uc3m | Universidad Carlos III de Madrid

Master Degree in Computational Social Science
Academic Year 2023-2024

Master Thesis

"Combating Food Waste in Spain's Autonomous Communities: The Deployment of Machine Learning Models to Predict Foodstuff Consumption"

Rianne Nienke Visscher

Alejandro Llorente Pinto

July 2024, Madrid

AVOID PLAGIARISM

The University uses the Turnitin Feedback Studio program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

ABSTRACT

Food waste is a significant global issue, contributing to environmental degradation and economic inefficiency. This thesis aims to address this problem by applying machine learning models to predict foodstuff consumption in Spain's autonomous communities, focusing on enhancing inventory management and reducing food waste. The study leverages two primary methods: a Global Model incorporating Decision Tree, Random Forest and Prophet algorithms, as well as a finely-tuned Prophet model. The Global Model, while computationally efficient and capable of providing robust generalizations across multiple time series, exhibited limitations in precision, particularly with the Prophet algorithm. In contrast, the singular Prophet model, although computationally expensive, delivered significantly better predictions, indicated by lower Mean Absolute Percentage Errors (MAPEs). The findings suggest that these models can be utilized for effective forecasting, though improvements in precision and model scalability are necessary.

Keywords: *Machine Learning, Spain, Food Waste, Foodstuff Consumption*

Contents

List of Figures

List of Tables

TABLE OF DEFINITIONS

Store Type	Description
Tradicional	Typical neighborhood shops not associated with large food & beverage distribution companies.
Cash & Carry	Larger self-service shops catering to bulk purchases, often targeted at businesses.
Droguerías	Convenience stores offering a variety of items including snacks, beverages, and essentials.
Hipermercado	Large retail stores combining supermarket and department store offerings under one roof.
Horeca	Hospitality sector including bars and restaurants.
Others	Miscellaneous small stores not classified into the above categories, such as specialty shops.

1. INTRODUCTION

Each year enormous amounts of food are being wasted all around the world. According to Eurostat (2023), around 131 kilograms of food is wasted per European inhabitant. All these kilos of wasted food is problematic for the planet and its population. Firstly, food waste is highly precarious in achieving nutrition security for millions of malnourished adults and children around the world. Of course wasting less food in the rich western world would not immediately feed the needy in low income countries. However, it would reduce the stress food production puts on scarce natural resources such as land, water and even biodiversity (Bagherzadeh, Inamura & Jeong, 2014). Another grande problem of food waste is that this food ends up at the landfills where it produces 3.3 billion tons of greenhouse gases making food waste the winning third in the competition of highest emission of greenhouse gases, right after the United States and China (FOA, 2013). Moreover, if the food supply chain would be more efficient, this could significantly reduce the costs of foodstuffs benefiting both businesses and consumers.

Food security and poverty are persisting issues in Spain. The financial crisis of 2008 caused many Spaniards to live in poverty causing food insecurity. Just when the standard of living seemed to be improving, the Covid-19 pandemic hit the country hard with its negative economic impact causing both food security and poverty to worsen again (Human Rights Watch, 2022). A study by Moragues Faus et al. (2022) reported that 13.3% of the population in Spain does not have consistent access to sufficient food in both quantity and quality. In other words, a reduction food prices would greatly assist in increasing the food security in Spain.

To maintain food security and limit the strain food production imposes on our planet with the world's population estimated to have grown to 9.7 billion people in 2050 (UN, n.d.), we need to become more frugal with our natural resources. We need to stop food waste. Lipinski et al. (2013, p.1), define food waste as: " […] food that is of good quality and fit for human consumption but that does not get consumed because it is discarded—either before or after it spoils." Also, Lipinski et al. (2013) propose five different stages in which both food waste can occur which are: 1) production, 2) handling and storage, 3) processing and packaging, 4) distribution and marketing, and 5) consumption. The aim of this paper is to investigate methods to reduce food waste in the distribution and marketing stage, which entails the discard of edible food due to it not having been

sold before the ‘use-by’ and ‘best before’ date or it not living up to the aesthetic quality standards, as well as food being wasted due to over purchasing by consumers, restaurants and caterers. Note that in this thesis the food waste in the consumption only encompasses restaurants. Precise estimates of future consumption can aid in reducing food waste as it will allow businesses to be better informed of the quantities of products they should buy at the retailer. In other words, with better estimates, food waste in the distribution and marketing stage and in the consumption stage can be reduced.

The goal of this thesis is to train machine learning models that will allow predicting the consumption of foodstuffs for different autonomous communities, sales channels and product categories in Spain. Large food businesses are already utilising machine learning to predict their future sales of foodstuffs. An often used method is the Oracle’s solution for Retail Demand Forecasting. However, softwares as such, are not easily manageable and require intensive customization which usually has to be done by consultancy firms (Tsoumakas, 2019). With this thesis, the author wishes to make a start on training machine learning models that can later be utilised as an informing tool for smaller businesses around Spain, ones that do not have the means for obtaining expensive software. However, one should note that due to the scope of this thesis as well as the quality of the data available, models and modelling techniques will only be investigated to provide insights for the future development of machine learning as such.

2. THEORETICAL FRAMEWORK

2.1 Machine Learning Models

Two different types of modelling strategies are applied to train the machine learning models. First the three machine learning algorithms: Decision Tree, Random Forest and Prophet which will be utilised in this paper are discussed. Then, the Global Modelling strategy from the ModelTime package is presented.

2.1.1 Decision Tree

The first machine learning algorithm trained in the global model is the Decision Tree. This model has been selected as Jeyarangani et al. (2023) showed that this algorithm performed best in predicting supermarket sales. A decision tree is a machine learning model used for both classification and regression that splits data into branches based on feature values, forming a tree-like structure. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a decision or prediction (Safavian & Landgrebe, 1996).

2.1.2 Random Forest

A random forest is an ensemble learning method used for classification and regression, which operates by constructing multiple decision trees during training. The output of a random forest is either the mode of the classifications (for classification tasks) or the mean prediction (for regression tasks) of the individual trees. It improves accuracy and controls over fitting by averaging multiple decision trees, thereby reducing variance and improving generalisation (Breiman, 2001). The Random Forest algorithm has not been widely used to predict foodstuff consumption. However, Ahmad, Mourshed and Rezgui (2017) did show that this algorithm can be effective in predicting energy consumption. As these time series are rather similar to the foodstuff time series, it can be expected that the Random Forest algorithm will perform well in predicting foodstuff consumption.

2.1.3 Prophet

The Prophet algorithm developed by Facebook is a machine learning model which can be applied to perform time series forecasting. It includes a decomposable time series model consisting of three main components: trend, seasonality and holidays which are integrated in the following equation: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$. The $g(t)$element captures the trend function, modelling the non-periodic changes in the time series values, $s(t)$ is the periodic change function and captures for example weekly or yearly seasonality, and $h(t)$ captures the holiday effects which are multi-day irregular occurrences. The element ϵ_t is the error term and accounts for idiosyncratic changes that are not captured by the model and is assumed to be normally distributed. The model is highly flexible as it allows for the accommodation of seasonality with multiple periods as well as the integration of different trend assumptions. An advantage compared to the more traditional models is that Prophet allows for measurements with irregular time spacings and handles missing values (Taylor & Letham, 2018). This model has not been widely applied in the prediction of foodstuffs which leaves a opportunity for this thesis to explore the functioning of this model in the prediction of foodstuffs.

2.1.4 Global Modeling

Global Forecasting (Multi-Series Modeling) involves developing a single predictive model that simultaneously considers multiple time series. This approach aims to capture the underlying patterns common across the series, reducing the impact of noise from individual series. It is computationally efficient, easier to maintain, and can provide robust generalisations across different time series, although it may compromise some detailed insights specific to each individual series (ModelTime, nd). To perform this global modelling, the ModelTime package in R is a tidy framework for time series forecasting and machine learning. It simplifies the process of building, tuning, and evaluating models, offering a unified interface and extensive model selection. It supports automatic pipelines, ensemble techniques, and provides interpretable results for effective time series analysis and forecasting. Three machine learning algorithms, Random Forest, Decision Tree and Prophet are trained within the global model.

2.2 Data attributes

The following attributes are of importance for informing the choice of machine learning algorithms.

2.2.1 Seasonality

Seasonality is crucial for time series prediction as it captures regular patterns that recur at specific intervals, such as daily, weekly, or yearly cycles. Identifying these patterns enhances the accuracy of forecasting models. Where, Prophet is designed to handle seasonality automatically, providing robust forecasts in the presence of recurring trends, Random Forest and Decision Tree algorithms can be configured to account for seasonality by incorporating time-related features.

2.2.2 Stationarity

Furthermore, according to Ramasubramanian and Singh (2017, pp.600): "Stationarity means that the autocorrelation for any particular lag remains the same regardless of time". A time series is stationary when: 1) The mean does not change over time, 2) The variance in constant, that is, the variability remains constant over time and 3) There is constant auto-correlation, or in other words, the relationship between values at different time lags is consistent over time. To measure this stationarity, the Dickey-Fuller test for stationarity will be applied. A simple Auto-regressive model (AR) can be defined as follows:

$$y_t = \rho y_{t-1} + u_t y_t = \rho y_{t-1} + u_t$$

Where y_t is the variable of interest, and in this case the equivalent units, t is a time index, in this case months with t_0 at 01-01-2018, ρ is a coefficient and u_t the error term. When $\rho = 1$, a unit root is present and the time series is not stationary.

2.2.3 ACF and PACF

The ACF diagnostics represent the autocorrelation at different lags. The definition of the autocorrelation (ACF) between time s and t is as follows:

$$R(s,t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s}$$

(Ramasubramanian & Singh, 2017, pp.604)

The PACF is the partial autocorrelation function and is suitable for understanding the relationship between observations in a time series with their lagged values, after removing the effects of intermediate lag. In other words, "Given a time series z_t, the partial autocorrelation of lag k, denoted by $\alpha(k)$, is the autocorrelation between z_t and $z_t + z_{t+k}$ with the linear dependence of z_t and z_{t+1} through z_{t+k-1} removed (Ramasubramanian & Singh, 2017, pp.608).

3. METHODOLOGY

3.1 The Data

The data utilised in this paper is longitudinal data ranging from January 2018 until November 2019. The source of this data could not be disclosed by the thesis supervisor who provided the data. The data set contains information on the quantity of equivalent units sold per product type in 9 Autonomous Communities in Spain: Extremadura, Castilla - La Mancha, Castilla y León, Cataluña, Balears, Illes, Comunitat Valenciana, Andalucía, País Vasco and Galicia trough 6 different sales channels: Tradicional, Cash & Carry, Droguerías, Hypermarket, Others (small stores) and Online. As the data for the channels 'Cash and Carry' and 'Other', were of low quality, these two sales channels have been deleted from the data set. Moreover, the products have been aggregated into product categories to reduce the number of time series and enhance the quality of the data for each time series as this aggregation causes each time series to contain more data points. Only product categories which belong to foodstuffs were retained. The primary data set was enriched by including longitudinal data obtained from the Instituto Nacional de Estadistica to enhance the predictive performance including: Gross Domestic Product per autonomous community, Production aggregation of the 1) wholesale and retail trade; 2) repair of motor vehicles and motorcycles; 3) transportation and storage; and 4) accommodation and food service activities as well as the Consumer Price Index of 1) food, 2) non-alcoholic beverages and 3) alcoholic beverages. As the variables GDP per autonomous community as well as production aggregation are yearly numbers, these two variables have been divided by 12 to convert the variables into monthly data, assuming that there are no severe differences in these variables between months. As these variables contain data until 2022, the NA values in the year 2023 are imputed through interpolation using the 'imputeTS' library. For each combination of Autonomous Community, sales channel and product category, an ID was generated to easily distinguish the different time series (each ID is one time series) and prepare the data for the global modelling as this method requires panel data. A data frame containing each possible combination of time point and ID was created and merged with the original data to detect missing time points in the time series. Time series with more than 10% missing time points were omitted from the analysis. Then, the missing time points in the data for the global model were imputed again through interpolation.

3.1.1 Descriptive Statistics

Figure 1 shows the sales in equivalent units per product type. For some of the product categories, there are clear trends in when the sales increase and decrease such as for example the *non-alcoholic drinks.* However, for other product categories the trends do not appear to be as clear. The trends within each of the Autonomous Communities are rather similar (figure 2). Figure 3 visualises the equivalent units sold of each product type by Autonomous Community. One should note that the number of equivalent units vary greatly but that the trends of each product category seem to be rather similar for each Autonomous Community. Figure 4 shows The aggregation of sold equivalent units by sales channel. Most of the products are sold through the sales channel *tradicional.* Depending on the Autonomous Community, the number of sales show clear seasonality or not. However, as the high sales of this channel drive up the scale on the y-axis, the graphs for the remaining sales channels are not interpretable. Therefore, figure 5 presents the aggregation of sold equivalent units by sales channel, excluding the channel *tradicional.* One should note there is a severe disturbance in the trends for the sales channel *horeca* (hospitality) in the year 2020 and the year 2021 as a result of the Covid-19 lockdowns.The external variable Consumer Price Index for the product categories *food, alcoholic beverages, non-alcoholic beverages* and *personal care products* is visualised in figure 6. In each Autonomous Community a stagnation in the upward trend is visible in the second half of 2021, after which there was a steep increase starting in 2022 with the steepest increase for *food.*

Figure 7 shows the GDP per Autonomous Community from 2018 until 2022. The trends seem to be similar for all Autonomous Communities. There was a drop in GDP in the year 2020 with Islas Baleares showing a more significant drop in GDP compared to the other Autonomous Communities. Figure 8 shows the production aggregation of combining all Autonomous Communities in Spain. In line with figure 7, this graph shows a steep decrease in production in the year 2020. The production aggregation in 2022 is higher compared to before the pandemic.

3.1.2 Seasonality

The seasonality plots (see annex A) show strong fluctuations in the number of equivalent units sold varying by months. On the other hand, the number of equivalent units sold does not seem to vary

Figure 1

Time Series Equivalent Units by Product

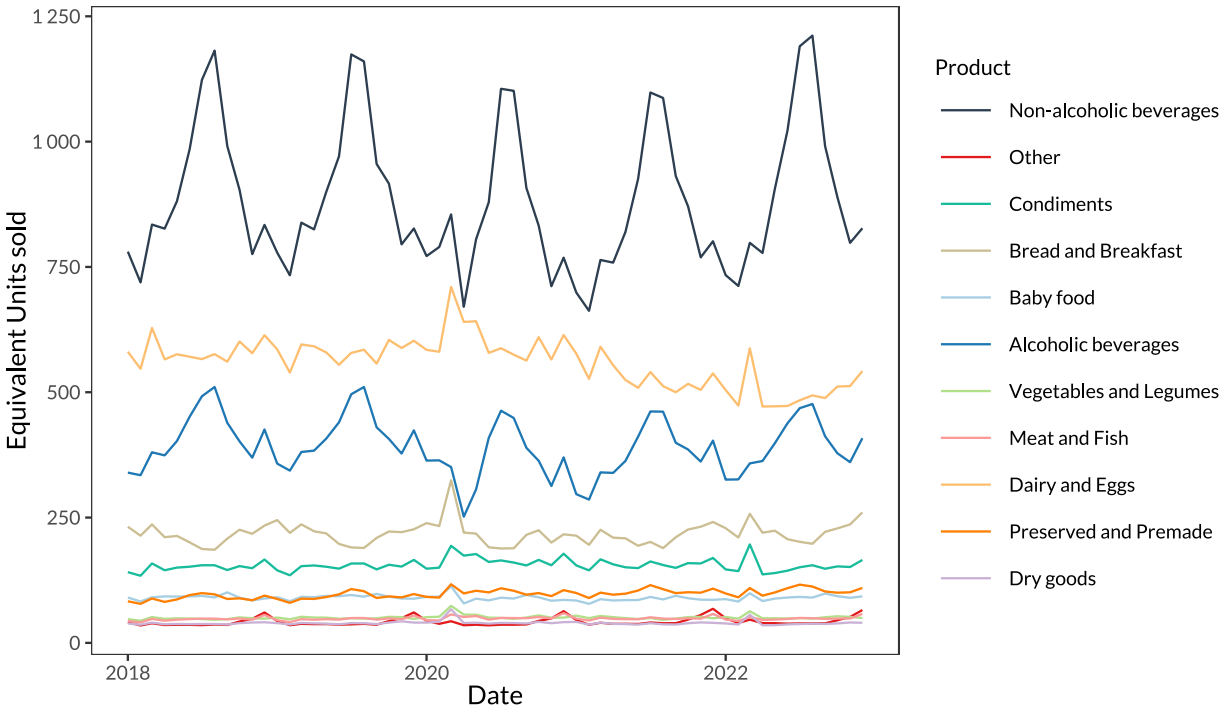


Figure 2

Time Series of the Aggregated Equivalent Units by Autonomous Community

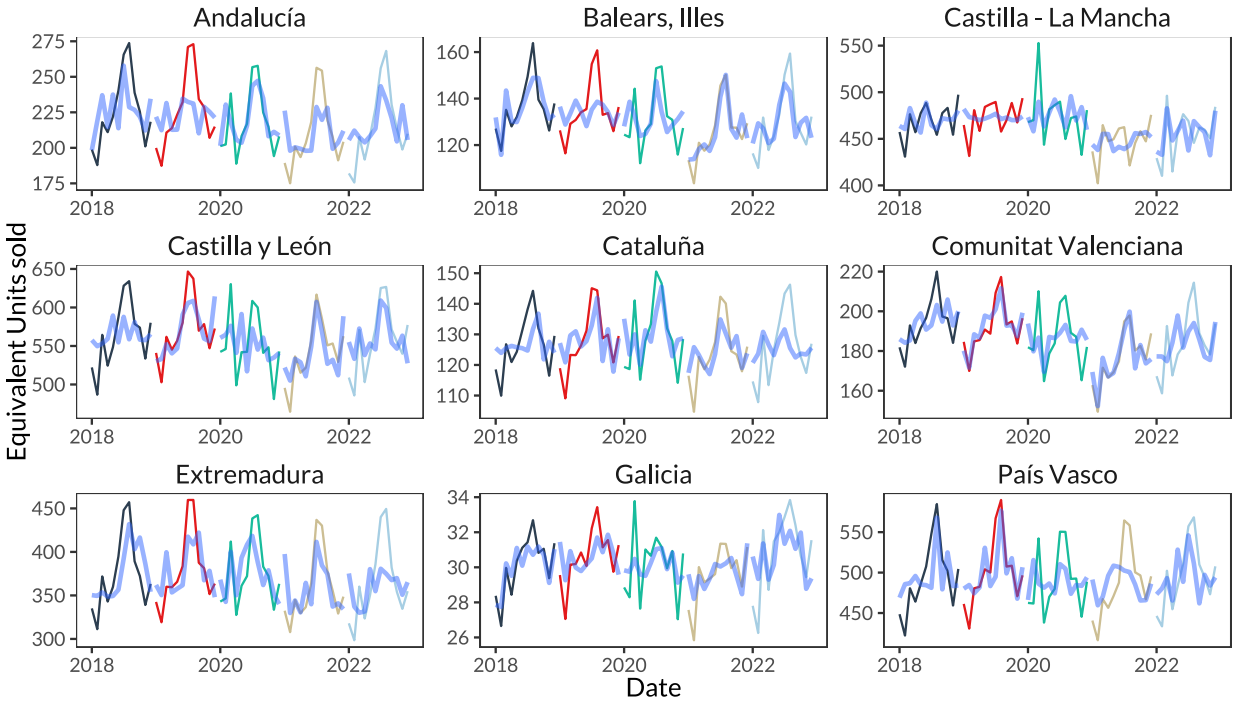


Figure 3

Time Series Equivalent Units by Product and Autonomous community

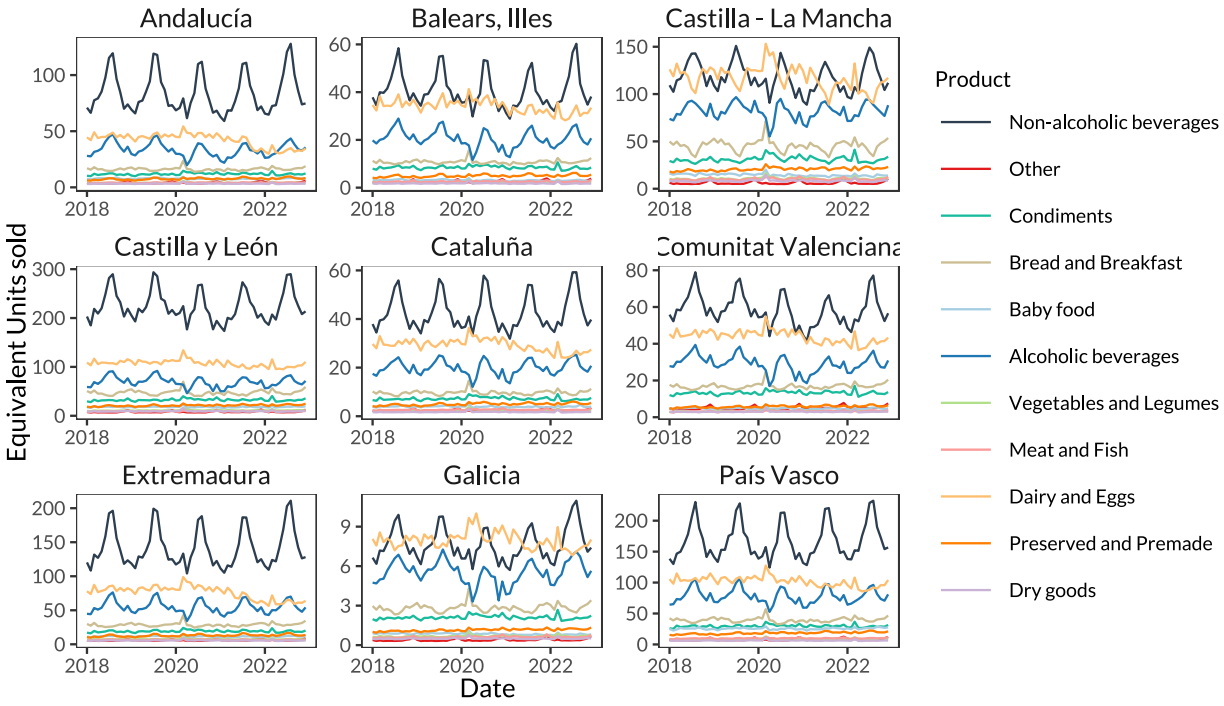


Figure 4

Time Series Equivalent Units by Sales Channel

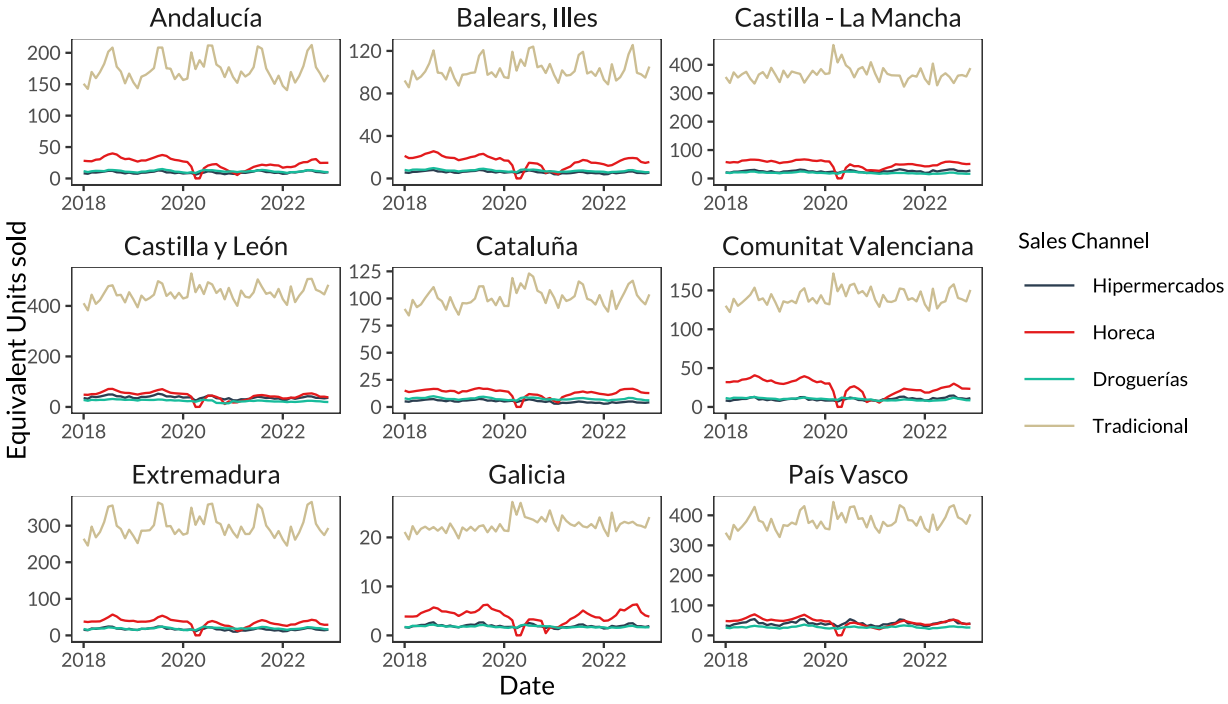


Figure 5

Time Series Equivalent Units by Sales Channel (excluding tradicional)

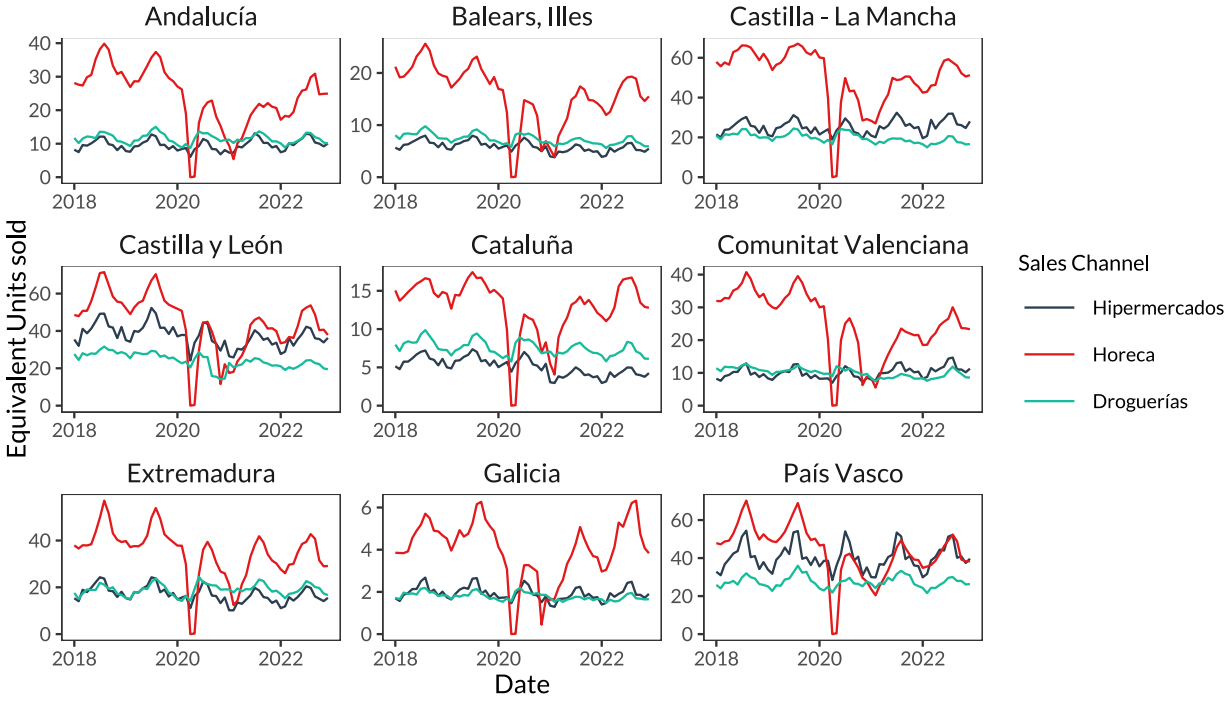


Figure 6

Consumer Price Index per Product Type for each Autonomous Community

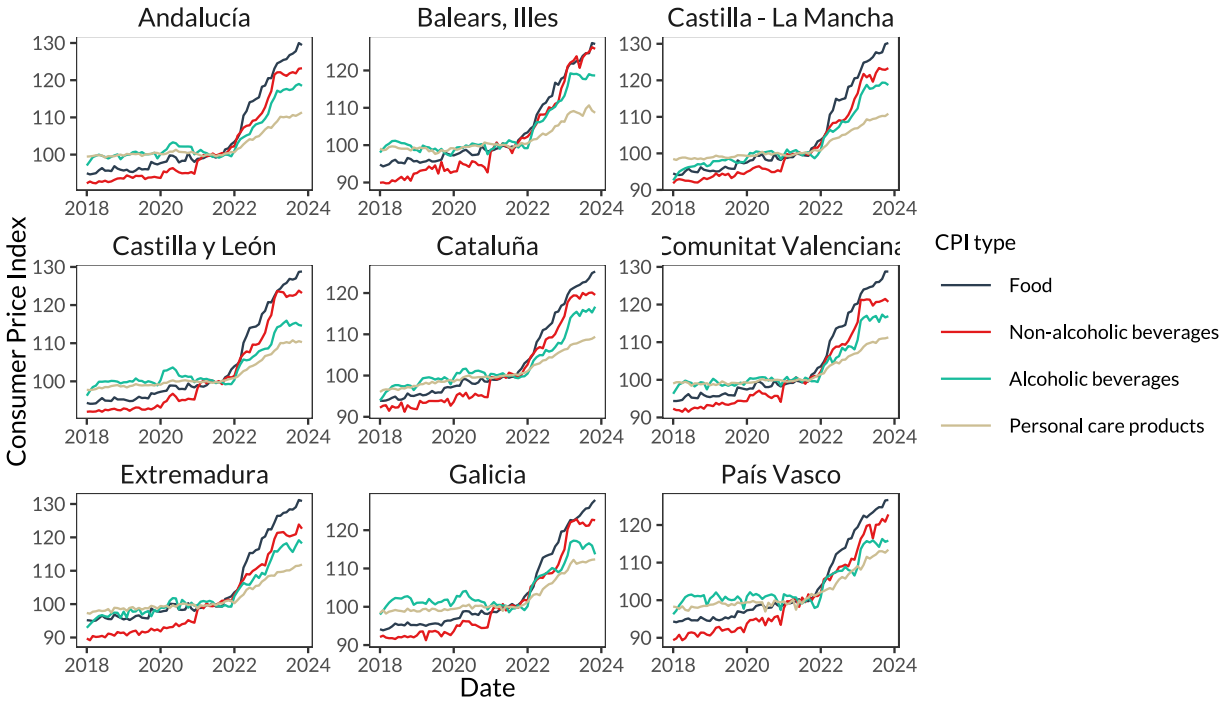


Figure 7

GDP per Autonomous Community

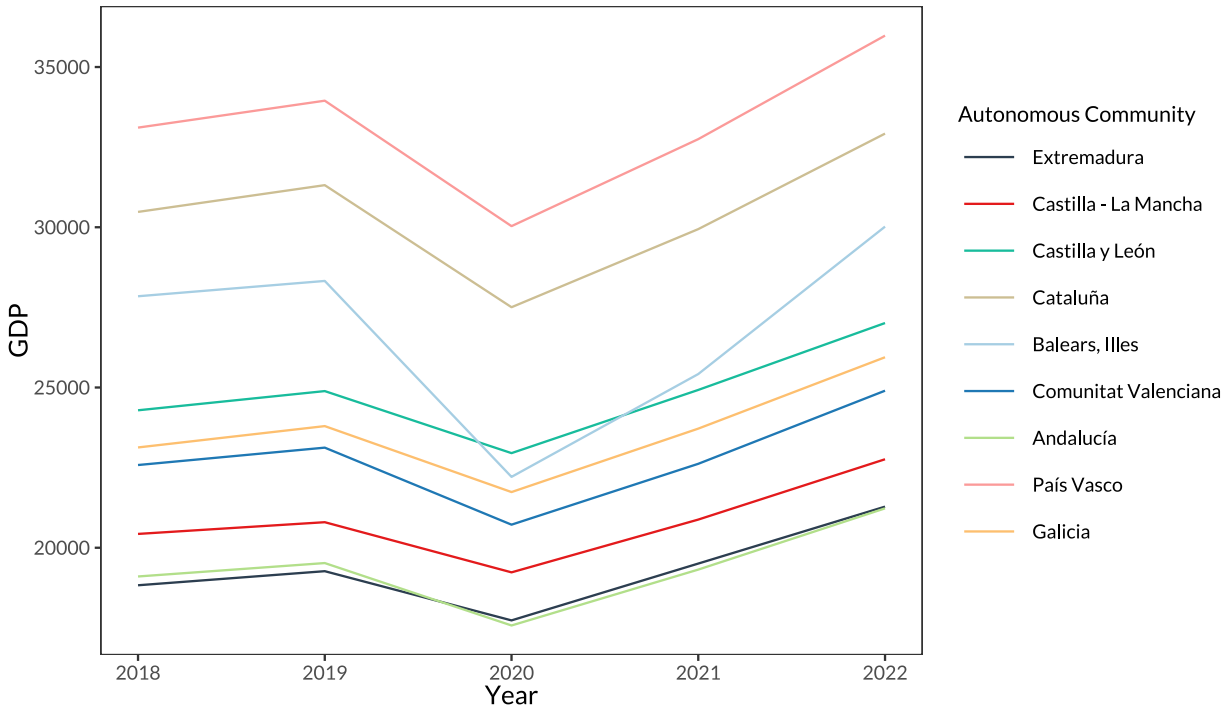
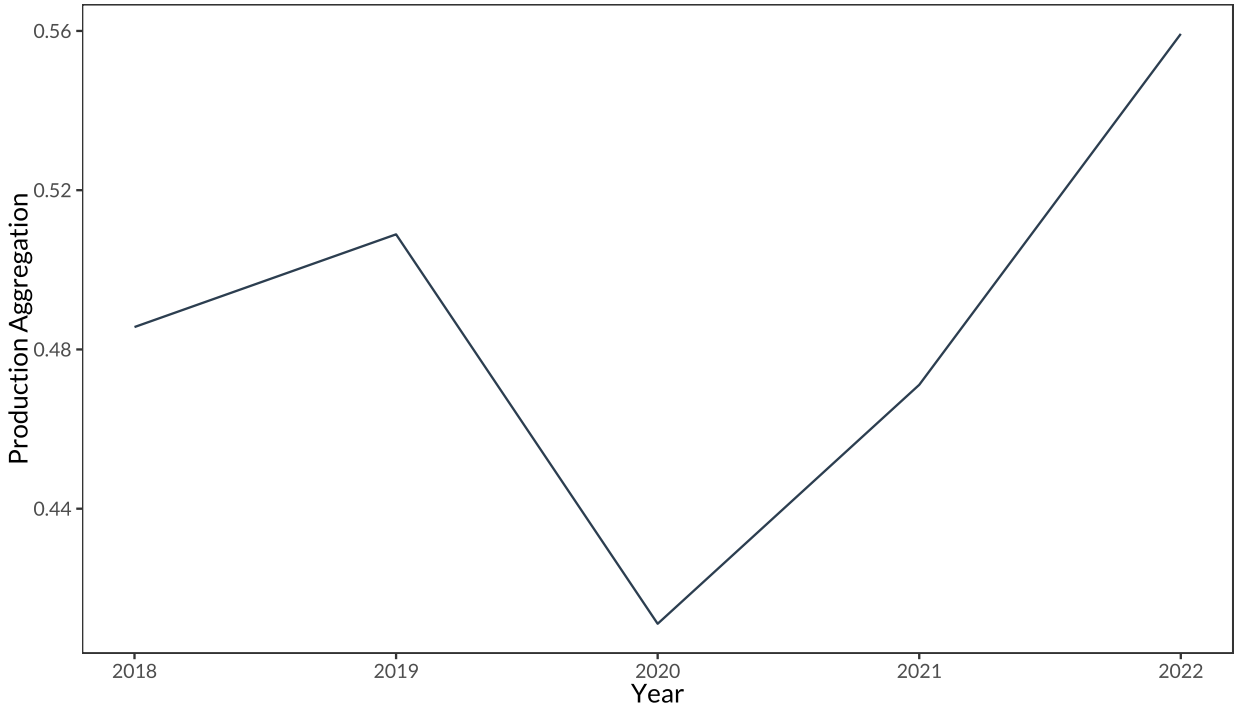


Figure 8

Time Series Production Aggregation



strongly over the years. However, the interquartile range did increase significantly in the year 2021 entailing that the differences in number of equivalent units sold in 2021 varied more strongly over the months in this year compared to the other years (see annex B)

3.1.3 Stationarity

The stationarity was computed for the aggregation of sold equivalent units. The null-hypothesis for the Dickey-fuller test is: h0: $\rho = 1$ (a unit root is present in the time series). Three types of the Dickey-fuller test are computed: 1) Test for a unit root, 2) Test for a unit root with drift, and 3) Test for a unit root with drift and deterministic time trend. All three tests were statistically significant ($p<0.01$) for each Autonomous Community. In other words, the time series of aggregated equivalent units is stationary and therefore suitable for time series analysis. Also, when inspecting the stationarity of sold equivalent units by product category, the time series seemed to be stationary.

3.1.4 ACF and PACF statistics

Both ACF and PACF seem to be rather low (see annex C). For this reason the Prophet algorithm is a suitable model as it is robust to low autocorrelation scenarios as it decomposes time series into trend, seasonality and holiday effects. Also, Random Forest and Decision Trees are applicable algorithms as they can handle non-linear relationships.

3.2 The Global Model

The global model is computed using the ModelTime package in R and includes the Decision Trees, Random Forest and Prophet algorithms.

3.2.1 Data preprocessing and splitting

First, for each ID (each time series), a future data frame was computed which was filled with NA for 12 months. The original data set was splitted into training and testing data with 12 months assigned to the testing data set. Four different recipes were created. Recipe 1 and 3 were created to fit the Prophet model, whereas recipe 2 and 3 were created to fit the Random Forest and Decision Trees algorithms from the Tidymodels package. The recipes 1 and 2 include the external regressors,

whereas these regressors are excluded in the recipes 3 and 4. This is done as it is expected that the regressors are not beneficial for enhancing the quality of all time series. The main difference between recipes 1 and 3 created for Prophet and 2 and 4 created for the Tidymodels algorithms is that Prophet requires the date as a date variable, whereas Tidymodels algorithms require the date as an ID. Also, as Prophet allows for holiday effects which are multi-day irregular occurrences, a data frame was created including the dates of different Covid-19 measures in Spain to account for the disruption in the trends during the pandemic. The final Global Model does not contain hyperparameter tuning. A variety of different hyperparameter tuning methods was tried but did not lead to an improvement of the model.

3.2.2 Training the Models

For training the *Decision Tree*, *Random Forest* and *Prophet models*, a separate workflow was created, including and excluding the external regressors. The Covid-19 measures were added to the holidays feature in the *Prophet* models. The engine used for the *Decision Trees* is obtained from the rpart package, the engine for *Random Forest* from the ranger package and *Prophet* from the Prophet package. For all three models, the mode is set to 'regression' as the outcome variable is continuous.

3.2.3 Accuracy, Refitting and Forecasting

The models were included into a ModelTime table and calibrated using the testing data. Both the local and the global accuracy was computed and the best model was selected for each time series based on the Mean Absolute Percentage Error:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100$$

To get an idea of which models work best on average for the product categories, the interquartile range, median and mean were computed for the aggregation of time series per product category. This method is chosen as due to the high number of time series, interpretation of the accuracy of individual time series is impossible. Hereafter, the models were refitted based on all training data.

3.3 The Prophet Model

To train the Prophet model, a function was created to execute the hyper parameter tuning, cross-validation, accuracy and prediction for each time series. The hyper parameter tuning inside the loop was performed using parallel processing, utilising 5 of the 20 cores of the PC. Due to the computational expensiveness of the hyperparameter tuning, the author was not able to run all models on her PC. Therefore, only time series of the categories 'non-alcoholic beverages', 'alcoholic beverages', 'vegetables and legumes', 'meat and fish' and 'dairy and eggs' from the autonomous communities: Castilla y León and País Vasco were trained in the Prophet model.

3.3.1 Specifying the Model

The Prophet model was specified using the primary data including the date (ds) and the equivalent units (y). Also, the following external regressors were added: 1) Spanish holidays, 2) GDP per month, 3) production aggregation, 4) Food Consumer Price Index, 5) Alcoholic drinks Consumer Price Index, 6) Non-alcoholic drinks Consumer Price Index. The Covid-19 interventions are included in the parameter 'holidays'.

3.3.2 Hyper Parameter Tuning

The three hyperparameters that were tuned are 'change point prior to scale', 'seasonality prior to scale' and 'holidays prior to scale'. The Prophet model contains more hyperparameters that could possibly be tuned. However, as the official Prophet documentation recommends only tuning these three hyperparameters and the author aims to keep the model as simple as possible to reduce computational expensiveness, only these three parameters were selected (Prophet, nd). Firstly, The 'change point prior to scale' is likely to be the most influential parameter as it dictates the flexibility of the trend, specifically determining the extent to which the trend can change at the trend change points. The values 0.001, 0.1 and 0.5 were tried. Secondly, the 'seasonality prior to scale' determines the flexibility of the seasonality. A higher value allows the model to accommodate significant fluctuations in seasonality, while a lower value constrains the seasonality to have a smaller magnitude. The values 0.01, 1 and 5 were tried. Thirdly, 'holidays prior to scale' regulates the flexibility to fit holiday effects similar to the 'seasonality prior to scale' and the values 0.01, 1

and 5 were tried.

3.3.3 Cross-validation

Cross-validation is used to determine the best hyper parameter with period set to 104 weeks, and the horizon set to 52 weeks. The initial period (which was set automatically by Prophet) defines the starting time frame of historical data used for model training, while the period sets the frequency of forecast generation, in this case 2 yearly intervals. The horizon specifies the future time span for predictions beyond the training data, indicating how far ahead the model forecasts. Then, the best hyper parameters were used to train the final model and the cross-validation method described above was used again to compute the accuracy. Finally, the trained model was used to predict the consumption for 1 year.

4. RESULTS

4.1 Global Model

The global model showed to be highly efficient as it was able to train 6 different models for 348 time series in only 5 minutes (processor: 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz, RAM: 16GB).

4.1.1 Accuracy

First of all, the global model showed high MAPE in the global accuracy for each of the three algorithms trained: Decision Tree, Random Forest and Prophet indicating that the global model is somewhat problematic for this type of time series as the MAPE values exceed the bounds of what can be expected. The aggregation of MAPE based on the local accuracy of the global model per product category, including and excluding the external regressors is shown in table 1. For the models including regressors, all categories obtain a median MAPE below 20% except the 'Preserved and Premade', 'Bread and Breakfast' and 'Vegetables and Legumes' categories. The product category with the lowest median MAPE is 'Dairy and Eggs', whereas the product category with the lowest mean MAPE is 'Dry goods'. The product category with the highest mean and median MAPE under the 20% threshold is 'alcoholic beverages'. Moreover, from the three models that were trained, the Decision Tree seems to be the most effective algorithm, as this algorithm has obtained the lowest mean and median MAPE per product category for the majority of the product categories. The Random Forest seems to be the best algorithm for 3 out of 13 product categories. The Prophet model does not seem to function well in the global model including regressors. However, Prophet did compute the four best predicted time series (see figure 9).

When inspecting the global model without regressors, in most cases, the models perform worse then the models including the regressors. However, in the case of 'Vegetables and Legumes', 'Personal Care', and 'Meat and Fish', the model without regressors outperforms the model with regressors.

For the distribution of the MAPEs, see annex D . However, one should note that the performance of the models is highly dependent on the data. The quality of the data for different time series differed drastically. Therefore, the accuracy metrics are not fully reliable for making inferences about the

goodness of fit of the algorithms for specific product categories in other contexts.

Figure 9

Predicted vs Acutual Values

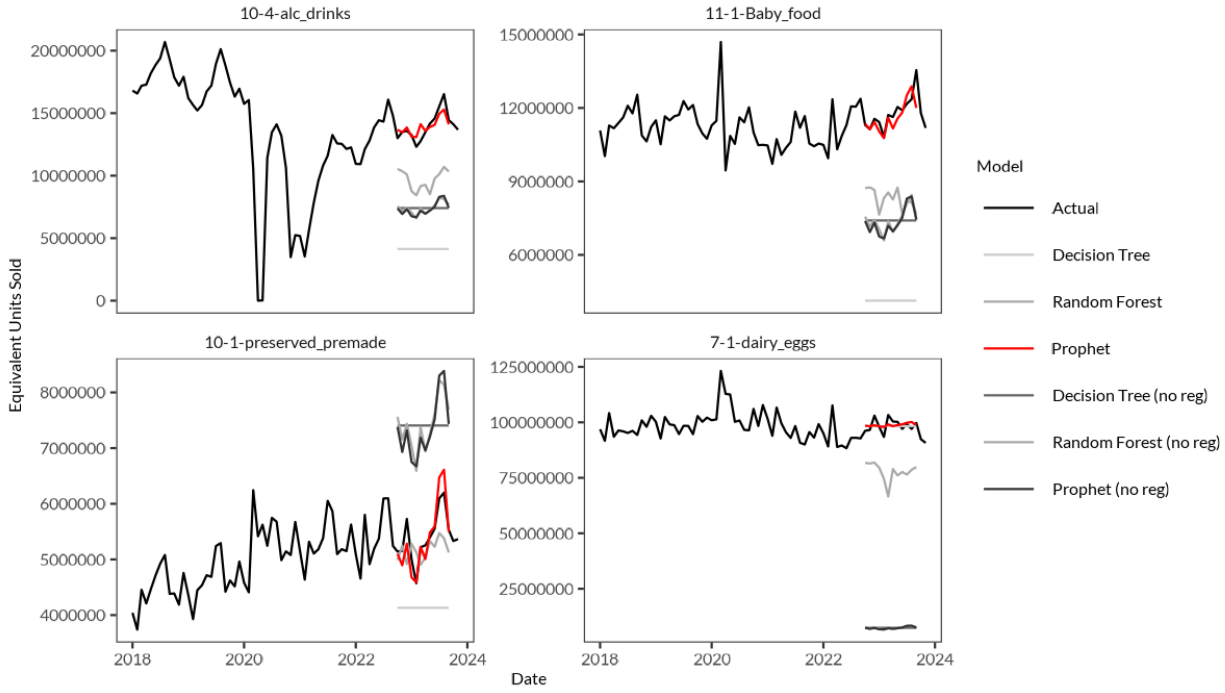


Table 1

Median MAPE Global Model

Product category	Model	Median	Model no regressors	Median no regressors
Condiments	Decision Tree	5.942	Random Forest	61.113
Dairy and Eggs	Decision Tree	6.023	Decision Tree	83.235
Dry goods	Random Forest	6.200	Decision Tree	5.078
Baby food	Decision Tree	7.056	Random Forest	46.133
Vegetables and Legumes	Decision Tree	8.961	Decision Tree	10.750
Alcoholic beverages	Decision Tree	10.210	Random Forest	50.456
Other	Decision Tree	10.719	Decision Tree	12.479
Bread and Breakfast	Decision Tree	11.060	Random Forest	64.449

Product category	Model	Median	Model no regressors	Median no regressors
Meat and Fish	Decision Tree	12.444	Random Forest	15.510
Non-alcoholic beverages	Prophet	12.590	Random Forest	53.974
Preserved and Premade	Prophet	145.651	Random Forest	60.888

Table 2

Mean MAPE Global Model

Product category	Model	Mean	Model no regressors	Median no regressors
Condiments	Decision Tree	5.942	Random Forest	53.250
Dry goods	Random Forest	6.200	Decision Tree	5.078
Dairy and Eggs	Decision Tree	6.348	Decision Tree	83.235
Baby food	Decision Tree	7.056	Random Forest	44.210
Vegetables and Legumes	Decision Tree	8.961	Decision Tree	10.750
Alcoholic beverages	Decision Tree	10.210	Random Forest	47.769
Other	Decision Tree	10.719	Decision Tree	19.603
Bread and Breakfast	Decision Tree	11.060	Random Forest	60.040
Meat and Fish	Decision Tree	12.444	Random Forest	15.510
Non-alcoholic beverages	Random Forest	14.494	Random Forest	42.252
Preserved and Premade	Prophet	586.252	Random Forest	57.147

4.2 Prophet Model

Due to the computational expensiveness of the Prophet model, only 28 time series were trained, tested and predicted. The loop including these three steps took 35 minutes to run. However, this extended training time does result in highly precise models.

4.2.1 Accuracy

The aggregation of MAPE per product category, showed that all categories obtain a median MAPE below 0.18% as and a mean MAPE below 1.16%. The product category with the lowest mean and median MAPE is ‘non-alcoholic beverages’ with 0.064%. The product category with the highest median MAPE is ‘vegetables and legumes’ being 0.176% and the highest mean MAPE is ‘dairy and eggs’ being 1.159% (see table 3). Overall, the MAPE values are decidedly low indicating a well predicting model.For the distribution of the MAPE per product category, see annex E

Table 3

Accuracy Prophet Model

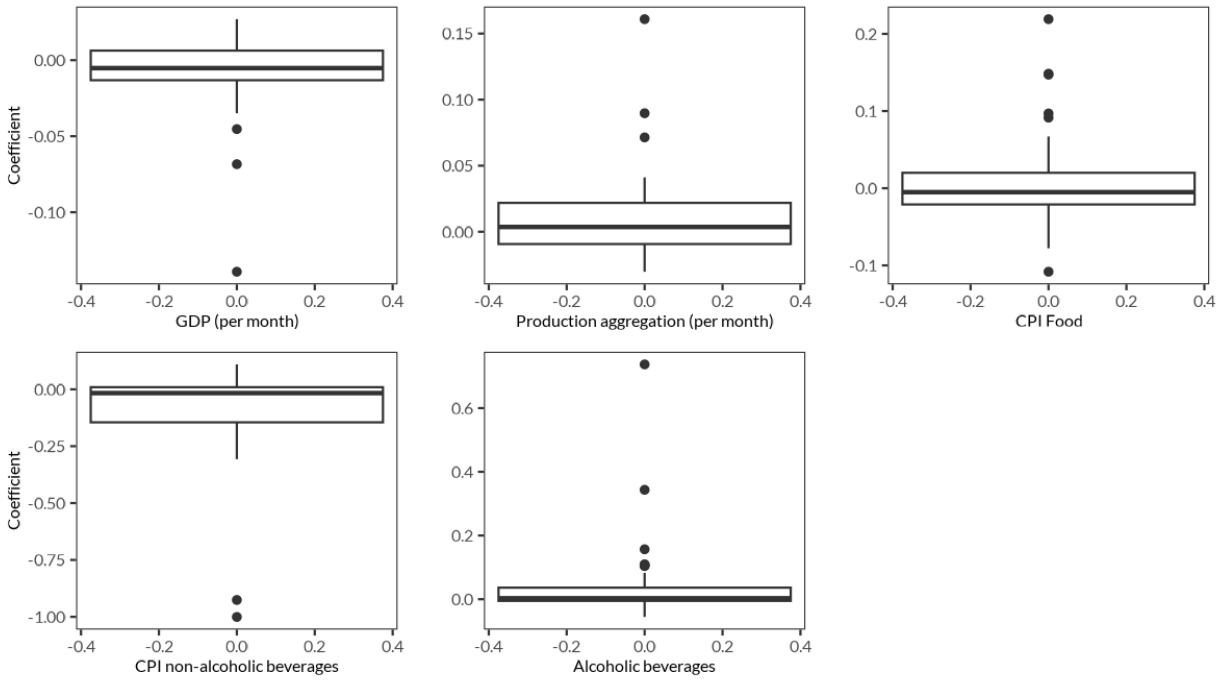
Product category	Median MAPE	Mean MAPE
Non-alcoholic beverages	0.064	0.064
Meat and Fish	0.086	0.120
Alcoholic beverages	0.109	0.154
Dairy and Eggs	0.131	1.159
Vegetables and Legumes	0.176	0.219

4.2.2 Feature Importance

The estimated beta coefficient for each external regressor in the model approximately represents the increase in the outcome variable *equivalent units* for each unit increase in the external regressor value. One should note that the coefficients are always based on the original data scale. The beta coefficients for all regressors in the model seem to be rather low. However, the highly differing scales of the regressor and the outcome variables complicates the interpretation of the beta coefficients. The confidence intervals of the beta coefficients do not cross 0 indicating that all regressors are meaningful to the model.

Figure 10

Distribution of the Coefficients of the External Regressors

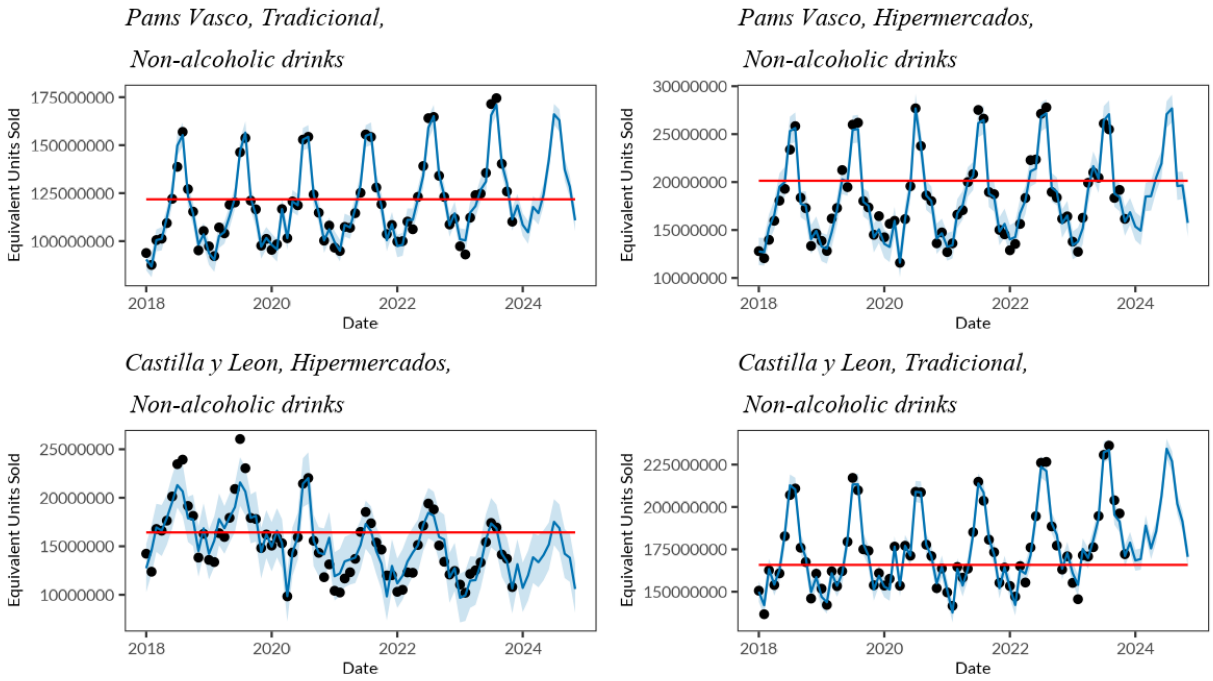


4.2.3 Prediction

In this section, the predictions for the time series ‘non-alcoholic drinks’ in the autonomous communities ‘Pais Vasco’ and ‘Castilla y Léon’ for the sales channels ‘Tradicional’ and ‘Hipermercados’, will be presented to illustrate the predictions by Prophet as not all time series can be discussed. Visualisations for the other time series can be found in the appendix. The forecast plots (figure 10) show a good fit on the existing data and a plausible fit for two years into the future with small confidence intervals.

Figure 10

Forecast plots: Non-Alcoholic Drinks



5. CONCLUSION

The aim of this thesis was to apply machine learning models to predict foodstuff consumption in Spain's autonomous communities. Accurate predictions of foodstuff consumption can help smaller businesses optimize inventory management, reducing food waste during distribution and marketing stages. Also, improved forecasting can lead to better planning, minimizing overproduction and spoilage, thereby contributing to environmental sustainability and food security.

Two methods were applied to build these predictive models. First, a global model was build in which three different machine learning models were trained: the *Decision Tree*, *Random Forest* and *Prophet* model. The benefit of a global model is that it is computationally efficient, easier to maintain, and can provide robust generalisations across different time series , although it may compromise some detailed insights specific to each individual series. In training the global model to predict consumption of foodstuffs in Spain, both the benefits and the downsides of the Global Model were encountered. On the one hand, the Global Model was able to train three different machine learning models for 348 different time series in a small amount of time. On the other

hand, the predictions by these three algorithms in the Global Model were not precise as was shown by the high mean and median MAPE of the aggregation of time series per product category. Also, the Prophet algorithm did not function well in the Global Model as it never provided the best model for one of the product categories. However, the four best predicted timeseries were predicted by Prophet. In other words, Prophet seems to vary greatly in effectiveness depending on the data. The singular Prophet model which was trained including an extensive process of hyper parameter tuning provided significantly better predictions then the the three machine learning models in the Global Model shown by the low MAPE of the aggregation of time series per product category. The downside of this method was that it is very computationally expensive which is why not all time series could be predicted.

The results of this thesis have shown that a Global Model including a Decision Tree, Random Forest and Porphet as well as a singular well tuned Prophet model can be utilised to predict the consumption of foodstuffs. However, future research needs to be done to increase the accuracy of the Global Model as well as the efficiency of the Prophet model to make the method scalable. Moreover, future work should focus on enhancing data quality and incorporating more additional external regressors that could influence food consumption patterns. Due to the computational limits of the PC of the researcher, there was a limit to the external regressors that could be added.

With this potential of effective forecasting of foodstuff consumption, policymakers and stakeholders should consider supporting the integration of advanced but also accessible forecasting models in the food distribution sector.

6. BIBLIOGRAPHY

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89.

Bagherzadeh, M., Inamura, M., & Jeong, H. (2014). *Food waste along the food chain*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Commission, E. (2023). *Food waste and food waste prevention - estimates - statistics explained*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Food_waste_and_food_waste_prevention_-_estimates.

HumanRightsWatch. (2022). *"We can't live like this": Spain's failure to protect rights amid rising pandemic-linked poverty*. https://www.hrw.org/report/2022/07/14/we-cant-live/spains-failure-protect-rights-amid-rising-pandemic-linked-poverty#:~:text=Human%20Rights%20Watch's%20findings%2C%20based,of%20living%20during%20the%20pandemic.

Jeyarangani, J., Gorla Buchayyagari, S., Chakala, R., & Mangala Venkata, S. (2023). Regressor based supermarket sales prediction using time series data. *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, 1–4.

Lipinski, B., Hanson, C., Waite, R., Searchinger, T., & Lomax, J. (2013). *Reducing food loss and waste*.

ModelTime. (n.d.). *Forecasting with global models • modeltime*. https://business-science.github.io/modeltime/articles/modeling-panel-data.html.

Moragues Faus, A., Magaña-González, R., Daniel, F., & Carasso, N. (2022). *Alimentando un futuro sostenible*. Barcelona.

Nations, U. (2024). *Population | united nations*. https://www.un.org/en/global-issues/population#:~:text=Our%20growing%20population&text=The%20world%E2%80%99s%20%20population%20is%20expected,billion%20in%20the%20mid%2D2080s.

Prophet. (n.d.). *Prophet | forecasting at scale.* https://facebook.github.io/prophet/.

Ramasubramanian, K., & Singh, A. (2017). *Machine learning using r*. Springer.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45.

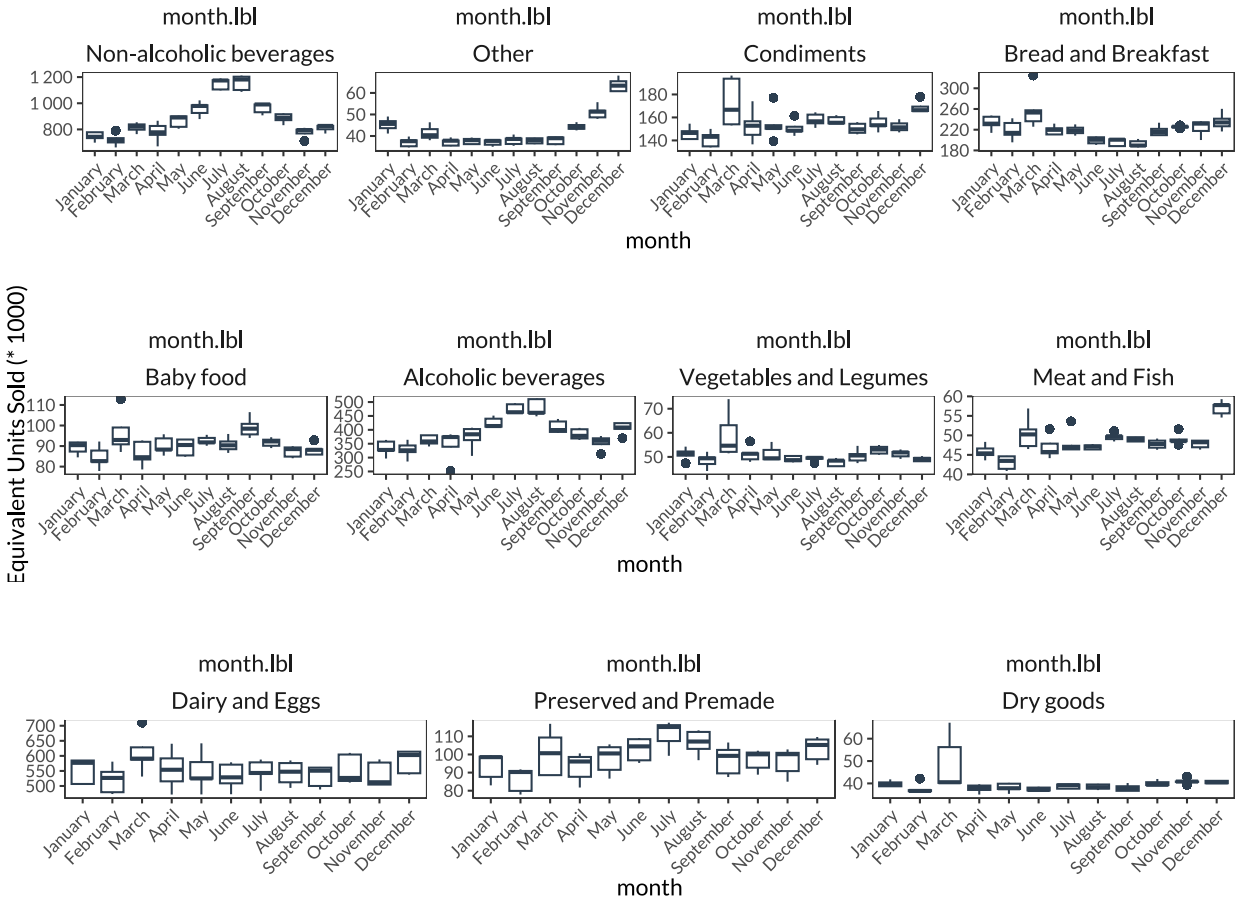
Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction. *Artificial*

Intelligence Review, *52*(1), 441–447.

7. APPENDICES

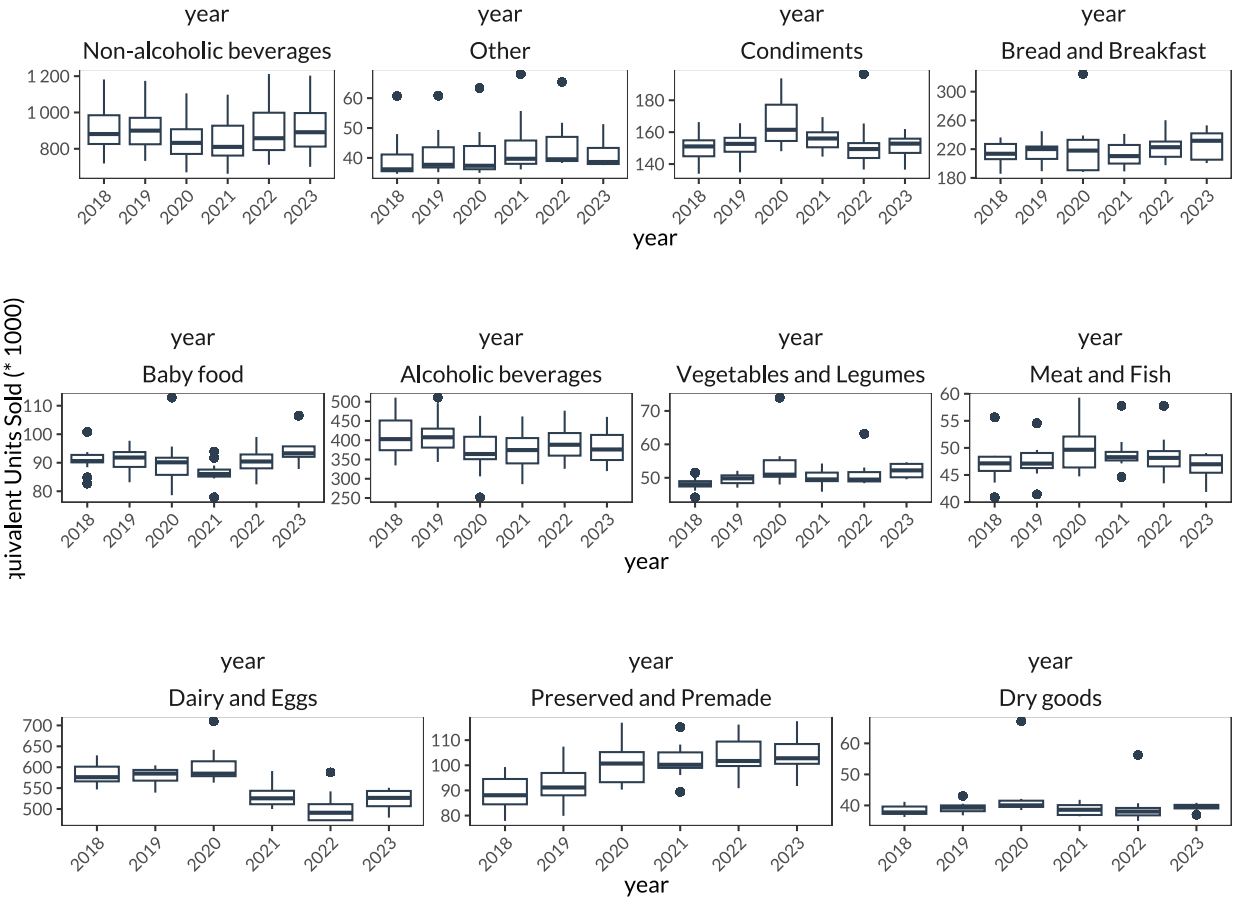
Annex A

Seasonality Diagnostics per Product Category (month)



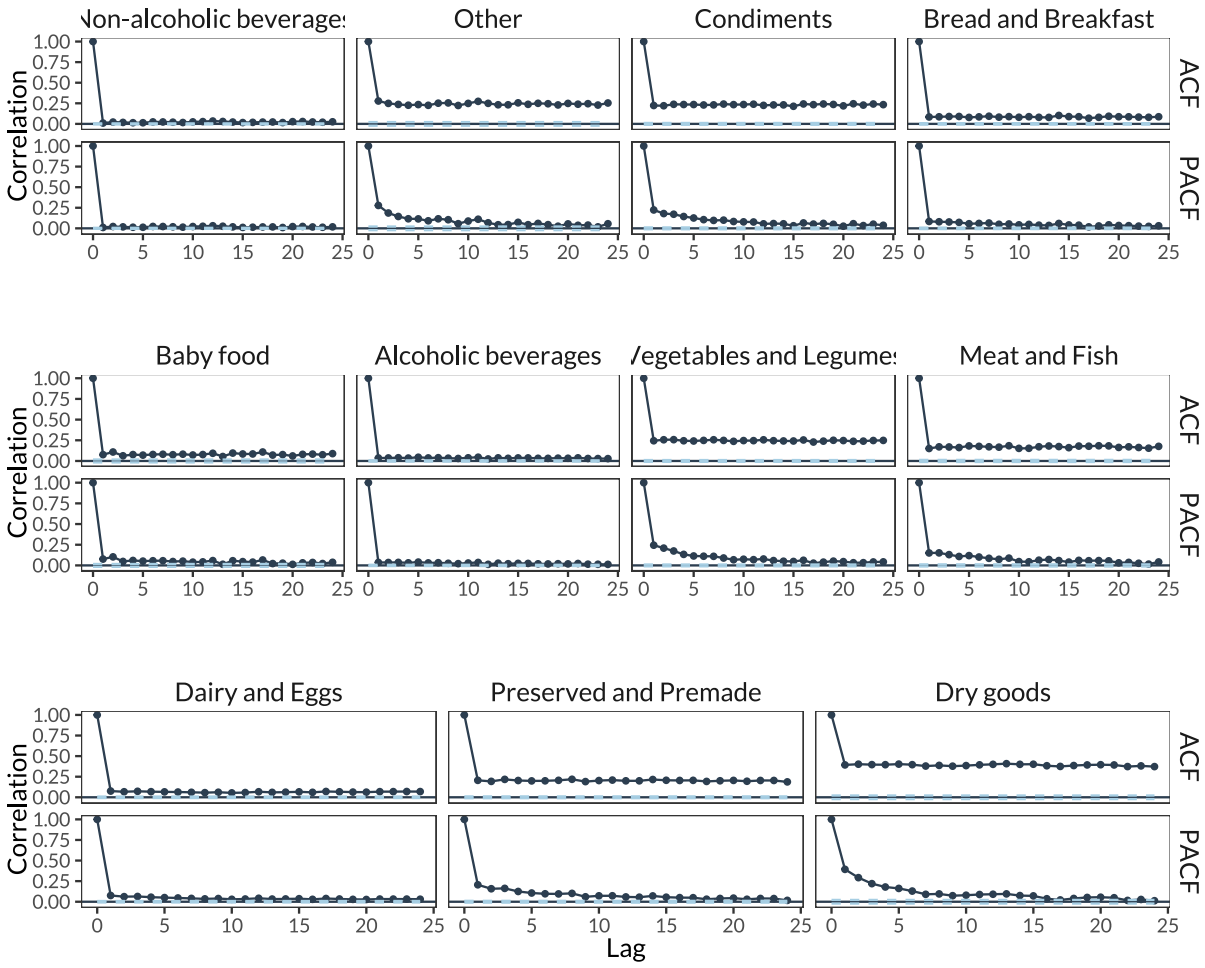
Annex B

Seasonality Diagnostics per Product Category (year)



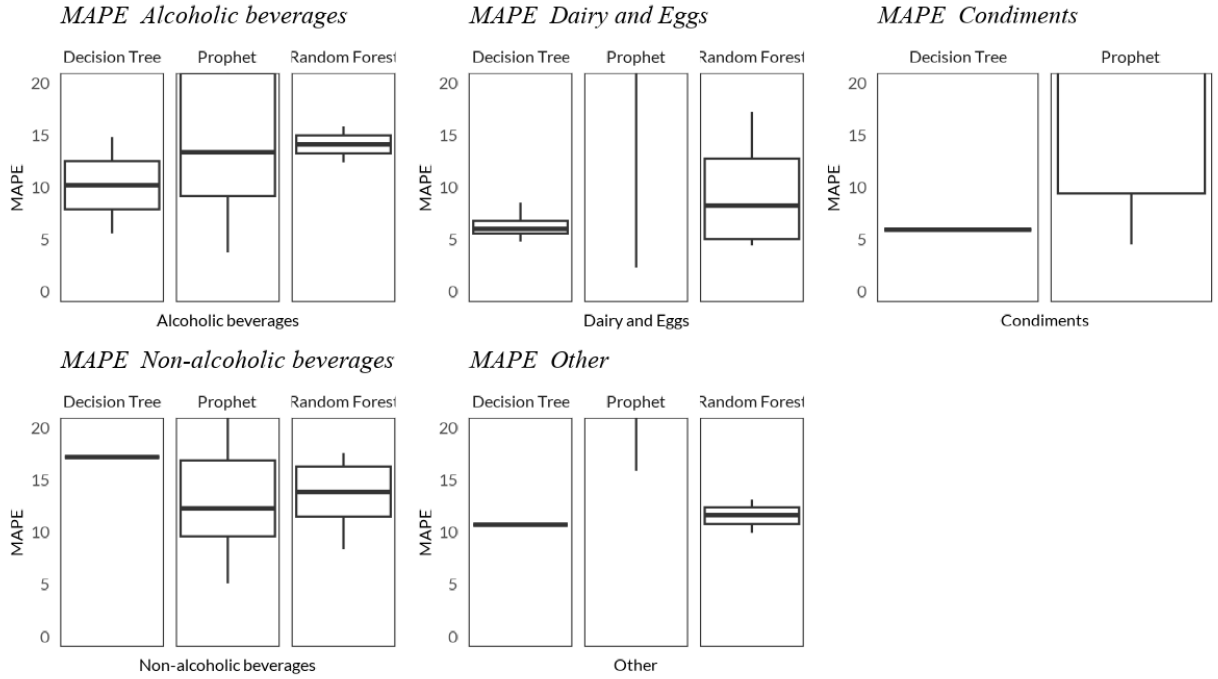
Annex C

ACF and PACF Diagnostics



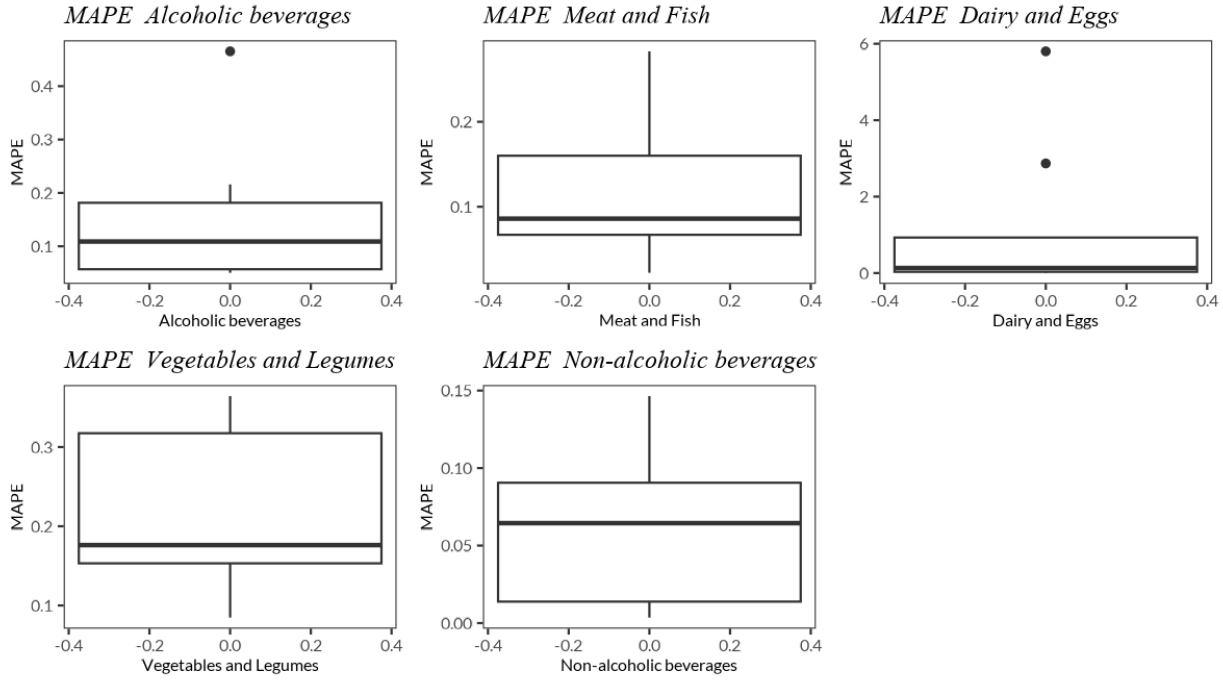
Annex D

Distribution MAPE Global Model with regressors



Annex E

Distribution of MAPE per Product Category



To replicate the thesis, please visit: https://github.com/Nienke Visscher/TFM_Computational_Social_Science.git