Málaga	-29,07	8,13	-12,07	-29,53	9,74	-11,41
Murcia	-24,57	10,08	-4,24	-23,27	12,56	-2,45
Navarra	-50,82	32,66	-5,84	-51,38	30,92	-6,74
Ourense	-19,94	18,77	0,57	-14,68	18,05	3,62
Palencia	-22,39	23,28	0,82	-18,49	27,07	4,03
Pontevedra	-16,82	11,58	-0,36	-15,88	13,09	-1,25
Rioja (La)	-24,21	12,10	-3,58	-23,41	15,79	-1,49
Salamanca	-34,98	18,00	-7,76	-30,80	25,68	-3,07
Segovia	-23,38	14,12	0,25	-20,82	17,15	2,15
Sevilla	-3,79	15,71	4,32	-5,41	15,82	3,79
Soria	-24,35	27,71	1,85	-24,04	29,18	3,49
Tarragona	-33,09	8,87	-8,12	-30,66	9,99	-7,53
Teruel	-19,93	16,38	-2,85	-18,87	15,81	-2,27
Toledo	-16,61	12,20	2,59	-16,86	8,75	-3,70
Valencia	-20,97	9,35	-4,94	-22,27	4,08	-5,73
Valladolid	-26,57	6,10	-6,14	-22,44	16,80	0,61
Vizcaya	-17,89	7,13	-3,47	-18,76	6,88	-3,43
Zamora	-25,08	24,26	1,72	-22,25	23,76	4,65
Zaragoza	-23,41	5,65	-5,43	-20,92	11,84	-0,62

V. Conclusions

In this study we have predicted the impact that COVID-19 has had on tobacco sales in Spain (in euros, in packs and in per capita packs) from January 2020 to December 2021, using ARIMA and SARIMA Machine Learning statistical models. Our estimates indicate that the greatest impact of COVID-19 on cigarette sales is observed in tourist provinces and those bordering France, where, in the months of border closures, sales were up to 66.74% lower than the forecast made. On the other hand, in the provinces bordering Gibraltar, the impact of COVID-19 was very slight (5.41%). The reasons why COVID-19 may impact tobacco sales may be public awareness, leisure restrictions, border closures, etc. However, it seems that the greatest impact of COVID-19 has been caused by the closure of borders.

Along these lines, in provinces such as Alicante/Alacant, Baleares (Illes), Girona, Guipúzcoa, Lleida, Málaga and Navarra, a strong impact of COVID-19 on tobacco sales has been observed. In addition, the least impact has been observed in Cádiz and Sevilla. If the national average impact is observed, in Spain COVID-19 has had almost no effect. Specifically, the average provincial impact in Spain is close to -2%. This is because the forecast made with the SARIMA models and Post COVID-19 sales are almost the same in most Spanish provinces.

The results seem to show that the closure of borders has had a marked impact on provincial tobacco sales in Spain. Therefore, it seems that the effect of tourism and cross-border purchases between Spain and France and Spain and Gibraltar have been altered by the border restrictions caused by COVID-19. Based on our predictions and forecasts, policymakers must make the right decisions about the tobacco price differentials observed between European countries where there is constant and abundant cross-border movement. To keep smoking under control, harmonized decisions by all countries must be made.

This work is not without limitations. A recent work reveals that Philip Morris International, the world's leading tobacco manufacturer, is using heated tobacco products (HTPs) to replace the traditional cigarette. The results achieved may be influenced by this phenomenon [48]. In addition, a recent study also indicates that the affordability of cigarettes is a key factor for their demand in Spain. For this reason, part of the "no loss" in Seville and Cádiz may be motivated by the affordability effect [49].

Given the limitations indicated, the lines of future research can be summarized in three. First, it is interesting to analyze whether HTPs are causing part of the gaps detected in this paper. Secondly, it would be important to analyze the role that affordability plays in the gaps detected. Finally, the behavior of substitute products must be analyzed to find out if part of the effects detected in this paper may be due to the consumption of other alternative products.

Appendix

A. Snapshots of the R Script For the Forecasting of the Time-Series Data

```
### STEP 1: Data collection,Import required, Read data
into dataframe, define the variable Per Capita Packs.
library(tidyverse)
library(forecast)
df = read.xlsx("../TFM/tobaccosales.xlsx")

colnames(df) = c("Province","Month","Euros","Packs",
"Year","Population")

df <- data.frame(df)

df$PercapitaPacks = df$Packs/df$Population
```

```
### STEP 2: Create the descriptive statistics table.

estdescriptiv1 = df %>%
  group_by(Province) %>%
  summarise(meanPacks = mean(Packs),
            sdPacks = sd(Packs),
            q1Packs = quantile(Packs, c(0.25)),
            q2Packs = quantile(Packs, c(0.5)),
            q3Packs = quantile(Packs, c(0.75)))
estdescriptiv2 = total %>%
  group_by(Province) %>%
  summarise(meanEuros = mean(Euros),
            sdEuros = sd(Euros),
            q1Euros = quantile(Euros, c(0.25)),
            q2Euros = quantile(Euros, c(0.5)),
            q3Euros = quantile(Euros, c(0.75)))
estdescriptiv3 = total %>%
  group_by(Province) %>%
  summarise(meanPacksperCapita = mean(PacksperCapita),
     sdPacksperCapita = sd(PacksperCapita),
     q1PacksperCapita = quantile(PacksperCapita, c(0.25)),
     q2PacksperCapita = quantile(PacksperCapita, c(0.5)),
     q3PacksperCapita = quantile(PacksperCapita, c(0.75)))

estdescriptiv <- cbind(estdescriptiv 1, estdescriptiv2,
estdescriptiv3)


### STEP 3: Convert data into date-time format and
create the dataset of train, test and post COVID-19
sales and build the model using auto.arima.

timeserielAlava = df %>%
  filter(Province == "Álava")

timeserieAlavaEurosTrain = ts(timeserieAlava[c(1:156),]
$Euros, start = c(2005,01), frequency = 12)

timeserieAlavaEurosTest = ts(serietemporalAlava[c(157:180),
]$Euros, start = c(2018,01) , frequency = 12)

totalAlavaEuros = ts(timeserieAlava[c(1:180),]$Euros,
start = c(2005,01), frequency = 12)

postcovid19AlavaEuros = ts(timeserieAlava[c(181:204),]
$Euros, start = c(2020,01), frequency = 12)

STAlavaEurosTrain = auto.arima(serietemporalAlavaEurosTrain)
STAlavaEurosTest = auto.arima(serietemporalAlavaEurosTest)
predAlavaEuros = forecast(auto.arima(totalAlavaEuros), 24)

### STEP 4. Graphical representation of the time series
(train, test, forecast and post COVID-19 sales).

plotAlavaEuros <-autoplot(timeserieAlavaEurosTrain,
series = "train") +
  autolayer(timeserieAlavaEurosTest, series = "test")+
  autolayer(predAlavaEuros, series = "prediction") +
  autolayer(postcovid19AlavaEuros, series =
"observed")+
  guides(colour = guide_legend(""))+
  labs(x = "Time",
       y = "Euros",
       title = "Alava")+
  scale_color_manual(labels = c("Post Covid-19 sales",
"Forecast", "Test data", "Training data"),
                     values = c("#333333", "#db8100",
"#7fb433", "#0098cd"))+
  theme_minimal()

### STEP 5. Calculate the provincial impact of COVID-19
on the Spanish tobacco market.

impactAlavaEuros <- ((postcovid19AlavaEuros
-predAlavaEuros$mean)/predAlavaEuros$mean)*100
```

B. Others Accuracy Metrics of the ML Models

APPENDIX TABLE I. Selected SARIMA Models for Frecasting Euros

Province	ME	RMSE	MAE	MPE	MASE
Alava	-1,24E-09	4,84E+05	4,06E+05	-8,16E-01	1,39E+00
Albacete	1,48E+04	7,41E+05	5,86E+05	-7,22E-01	1,67E+00
Alicante	1,92E+05	5,40E+06	4,63E+06	-1,03E+00	2,77E+00
Almería	2,52E+04	1,70E+06	1,37E+06	-1,11E+00	3,16E+00
Asturias	-6,21E-09	1,77E+06	1,39E+06	-7,72E-01	2,14E+00
Ávila	5,35E+03	6,19E+05	4,91E+05	-3,01E+00	2,81E+00
Badajoz	-2,48E-09	1,12E+06	9,02E+05	-8,55E-01	2,01E+00
Balears	-1,31E+05	4,06E+06	3,42E+06	-2,50E+00	2,79E+00
Barcelona	-3,48E-08	7,78E+06	6,42E+06	-6,72E-01	1,81E+00
Burgos	-1,86E-09	8,14E+05	6,53E+05	-1,44E+00	2,94E+00
Cáceres	0,00E+00	9,77E+05	7,55E+05	-1,40E+00	2,22E+00
Cádiz	5,29E+04	2,27E+06	1,76E+06	-1,88E+00	1,83E+00
Cantabria	2,38E+03	1,40E+06	1,04E+06	-1,32E+00	2,72E+00
Castellón	2,32E+04	1,73E+06	1,31E+06	-1,71E+00	2,71E+00
Ciudad Real	-2,48E-09	9,32E+05	7,02E+05	-9,58E-01	2,58E+00
Córdoba	-4,97E-09	9,79E+05	7,72E+05	-7,15E-01	1,68E+00
Coruña (A)	-4,97E-09	1,90E+06	1,53E+06	-8,53E-01	2,52E+00
Cuenca	1,04E+04	5,92E+05	4,45E+05	-1,48E+00	2,57E+00
Girona	1,25E+05	1,07E+07	8,59E+06	-6,38E+00	2,39E+00
Granada	-7,45E-09	1,59E+06	1,28E+06	-8,76E-01	2,22E+00
Guadalajara	6,80E+03	4,46E+05	3,65E+05	-7,61E-01	1,96E+00
Guipúzcoa	6,36E+04	2,62E+06	2,12E+06	-1,37E+00	1,70E+00
Huelva	3,21E+04	1,31E+06	1,05E+06	-1,31E+00	2,28E+00
Huesca	1,54E+04	8,19E+05	5,96E+05	-1,77E+00	2,27E+00
Jaén	-3,73E-09	8,11E+05	6,57E+05	-5,22E-01	1,92E+00
León	-1,86E-09	1,07E+06	8,18E+05	-1,37E+00	2,54E+00
Lleida	-1,24E-09	1,33E+06	1,16E+06	-1,48E+00	1,18E+00
Lugo	-1,55E-09	6,57E+05	5,12E+05	-1,14E+00	2,03E+00
Madrid	-2,98E-08	7,38E+06	6,14E+06	-4,82E-01	1,61E+00
Málaga	1,70E+05	4,27E+06	3,71E+06	-1,34E+00	2,76E+00
Murcia	-2,48E-09	2,46E+06	2,02E+06	-6,98E-01	1,76E+00
Navarra	5,65E+04	2,41E+06	1,93E+06	-1,41E+00	2,72E+00
Ourense	-3,10E-10	6,07E+05	4,55E+05	-1,08E+00	2,25E+00
Palencia	7,06E+03	3,80E+05	2,95E+05	-1,10E+00	1,57E+00
Pontevedra	-6,21E-10	1,87E+06	1,41E+06	-1,34E+00	2,28E+00
Rioja (La)	-1,86E-09	5,70E+05	4,62E+05	-9,08E-01	3,15E+00
Salamanca	8,86E+03	7,89E+05	6,23E+05	-1,48E+00	1,91E+00
Segovia	2,00E+03	3,81E+05	2,89E+05	-1,64E+00	2,25E+00
Sevilla	-1,24E-09	2,00E+06	1,58E+06	-7,26E-01	1,14E+00
Soria	4,59E+03	2,20E+05	1,75E+05	-1,24E+00	1,57E+00
Tarragona	8,43E+04	3,09E+06	2,43E+06	-2,37E+00	3,13E+00
Teruel	5,78E+03	4,25E+05	3,33E+05	-1,80E+00	2,10E+00
Toledo	1,64E+04	1,10E+06	8,92E+05	-6,52E-01	1,74E+00
Valencia	-2,48E-09	3,72E+06	3,01E+06	-6,02E-01	1,89E+00
Valladolid	-1,24E-09	7,44E+05	6,17E+05	-6,92E-01	1,60E+00
Vizcaya	-6,21E-09	1,26E+06	1,09E+06	-4,04E-01	1,50E+00
Zamora	-1,09E-09	4,88E+05	3,61E+05	-2,01E+00	2,23E+00
Zaragoza	-7,45E-09	1,40E+06	1,17E+06	-5,70E-01	1,39E+00

APPENDIX TABLE II. SELECTED SARIMA MODELS FOR FORECASTING PACKS

Province	ME	RMSE	MAE	MPE	MASE
Alava	1,76E-14	3,78E-01	3,19E-01	-8,26E-01	1,35E+00
Albacete	9,69E-03	4,86E-01	3,85E-01	-7,16E-01	1,69E+00
Alicante	2,58E-02	7,44E-01	6,36E-01	-1,01E+00	2,97E+00
Almería	8,99E-03	6,42E-01	5,21E-01	-1,11E+00	3,44E+00
Asturias	-1,17E-15	4,12E-01	3,26E-01	-7,59E-01	2,29E+00
Ávila	8,21E-03	9,38E-01	7,44E-01	-3,01E+00	2,82E+00
Badajoz	-1,28E-15	4,30E-01	3,46E-01	-8,53E-01	1,99E+00
Balears	3,70E-02	1,12E+00	9,20E-01	-2,33E+00	2,83E+00
Barcelona	-1,31E-15	3,62E-01	2,98E-01	-6,84E-01	1,62E+00
Burgos	-2,11E-15	5,57E-01	4,48E-01	-1,43E+00	2,93E+00
Cáceres	3,48E-14	6,14E-01	4,77E-01	-1,38E+00	2,27E+00
Cádiz	1,10E-02	4,81E-01	3,72E-01	-1,86E+00	1,87E+00
Cantabria	1,00E-03	5,94E-01	4,44E-01	-1,31E+00	2,64E+00
Castellón	1,04E-02	7,84E-01	5,94E-01	-1,68E+00	2,69E+00
Ciudad Real	-5,52E-14	4,83E-01	3,64E-01	-9,65E-01	2,63E+00
Córdoba	-1,92E-15	3,21E-01	2,53E-01	-7,14E-01	1,70E+00
Coruña (A)	-1,13E-14	4,13E-01	3,32E-01	-8,47E-01	2,51E+00
Cuenca	1,28E-02	7,25E-01	5,44E-01	-1,47E+00	2,58E+00
Girona	4,26E-02	3,63E+00	2,93E+00	-6,20E+00	2,66E+00
Granada	-2,61E-15	4,51E-01	3,58E-01	-8,75E-01	2,23E+00
Guadalajara	6,55E-03	4,50E-01	3,66E-01	-7,54E-01	1,99E+00
Guipúzcoa	2,19E-02	9,24E-01	7,42E-01	-1,35E+00	1,80E+00
Huelva	1,58E-02	6,55E-01	5,25E-01	-1,29E+00	2,48E+00
Huesca	1,73E-02	9,17E-01	6,69E-01	-1,74E+00	2,29E+00
Jaén	3,01E-13	3,27E-01	2,65E-01	-5,16E-01	1,97E+00
León	-5,04E-14	5,40E-01	4,15E-01	-1,36E+00	2,59E+00
Lleida	-2,92E-15	7,98E-01	6,94E-01	-1,48E+00	1,21E+00
Lugo	-1,42E-15	4,61E-01	3,59E-01	-1,13E+00	2,02E+00
Madrid	-1,31E-15	2,93E-01	2,45E-01	-4,94E-01	1,47E+00
Málaga	2,58E-02	6,66E-01	5,81E-01	-1,33E+00	3,38E+00
Murcia	3,57E-13	4,44E-01	3,69E-01	-7,06E-01	1,73E+00
Navarra	2,16E-02	9,57E-01	7,69E-01	-1,41E+00	3,19E+00
Ourense	-1,55E-15	4,52E-01	3,41E-01	-1,06E+00	2,35E+00
Palencia	1,06E-02	5,60E-01	4,35E-01	-1,09E+00	1,57E+00
Pontevedra	-1,02E-15	4,93E-01	3,72E-01	-1,33E+00	2,23E+00
Rioja (La)	-5,18E-16	4,60E-01	3,75E-01	-9,08E-01	2,79E+00
Salamanca	6,41E-03	5,65E-01	4,47E-01	-1,47E+00	1,91E+00
Segovia	3,26E-03	6,05E-01	4,59E-01	-1,63E+00	2,23E+00
Sevilla	-7,77E-16	2,75E-01	2,17E-01	-7,33E-01	1,16E+00
Soria	1,21E-02	5,82E-01	4,62E-01	-1,24E+00	1,54E+00
Tarragona	2,66E-02	9,92E-01	7,81E-01	-2,32E+00	3,16E+00
Teruel	1,06E-02	7,68E-01	6,02E-01	-1,78E+00	2,08E+00
Toledo	6,18E-03	4,19E-01	3,36E-01	-6,47E-01	1,82E+00
Valencia	3,33E-16	3,85E-01	3,13E-01	-5,98E-01	1,98E+00
Valladolid	-2,78E-16	3,56E-01	2,95E-01	-6,88E-01	1,61E+00
Vizcaya	5,18E-16	2,76E-01	2,39E-01	-4,08E-01	1,44E+00
Zamora	1,39E-14	6,40E-01	4,76E-01	-1,99E+00	2,50E+00
Zaragoza	-1,33E-15	3,72E-01	3,11E-01	-5,76E-01	1,42E+00

APPENDIX TABLE III. SELECTED SARIMA MODELS FOR FORECASTING PER CAPITA PACKS

Province	ME	RMSE	MAE	MPE	MASE
Alava	4,57E+02	1,76E-14	3,78E-01	3,19E-01	1,35E+00
Albacete	3,11E+02	9,69E-03	4,86E-01	3,85E-01	1,69E+00
Alicante	3,87E+02	2,58E-02	7,44E-01	6,36E-01	2,97E+00
Almería	2,73E+02	8,99E-03	6,42E-01	5,21E-01	3,44E+00
Asturias	2,95E+02	-1,17E-15	4,12E-01	3,26E-01	2,29E+00
Ávila	3,20E+02	8,21E-03	9,38E-01	7,44E-01	2,82E+00
Badajoz	2,91E+02	-1,28E-15	4,30E-01	3,46E-01	1,99E+00
Balears	5,11E+02	3,70E-02	1,12E+00	9,20E-01	2,83E+00
Barcelona	2,66E+02	-1,31E-15	3,62E-01	2,98E-01	1,62E+00
Burgos	4,78E+02	-2,11E-15	5,57E-01	4,48E-01	2,93E+00
Cáceres	3,13E+02	3,48E-14	6,14E-01	4,77E-01	2,27E+00
Cádiz	2,29E+02	1,10E-02	4,81E-01	3,72E-01	1,87E+00
Cantabria	2,98E+02	1,00E-03	5,94E-01	4,44E-01	2,64E+00
Castellón	3,30E+02	1,04E-02	7,84E-01	5,94E-01	2,69E+00
Ciudad Real	2,92E+02	-5,52E-14	4,83E-01	3,64E-01	2,63E+00
Córdoba	3,02E+02	-1,92E-15	3,21E-01	2,53E-01	1,70E+00
Coruña (A)	2,67E+02	-1,13E-14	4,13E-01	3,32E-01	2,51E+00
Cuenca	4,12E+02	1,28E-02	7,25E-01	5,44E-01	2,58E+00
Girona	6,02E+02	4,26E-02	3,63E+00	2,93E+00	2,66E+00
Granada	2,84E+02	-2,61E-15	4,51E-01	3,58E-01	2,23E+00
Guadalajara	2,96E+02	6,55E-03	4,50E-01	3,66E-01	1,99E+00
Guipúzcoa	4,44E+02	2,19E-02	9,24E-01	7,42E-01	1,80E+00
Huelva	3,33E+02	1,58E-02	6,55E-01	5,25E-01	2,48E+00
Huesca	3,50E+02	1,73E-02	9,17E-01	6,69E-01	2,29E+00
Jaén	3,63E+02	3,01E-13	3,27E-01	2,65E-01	1,97E+00
León	2,72E+02	-5,04E-14	5,40E-01	4,15E-01	2,59E+00
Lleida	5,40E+02	-2,92E-15	7,98E-01	6,94E-01	1,21E+00
Lugo	3,12E+02	-1,42E-15	4,61E-01	3,59E-01	2,02E+00
Madrid	2,61E+02	-1,31E-15	2,93E-01	2,45E-01	1,47E+00
Málaga	3,27E+02	2,58E-02	6,66E-01	5,81E-01	3,38E+00
Murcia	2,55E+02	3,57E-13	4,44E-01	3,69E-01	1,73E+00
Navarra	4,87E+02	2,16E-02	9,57E-01	7,69E-01	3,19E+00
Ourense	2,16E+02	-1,55E-15	4,52E-01	3,41E-01	2,35E+00
Palencia	4,47E+02	1,06E-02	5,60E-01	4,35E-01	1,57E+00
Pontevedra	2,29E+02	-1,02E-15	4,93E-01	3,72E-01	2,23E+00
Rioja (La)	4,12E+02	-5,18E-16	4,60E-01	3,75E-01	2,79E+00
Salamanca	4,52E+02	6,41E-03	5,65E-01	4,47E-01	1,91E+00
Segovia	3,46E+02	3,26E-03	6,05E-01	4,59E-01	2,23E+00
Sevilla	2,55E+02	-7,77E-16	2,75E-01	2,17E-01	1,16E+00
Soria	3,95E+02	1,21E-02	5,82E-01	4,62E-01	1,54E+00
Tarragona	4,21E+02	2,66E-02	9,92E-01	7,81E-01	3,16E+00
Teruel	3,73E+02	1,06E-02	7,68E-01	6,02E-01	2,08E+00
Toledo	3,07E+02	6,18E-03	4,19E-01	3,36E-01	1,82E+00
Valencia	2,92E+02	3,33E-16	3,85E-01	3,13E-01	1,98E+00
Valladolid	3,98E+02	-2,78E-16	3,56E-01	2,95E-01	1,61E+00
Vizcaya	2,73E+02	5,18E-16	2,76E-01	2,39E-01	1,44E+00
Zamora	3,05E+02	1,39E-14	6,40E-01	4,76E-01	2,50E+00
Zaragoza	3,26E+02	-1,33E-15	3,72E-01	3,11E-01	1,42E+00

Acknowledgment

The authors are thanked for the support of Antonio Golpe, full professor at the University of Huelva and professor of Big Data Analysis in the Master's in Business Intelligence at Universidad Internacional de La Rioja (UNIR).

References

[1] I. Papanicolas,L.R. Woskie and A.K. Jha, "Health care spending in the United States and other high-income countries". *Jama*, vol. 319, no. 10, pp. 1024-1039, 2018, doi: 10.1001/jama.2018.1150.

[2] P. Cadahia, A. Golpe, J.M. Martín-Álvarez and E. Asensio, "Measuring anomalies in cigarette sales using official data from Spanish provinces: Are the anomalies detected by the Empty Pack Surveys (EPSs) used by Transnational Tobacco Companies (TTCs) the only anomalies?". *Tobacco Induced Diseases*, vol. 19, no. 98, 2021, doi: 10.18332/tid/143321.

[3] A. Almeida, A. Golpe and J.M. Martín-Álvarez, "A spatial analysis of the Spanish tobacco consumption distribution: Are there any consumption clusters?". *Public Health*, vol. 186, 2020, doi: 10.1016/j.puhe.2020.06.040.

[4] J.M. Martín-Álvarez, A. Golpe, J. Iglesias and R. Ingelmo, "Price and income elasticities of demand for cigarette consumption: what is the association of price and economic activity with cigarette consumption in Spain from 1957 to 2016?", *Public Health*, vol. 185, 275-282, 2020, doi: 10.1016/j.puhe.2020.05.059.

[5] J.M. Martín-Álvarez, A. Almeida, A. Galiano, A. and A. Golpe, "Asymmetric behavior of tobacco consumption in Spain across the business cycle: a long-term regional analysis", *International Journal of Health Economics and Management*, vol. 20, 391-421, 2020, doi: 10.1007/s10754-020-09286-y.

[6] J.M. Martín-Álvarez., A. Almeida, A. Golpe, and J.C. Vides, "The influence of cigarette price on the cigarette consumption in Spain: a Logarithmic Mean Divisia Index analysis from 1957 to 2018", *Revista Espanola de Salud Publica*, vol. 95, 2021, e202102026.

[7] A. Almeida, A. Golpe, J. Iglesias and J.M. Martín-Álvarez, "The price elasticity of cigarettes: new evidence from Spanish regions, 2002–2016", *Nicotine and Tobacco Research*, vol. 23, no. 1, 48-56, 2021, doi: 10.1093/ntr/ntaa131.

[8] R. Fu, A. Kundu, N. Mitsakakis, T. Elton-Marshall, W. Wang, S. Hill and M.O. Chaiton, "Machine learning applications in tobacco research: a scoping review". *Tobacco Control,* vol. 32, no. 1, pp. 99-109, 2023, doi: 10.1136/tobaccocontrol-2020-056438.

[9] K.P., Murphy, "Machine learning: a probabilistic perspective". *MIT press*, 2012.

[10] A.L. Beam and I.S. Kohane, "Big data and machine learning in health care". *Jama*, vol. 319, no. 13, pp. 1317-1318, 2018, doi: 10.1001/jama.2017.18391.

[11] T. Hastie, R. Tibshirani and J.H. Friedman, "The elements of statistical learning: data mining, inference, and prediction". *Springer New York*, vol. 2, pp. 1-758, 2009, doi: 10.1007/978-0-387-21606-5.

[12] J.M. Reps, P.R. Rijnbeek and P.B. Ryan, "Supplementing claims data analysis using self-reported data to develop a probabilistic phenotype model for current smoking status". *Journal of Biomedical Informatics*, vol. 97, pp. 103264, 2019, doi: 10.1016/j.jbi.2019.103264.

[13] P. Mamoshina, K. Kochetov, F. Cortese, A. Kovalchuk, A. Aliper, E. Putin and A. Zhavoronkov, "Blood biochemistry analysis to detect smoking status and quantify accelerated aging in smokers". *Scientific Reports*, vol. 9, no. 1, pp. 1-10, 2019, doi: 10.1038/s41598-018-35704-w.

[14] S. Huda, J. Yearwood and R. Borland, "Cluster based rule discovery model for enhancement of government's tobacco control strategy". *4th International Conference on Network and System Security IEEE*, pp. 383-390, 2010, doi:10.1109/NSS.2010.14.

[15] N. Kim, D.E. McCarthy, W.Y. Loh, J.W. Cook, M.E. Piper, T.R. Schlam and T.B. Baker, "Predictors of adherence to nicotine replacement therapy: Machine learning evidence that perceived need predicts medication use". *Drug and Alcohol Dependence*, vol. 205, pp. 107668, 2019, doi: 10.1016/j.drugalcdep.2019.107668

[16] A. Dumortier, E. Beckjord, S. Shiffman and E. Sejdić, "Classifying smoking urges via machine learning". *Computer Methods and Programs in Biomedicine*, vol. 137, pp. 203-213, 2016, doi: 10.1016/j.cmpb.2016.09.016.

[17] L.N. Coughlin, A.N. Tegge, C.E. Sheffer and W.K. Bickel, "A machine-learning approach to predicting smoking cessation treatment outcomes". *Nicotine and Tobacco Research*, vol. 22, no. 3, pp. 415-422, 2020, doi: 10.1093/ntr/nty259.

[18] K. Davagdorj, J.S. Lee, K.H. Park and K.H. Ryu, "A machine-learning approach for predicting success in smoking cessation intervention". *10th International Conference on Awareness Science and Technology IEEE*, pp. 1-6, 2019, doi: 10.1109/ICAwST.2019.8923252.

[19] A. Singh and H. Katyan, "Classification of nicotine-dependent users in India: a decision-tree approach". *Journal of Public Health*, vol. 27, no. 4, pp. 453-459, 2019, doi: 10.1007/s10389-018-0973-x.

[20] L. Clancy, S. Gallus, J. Leung and C.O. Egbe, "Tobacco and COVID-19: Understanding the science and policy implications". *Tobacco Induced Diseases*, vol. 18., 2020, doi: 10.18332/tid/131035.

[21] Y. Saloojee and A. Mathee, "COVID-19 and a temporary ban on tobacco sales in South Africa: impact on smoking cessation". *Tobacco Control*, vol. 31, no. 2, pp. 207-210, 2022, doi: 10.1136/tobaccocontrol-2020-056293.

[22] B.P. Lee, J.L. Dodge, A. Leventhal, N.A. Terrault, "Retail alcohol and tobacco sales during COVID-19". Annals of internal medicine, vol. 174, no. 7, pp. 1027-1029, 2021, doi: 10.7326/M20-7271.

[23] D. Yach, "Tobacco use patterns in five countries during the COVID-19 lockdown". *Nicotine & Tobacco Research*, vol. 22, no. 9, pp. 1671-1672, 2020, doi: 10.1093/ntr/ntaa097.

[24] P. Driezen, K.A. Kasza, S. Gravely, M.E. Thompson, G.T. Fong, K.M. Cummings and A. Hyland, "Was COVID-19 associated with increased cigarette purchasing, consumption, and smoking at home among US smokers in early 2020? Findings from the US arm of the International Tobacco Control (ITC) Four Country Smoking and Vaping Survey". *Addictive Behaviors*, vol. 129, pp. 107276, 2022, doi: 10.1016/j.addbeh.2022.107276.

[25] S. Asare, A. Majmundar, F. Islami, P. Bandi, S. Fedewa, L.J. Westmaas and N. Nargis, "Changes in cigarette sales in the United States during the COVID-19 pandemic". *Annals of Internal Medicine*, vol. 175, no. 1, pp. 141-143, 2022, doi: 10.7326/M21-3350.

[26] J. Kim and S. Lee, "Impact of the COVID-19 pandemic on tobacco sales and national smoking cessation services in Korea". *International Journal of Environmental Research and Public Health*, vol. 19, no. 9, pp. 5000, 2022, doi: 10.3390/ijerph19095000.

[27] I.B., Ahluwalia, M. Myers and J.E. Cohen, "COVID-19 pandemic: an opportunity for tobacco use cessation". *The Lancet Public Health*, vol. 5, no. 11, pp. e577, 2020, doi: 10.1016/S2468-2667(20)30236-X.

[28] M. Hefler and C.E. Gartner, "The tobacco industry in the time of COVID-19: time to shut it down?". *Tobacco Control*, vol. 29, no. 3, pp. 245-246, 2020, doi: 10.1136/tobaccocontrol-2020-055807.

[29] T.K. Burki, "Tobacco industry capitalises on the COVID-19 pandemic". *The Lancet Respiratory Medicine*, vol. 9, no. 10, pp. 1097-1098, 2021, doi: 10.1016/S2213-2600(21)00361-1.

[30] R. Álvarez, N. Vicente, L. Polo, P. Ríos, P. Ferrández, A.M. Furió, O Monteagudo, R. Dalmau, J. Doncel, S. Justo, J. Rey, C. González and C. Gómez-Chacón, "Tobacco use in Spain during COVID-19 lockdown: an evaluation through social media". *Revista Espanola de Salud Publica*, vol. 95, 2021, PMID: 33724261.

[31] E.P. Esplá, C.C. Faus, A.J. Baldó, I.B. Enrique and E.C. Vives, "COVID-19 and smoking: an opportunity to quit". *Archivos de Bronconeumologia*, vol. 57, no. 12, pp. 784, 2021, doi: 10.1016/j.arbr.2021.10.009.

[32] J.M. Suelves, B. Gomez-Zuniga and M. Armayones, "Changes in smoking behaviour due to the COVID-19 pandemic in Spain". *Tobacco Prevention & Cessation*, vol. 7(Supplement), no. 55, 2021, doi: 10.18332/tpc/143664.

[33] A. Estévez-Danta, L. Bijlsma, R. Capela, R. Cela, A. Celma, F. Hernández and J.B. Quintana, "Use of illicit drugs, alcohol and tobacco in Spain and Portugal during the COVID-19 crisis in 2020 as measured by wastewater-based epidemiology". *Science of the Total Environment*, vol. 836, pp. 155697, 2022, doi:10.1016/j.scitotenv.2022.155697.

[34] C. Martínez-Cao, L. de La Fuente-Tomas, I. Menéndez-Miranda, A. Velasco, P. Zurrón-Madera, L. García-Álvarez and J. Bobes, "Factors associated with alcohol and tobacco consumption as a coping strategy to deal with the coronavirus disease (COVID-19) pandemic and lockdown in Spain". *Addictive Behaviors*, vol. 121, pp. 107003, 2021, doi: 10.1016/j.addbeh.2021.107003.

[35] J.C. Vázquez, J. C. and D. Redolar-Ripoll, "COVID-19 outbreak impact in

Spain: A role for tobacco smoking?". *Tobacco Induced Diseases*, vol. 18, no. 30, 2020, doi: 10.18332/tid/120005.

[36] J.C. Vázquez and D. Redolar-Ripoll,"Epidemiological data from the COVID-19 outbreak in Spain for the promotion of tobacco smoking cessation policies". *Tobacco Use Insights*, vol. 13, pp. 1179173X20924028, 2020, doi: 10.1177/1179173X20924028.

[37] O. Parra, C. Suárez and L. Martínez, "Análisis comparativo de las técnicas de series de tiempo ARIMA y ANFIS para pronosticar tráfico Wimax", *Ingeniería*, vol. 12, no. 2, 73-79, 2007, doi: 10.14483/23448393.2166.

[38] R.A. Yaffee, and M. McGee, "An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®". *Elsevier*, 2000.

[39] E.P. Box George, M. Jenkins Gwilym, C. Reinsel Gregory and M. Ljung Greta, "Time series analysis: forecasting and control". *San Francisco: Holden Bay*, 1979.

[40] M.S.M. Kasihmuddin, M.A. Mansor, S. A. Alzaeemi and S. Sathasivam, "Satisfiability logic analysis via radial basis function neural network with artificial bee colony algorithm". *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2021, doi: 10.9781/ijimai.2020.06.002.

[41] A. Gupta, K. Ghanshala and R.C. Joshi, "Machine Learning Classifier Approach with Gaussian Process, Ensemble boosted Trees, SVM, and Linear Regression for 5G Signal Coverage Mapping". *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 6, No 6, 2021, doi: 10.9781/ijimai.2021.03.004.

[42] R Core Team, "R: A language and environment for statistical computing". *R Foundation for Statistical Computing, Vienna, Austria*, 2021, https://www.R-project.org/

[43] RStudio Team. "RStudio: Integrated Development for R". *RStudio, PBC, Boston, MA URL*, 2020, http://www.rstudio.com/

[44] H. Wickham, M. Averick, J. Bryan, W. Chang, L.D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T.L. Pedersen, E. Miller, S.M. Bache, K. Müller, J. Ooms, D. Robinson D, D.P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, "Welcome to the tidyverse." *Journal of Open Source Software*, vol. 4, no. 43, pp. 1686, 2019, doi: 10.21105/joss.01686.

[45] R.J. Hyndman, Y. Khandakar, "Automatic time series forecasting: the forecast package for R". *Journal of Statistical Software*, vol. 27, no. 3, pp. 1–22, 2008, doi: 10.18637/jss.v027.i03.

[46] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, F. Yasmeen, "forecast: Forecasting functions for time series and linear models". *R package version 8.16*, 2022, https://pkg.robjhyndman.com/forecast/

[47] R. Gomajee, H. Torregrossa, C. Bolze, M. Melchior and F.E.K. Lesueur, "Decrease in cross-border tobacco purchases despite intensification of antitobacco policies in France". *Tobacco Control*, vol. 30, no. 4, pp. 428-433, 2021, doi: 10.1136/tobaccocontrol-2019-055540.

[48] A. Golpe, J.M. Martín-Álvarez, A. Galiano and E. Asensio, "Effect of IQOS introduction on Philip Morris International cigarette sales in Spain: a Logarithmic Mean Divisa Index decomposition approach", *Gaceta Sanitaria*, vol. 36, 293-300, 2022, doi: 10.1016/j.gaceta.2021.12.007.

[49] P. Cadahia, A. Golpe, J.M. Martín-Álvarez, E. Asensio, "The importance of price, income, and affordability in the demand for cigarettes in Spain", *Addicta: The Turkish Journal on Addictions*, vol. 9, no. 3, 241-251, 2022, doi: 10.5152/ADDICTA.2022.22054.

Andoni Andueza

Andoni Andueza is a Graduated in Business Administration by University and Msc in Business Intelligence by Universidad Internacional de La Rioja (UNIR). His master's thesis focused on the analysis of time series. He has developed different dashboard for any companies for business analysis. He has worked in University of Mondragon (MU) as People Analyst. Currently, he is People Analyst at ULMA.

Miguel Ángel Del Arco-Osuna

Miguel Ángel Del Arco-Osuna is a Telecommunication Engineer by University of Seville (US) and Graduated in Business Administration by Distance Learning University (UNED). He is currently professor in Quantitative Methods for Economics and Business at Universidad Internacional de La Rioja (UNIR) and PhD Candidate in Economics.

Bernat Fornés

Bernat Fornés is a Graduated in Business Administration by University of Vic (UVic-UCC) and Msc in Business Intelligence by Universidad Internacional de La Rioja (UNIR). For the Graduated in Business Administration, his final project was related to the analysis of different areas of Fintech. His master's thesis focused on the analysis of time series. He has developed different dashboard for any companies for business analysis. Currently, he is treasurer for a local council.

Rubén González Crespo

Dr. Rubén González Crespo has a PhD in Computer Science Engineering. Currently he is Vice Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is member from different committees at ISO Organization. Finally, He has published more than 200 papers in indexed journals and congresses.

Juan Manuel Martín-Álvarez

Dr. Juan Manuel Martín-Álvarez has a Phd in Economics with focus in quantitative analysis for decision making. He has extensive experience as a teacher in public and private universities in the areas of Accounting, Finance, Statistics and Econometrics. He is Associate Professor in Quantitative Methods for Economics and Business at Universidad Internacional de La Rioja (UNIR). Currently, he is head of the Msc in Business Intelligence at Universidad Internacional de La Rioja. Finally, he has published more than 10 Health Economics papers with special focus on tobacco use in indexed journals.