

# Business Economic and Financial Data Project

Flavio Kaci  
2124007

Antonio Mattesco  
2104368

Alberto Calabrese  
2103405

## 1 Introduction

The culinary arts have always held a special place in the cultural and economic landscape of Italy, with each region offering its unique contributions to the nation's gastronomic heritage. In Treviso, a city renowned for its historical charm and vibrant culinary traditions, one laboratory has distinguished itself through its artisanal approach to preparing delicacies for local gastronomy shops. Among its offerings, the signature dish—Baccalà—has become a hallmark of excellence and a key driver of demand. This report aims to delve into the time series data surrounding this iconic product, shedding light on its production and consumption patterns while exploring predictive insights.

The analysis is rooted in data collected from the laboratory's operations, focusing specifically on the production and supply metrics for Baccalà. These metrics are critical not only for optimizing internal processes but also for ensuring consistent supply to meet the expectations of gastronomy shops and, by extension, the customers they serve. By leveraging advanced statistical techniques and predictive modeling, the goal is to identify trends, seasonality, and potential anomalies in the data, thereby enabling the laboratory to make informed decisions regarding inventory management and resource allocation.

A key objective of this report is to develop a robust short-term forecasting model. Accurate predictions are essential for maintaining the delicate balance between supply and demand, especially given the perishable nature of food products and the seasonal fluctuations that characterize the culinary industry. The predictive insights derived from this analysis will provide the laboratory with actionable intelligence, facil-

itating strategic planning and enhancing its capacity to respond proactively to market dynamics.

In addition to its operational benefits, this analysis seeks to underscore the broader implications of data-driven decision-making within the food industry. By adopting a methodical approach to time series analysis, the laboratory not only strengthens its competitive position but also contributes to the sustainability of local gastronomy. This aligns with a growing emphasis on efficiency and resilience within the supply chain, particularly in an era marked by evolving consumer preferences and increasing awareness of food waste.

Through this report, our aim is to combine the richness of tradition with the precision of modern analytics, providing a comprehensive understanding of the factors that influence the production and distribution of Baccalà in Treviso. By examining historical data and projecting future trends, we seek to equip the laboratory with the tools it needs to continue thriving in a competitive and ever-changing market.

## 2 Data

### 2.1 Data Gathering

The data analyzed in this report were sourced from the archival records of the billing software utilized by the laboratory. This software has proven instrumental in compiling a comprehensive dataset on Baccalà sales, spanning a period from January 2021 to December 2024. The dataset consists of 48 monthly observations, providing a robust foundation for time-series analysis.

Each observation in the dataset is structured into three columns:

- **Date:** Represents the specific month and year of the recorded sale.
- **Baccalà\_Mantecato:** Indicates the quantity (in kg) of Baccalà Mantecato sold during the corresponding period.
- **Baccalà\_Vicentina:** Indicates the quantity (in kg) of Baccalà alla Vicentina sold during the corresponding period.

This dataset offers valuable information on the dynamics of Baccalà production and consumption, allowing the identification of patterns and trends that can inform strategic decision making. By examining these variables over time, we aim to uncover the underlying factors influencing demand and develop predictive models to improve operational efficiency.

## 2.2 Data Preprocessing

To ensure the data was ready for analysis, several preprocessing steps were undertaken. Firstly, the "Date" column was converted into an appropriate date format, allowing for accurate temporal analysis. Additional features such as "Month" and "Year" were derived from the "Date" column to facilitate the identification of seasonal trends and interannual variations. A trend variable was also introduced to capture the temporal progression of the dataset.

Missing values were checked and confirmed to be absent, ensuring the dataset's integrity. Alongside the primary dataset, an external dataset containing information on salmon consumption in Italy was integrated. This dataset, sourced from EUMOFA, provided monthly observations on salmon sales, starting from January 2009. The inclusion of this dataset was motivated by its potential to serve as an explanatory variable in modeling Baccalà sales.

The salmon dataset included three key variables:

**Date:** The month and year of salmon consumption, starting from November 2020. A two-month lag was applied to align with future forecasting scenarios.

- **kg\_std:** The standardized quantity of salmon consumed (in kilograms), used to simplify comparative analysis.

- **kg:** The raw quantity of salmon consumed (in kilograms).

These data were aggregated and aligned with the Baccalà dataset to ensure temporal consistency and compatibility. While other external variables, such as the NIC for fish products and production prices sourced from Istat, were initially considered, they were excluded due to their lack of statistical significance or mismatched temporal resolution.

The combined dataset now offers a comprehensive view of Baccalà sales alongside relevant external factors, setting the stage for robust time series analysis and predictive modeling.

The resulting dataset (Tab. 1), looks like:

Date	Mant.	Vic.	M.	Y.	Trend	Fish Cons
2021-01-01	36.4	6.1	01	2021	1	0.8904992
2021-02-01	36.4	5.4	02	2021	2	2.7444439
2021-03-01	31.2	6.1	03	2021	3	0.6030702

Table 1: Resulting dataset

## 3 Explanatory Analysis

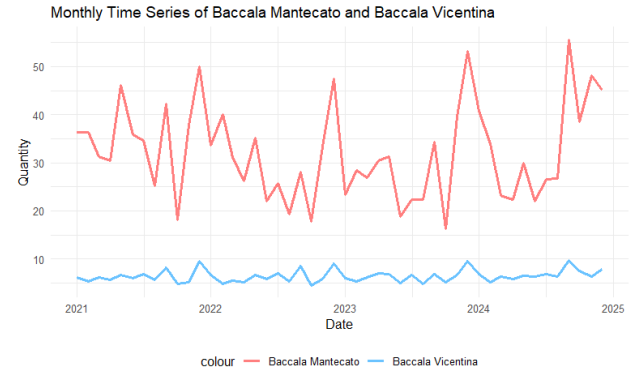


Figure 1

The exploratory phase of the analysis begins with visualizing the time series data for Baccalà Mantecato and Baccalà Vicentina. This visualization (Fig. 1) provides a comparative view of the monthly sales trends for both products. Notably, the sales

quantities of Baccalà Mantecato consistently exceed those of Baccalà Vicentina across all observed periods. Furthermore, Baccalà Mantecato demonstrates greater variability in sales, with a wider range of values, while Baccalà Vicentina exhibits a more stable pattern. Both products, however, show a pronounced increase in sales during the final months of each year, particularly in December.

To further explore yearly patterns, separate time series plots grouped by year were created for each product (Fig. 2). These plots reveal consistent seasonal trends, with sales peaking in the latter months of the year, particularly in September and December. For Baccalà Mantecato, the data also suggest a shift over time: during the first year (2021), sales volumes in off-peak months were generally higher compared to subsequent years, while peak-month sales have increased in recent periods.

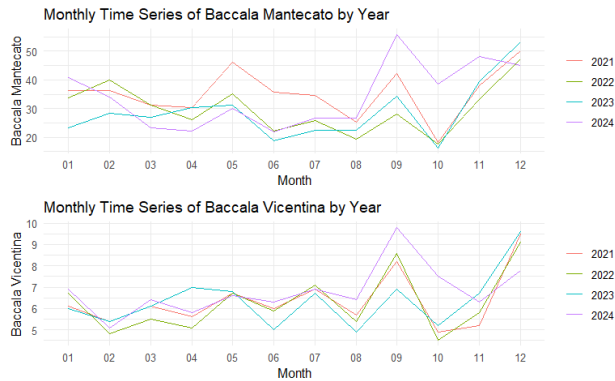


Figure 2

The autocorrelation functions (ACF) of the time series were analyzed to examine the presence of seasonality and lagged relationships (Fig. 3, Fig. 4).

Most autocorrelations fall within the confidence bands, indicating an absence of significant correlation for most lags. However, the sinusoidal pattern observed within the bands suggests periodic fluctuations, with a notable peak at lag 12, supporting the hypothesis of annual seasonality. This periodic effect will be further evaluated by analyzing the residuals of future forecasting models to confirm or refute the presence of seasonality.

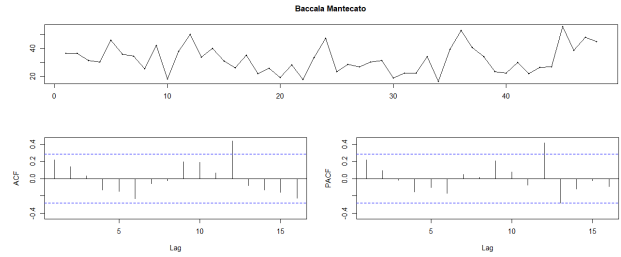


Figure 3

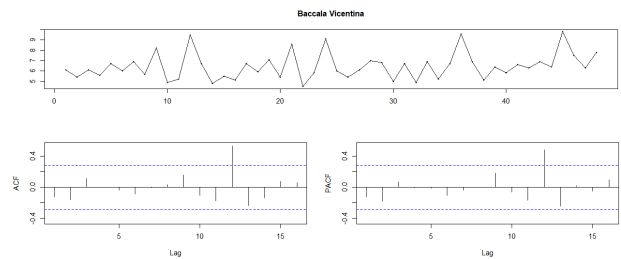


Figure 4

The insights gained from this exploratory analysis lay the groundwork for constructing robust predictive models that incorporate both seasonal and trend components. These models aim to enhance the laboratory's ability to forecast sales and optimize inventory management, ultimately contributing to its operational efficiency and market responsiveness.

## 4 Train-Test Split

To evaluate the predictive performance of forecasting models, the dataset was divided into training and testing subsets.

The training set includes observations from the initial 80% of the time periods, while the test set consists of the final 20%. By design, this approach preserves the temporal structure of the data, preventing data leakage and ensuring that predictions are based solely on prior information.

For visualization purposes, the split was applied to both Baccalà Mantecato and Baccalà Vicentina. Two distinct plots (Fig. 5, Fig. 6) illustrate the sep-

aration between training and testing data for each product. These plots reveal the temporal trends in sales, highlighting the segments used for model training and evaluation.

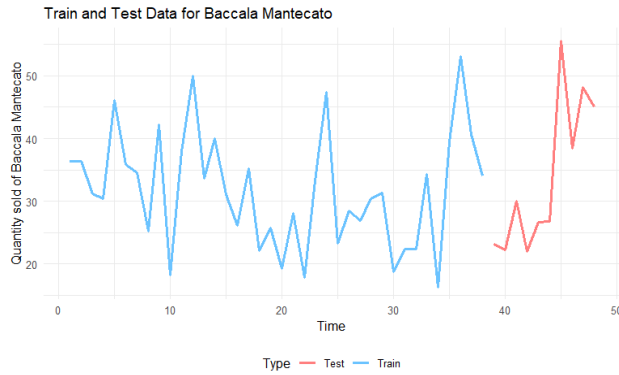


Figure 5

Baccalà Mantecato (Fig. 5): The training data (in blue) exhibits the characteristic variability and seasonal trends observed in the explanatory analysis. The testing data (in red) includes more recent observations, capturing the peak sales period in September.

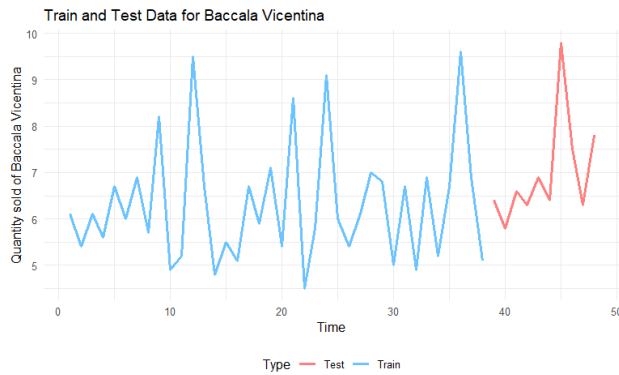


Figure 6

Baccalà Vicentina (Fig. 6): Similarly, the training data reflects the stable yet seasonal nature of sales, while the testing data emphasizes the end-of-year peaks.

Notably, in 2024, the peak sales for Baccalà occurred in September rather than December, deviating from historical patterns. This anomaly may present challenges for forecasting models, as it represents a shift in seasonal behavior for both the variants.

By implementing this train-test split, the analysis sets the stage for rigorous model development and validation.

## 5 Modelling

### 5.1 Linear Regression Models

The modeling phase began with the development of linear regression models for both Baccalà Mantecato and Baccalà Vicentina. Each variant required a distinct approach due to their different underlying patterns and relationships with predictor variables.

#### 5.1.1 Model for Baccalà Mantecato

For Baccalà Mantecato, the initial model incorporated three key predictors: a trend variable, monthly seasonality indicators, and fish consumption data. The model demonstrated strong explanatory power, with an  $R^2$  value of 0.9178, indicating that approximately 91.78% of the variability in Baccalà Mantecato sales could be explained by these predictors.

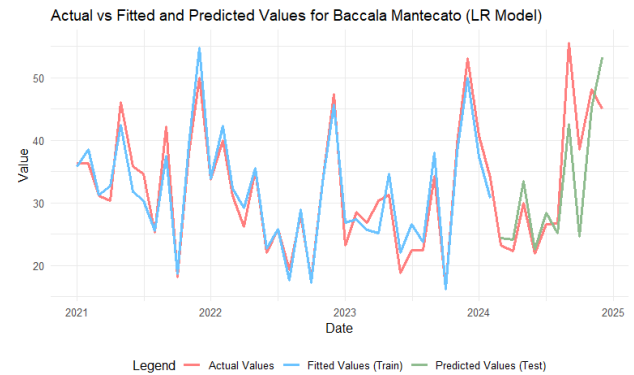


Figure 7

Analysis of the coefficients revealed several important relationships:

The trend variable showed a positive coefficient of 0.18, suggesting that, *ceteris paribus*, Baccalà Mantecato sales increase by 0.18 units per time period. Fish consumption demonstrated a strong positive relationship with a coefficient of 7.79, indicating that for each unit increase in fish consumption, Baccalà Mantecato sales rise by approximately 7.79 kg. For example, if fish consumption in Italy increases by one unit in October 2021, the quantity sold in December 2021 would increase by 7.79 kg, assuming all other variables remain constant. The monthly indicators captured significant seasonal effects, with December showing a substantial positive impact and months like February and October displaying negative coefficients relative to the January baseline.

Model selection was performed by systematically removing variables and comparing performance metrics ( $AIC$  and adjusted  $R^2$ ) across different specifications:

- Model with trend and fish consumption only
- Model with trend and monthly seasonality only
- Model with monthly seasonality and fish consumption only
- Full model with all predictors

The comparative analysis confirmed that the full model, incorporating all predictors, provided the best fit according to both  $AIC$  and adjusted  $R^2$  criteria.

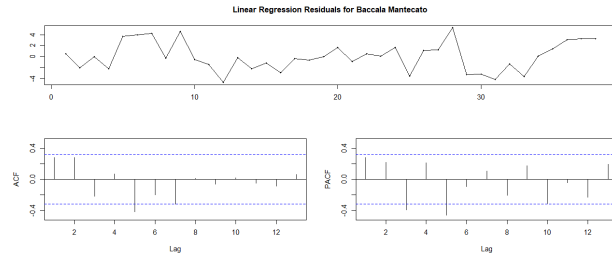


Figure 8

Residual analysis (Fig. 8) revealed no concerning patterns, with residuals appearing randomly scattered, suggesting that the model's assumptions of linearity, constant variance, and independence were reasonably satisfied. The Durbin-Watson test for autocorrelation yielded a  $P$ -value below 0.05, supporting

the null hypothesis that the residuals' autocorrelation is *zero*.

### 5.1.2 Model for Baccalà Vicentina

The modeling approach for Baccalà Vicentina followed a similar initial framework but led to different conclusions. The initial full model included the same three predictors: trend, monthly seasonality, and fish consumption. However, the analysis revealed that fish consumption did not significantly contribute to the model's performance. Through iterative model refinement:

Fish consumption was removed, leading to an improvement in adjusted  $R^2$  from 0.8261 to 0.8305. The trend variable was subsequently eliminated due to its high p-value. The final model retained only monthly seasonality indicators

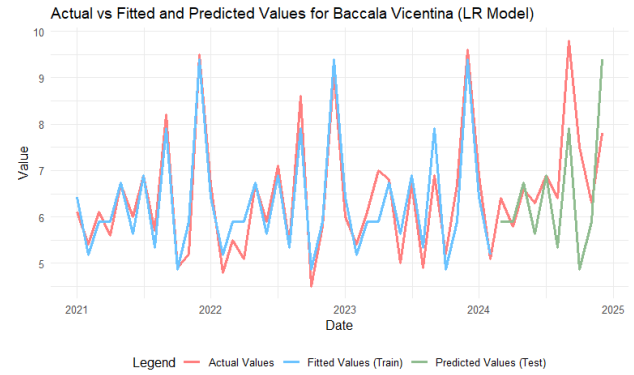


Figure 9

The simplified model demonstrated strong statistical significance with an F-statistic of 18.21 ( $p$ -value =  $1.684e - 09$ ). This suggests that variations in Baccalà Vicentina sales are primarily driven by seasonal effects, without significant influence from trend or fish consumption patterns. Comparison of model performance metrics confirmed the superiority of the seasonality-only model over the full specification, with improved  $AIC$  and adjusted  $R^2$  values. The analysis of the historical sales data for Baccalà Vicentina in kilograms reveals notable seasonal effects. December shows a significant positive im-

impact on sales (+2.975), likely driven by higher demand during the holiday season. In contrast, October (−1.558) and February (−1.250) exhibit substantial declines compared to January. Additionally, August records a notable decrease (−1.092), potentially reflecting lower summer demand. These results highlight clear seasonal patterns in sales trends. Residual analysis (Fig. 10) showed a problematic scatter, indicating problematic model fit.

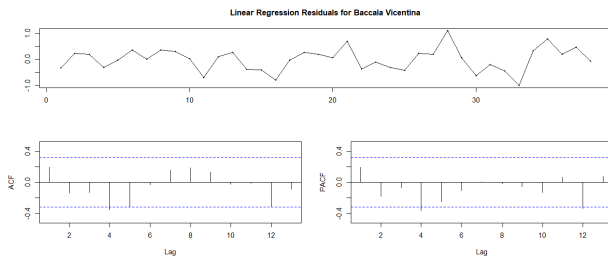


Figure 10

This idea is confirmed by the Durbin-Watson test, which suggested the presence of some autocorrelation in the residuals, as indicated by a p-value (0.18) exceeding the 0.05 significance level. The distinct modeling outcomes for the two variants highlight the importance of tailored approaches in time series analysis, as similar products may exhibit different underlying patterns and relationships with predictor variables.

## 5.2 SARIMA Models

Following the linear regression analysis, we explored various time series models for both variants of Baccalà. Initial attempts with ARMA and ARIMA models showed poor performance due to the pronounced seasonal patterns in the data. This led us to focus on Seasonal Autoregressive Integrated Moving Average (SARIMA) models, which explicitly account for seasonality in the time series.

### 5.2.1 Model for Baccalà Mantecato

Initial analysis of the Baccalà Mantecato time series confirmed the presence of both trend and seasonal

components identified in the regression analysis. After transforming the data into a time series object, we examined the autocorrelation (*ACF*) and partial autocorrelation (*PACF*) functions of the first-differenced series.

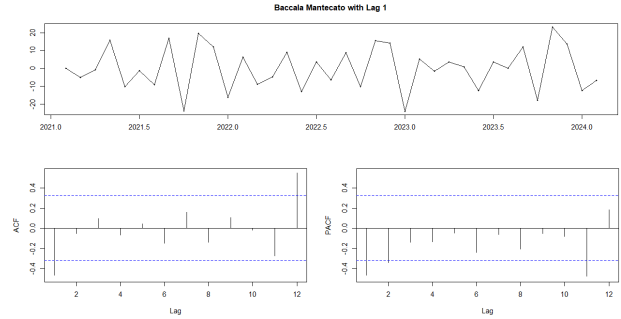


Figure 11

Both *ACF* and *PACF* plots (Fig. 11) revealed a distinctive sinusoidal pattern, with values generally within the confidence bands except for a significant spike at lag 12, strongly suggesting annual seasonality in the data. To address these patterns, three SARIMA models were evaluated:

- **SARIMA(1,1,0)(0,1,0)[12]**: Includes non-seasonal differencing ( $d = 1$ ), one autoregressive term, and seasonal differencing.
- **SARIMA(0,1,1)(0,1,0)[12]**: Incorporates a moving average term instead of an autoregressive term.
- **SARIMA(0,1,0)(0,1,0)[12]**: Uses only differencing components.

While the *SARIMA*(0,1,1)(0,1,0)[12] model showed the lowest *AIC* value, suggesting better in-sample fit, evaluation on the test set revealed that the simpler *SARIMA*(0,1,0)(0,1,0)[12] model achieved superior out-of-sample performance with lower Mean Squared Error (*MSE*). This finding highlights the importance of model parsimony and the potential risks of overfitting. Analysis of the residuals for the selected model (Fig. 12) showed some patterns that suggest the fit isn't perfect.

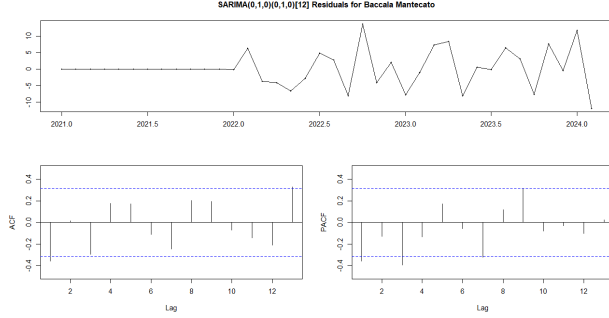


Figure 12

However, given its superior test set performance, this model was retained as the final specification for Baccalà Mantecato.

### 5.2.2 Model for Baccalà Vicentina

For Baccalà Vicentina, the time series analysis revealed strong seasonal patterns, with the *ACF* and *PACF* functions showing significant spikes at lag 12, confirming the annual seasonality identified in the regression analysis.

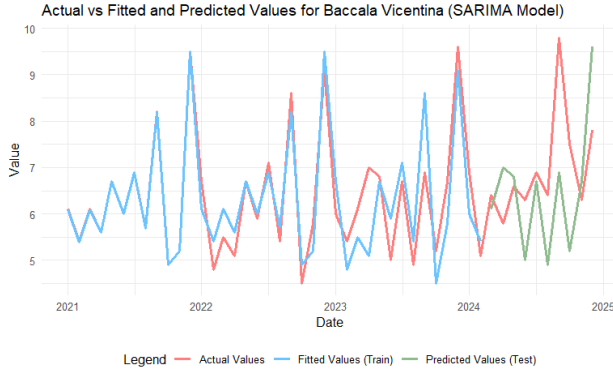


Figure 13

Unlike Baccalà Mantecato, the series showed less evidence of trend components. After examining various specifications, a  $SARIMA(0,0,0)(0,1,0)[12]$  model was selected. While models with higher orders of seasonal differencing ( $D = 2$ ) produced lower *AIC*

values, they led to higher *MSE* on the test set, indicating overfitting. The chosen model represents a balance between complexity and predictive accuracy.

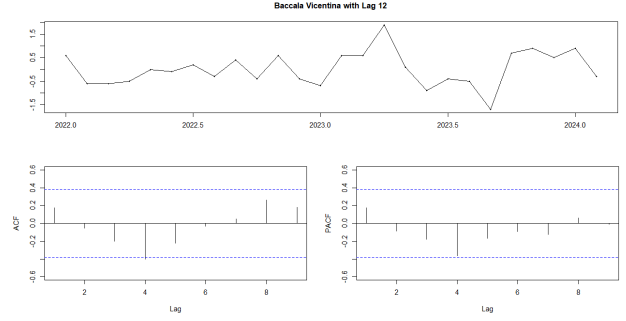


Figure 14

The residual analysis (Fig. 14) of the final model showed some remaining patterns, suggesting that while the model captures the main features of the series, there might be additional structure in the data. However, attempts to incorporate additional ARIMA components did not yield improved performance, supporting the retention of the simpler specification.

These SARIMA models complement the linear regression analysis by explicitly modeling the time series structure of the data, particularly the seasonal patterns that are crucial for both variants of Baccalà. The different specifications required for each variant further emphasize the distinct temporal dynamics of these products, despite their related nature.

## 5.3 SARIMAX Models

To extend our previous analysis with SARIMA models, we included fish consumption as an external regressor (*xreg*) to assess whether it could enhance the predictive performance of the models. The rationale behind this choice is rooted in exploring potential relationships between fish consumption patterns and the demand for the specific products analyzed. Below, we describe the steps and results obtained for each product category.

### 5.3.1 Baccalà Mantecato

We tested three SARIMAX models using different parameter configurations, maintaining consistency with the previously defined SARIMA structures. Each model incorporated fish consumption data as an external regressor:

- $SARIMAX(1, 1, 0)(0, 1, 0)[12]$
- $SARIMAX(0, 1, 1)(0, 1, 0)[12]$
- $SARIMAX(0, 1, 0)(0, 1, 0)[12]$

The comparison between the SARIMA and SARIMAX models was performed using  $AIC$  values as the primary metric. Notably, all SARIMAX models exhibited improved  $AIC$  scores compared to their SARIMA counterparts. However, recognizing the limitations of relying solely on  $AIC$ , we also evaluated the models using Mean Squared Error ( $MSE$ ) on the test set predictions.

The  $MSE$  analysis revealed a significant improvement for all three SARIMAX models, with reductions of approximately 50% compared to the SARIMA models. Considering both  $AIC$  and  $MSE$ , we selected the  $SARIMAX(0, 1, 0)(0, 1, 0)[12]$  model as the most suitable for forecasting Baccalà Mantecato.

To visualize the performance, we plot the actual versus predicted values for both SARIMA and SARIMAX on the test set.

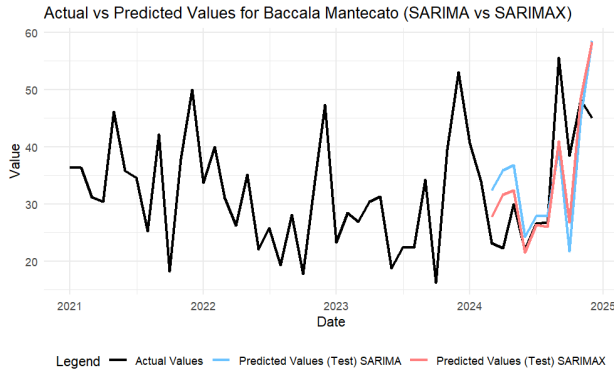


Figure 15

The plot (Fig. 15) highlights the substantial improvement achieved with SARIMAX, confirming the

utility of including fish consumption as an external regressor in this context.

Despite these positive results, the residual analysis of the selected SARIMAX model (Fig. 16) indicates some remaining autocorrelation, suggesting a potential need for an autoregressive component.

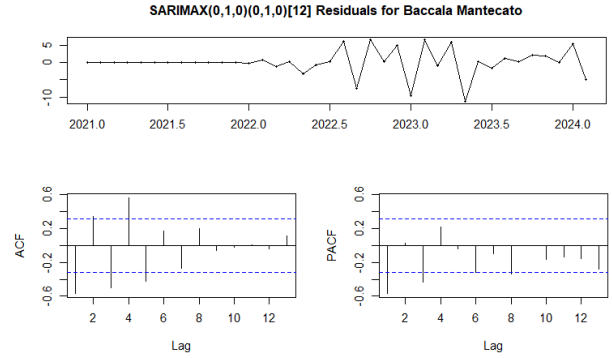


Figure 16

However, introducing this component results in overfitting and worse test set performance, leading us to accept the slight loss of information in favor of generalization.

### 5.3.2 Baccalà Vicentina

For Baccalà Vicentina, we applied a similar approach, comparing the previously selected  $SARIMA(0, 1, 1)(0, 1, 0)[12]$  model with its SARIMAX counterpart, which included fish consumption as an external regressor:

$$SARIMAX(0, 1, 1)(0, 1, 0)[12]$$

The  $AIC$  and  $MSE$  evaluations for this product category yielded surprising results. Unlike Baccalà Mantecato, the SARIMA model consistently outperformed the SARIMAX model across both metrics. This discrepancy suggests that the inclusion of fish consumption data did not enhance the model's predictive capabilities for Baccalà Vicentina.

A closer examination of the prediction plots (Fig. 17) revealed that both models produced similar trends. However, the SARIMAX model appeared



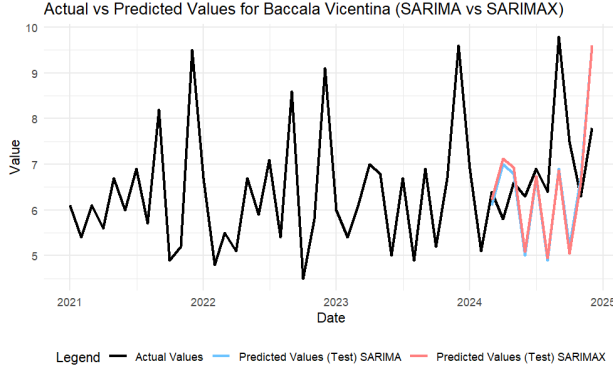


Figure 17

to overestimate quantities during the test period, coinciding with a recurring December peak absent in the test set. This overestimation likely contributed to the inferior performance of the SARIMAX model, as indicated by the higher  $AIC$  and  $MSE$  values.

Consequently, we decided to retain the original SARIMA model for Baccalà Vicentina, as the inclusion of fish consumption did not provide a tangible benefit and introduced a risk of overfitting.

The comparative analysis between SARIMA and SARIMAX models demonstrates the importance of context-specific considerations when incorporating external regressors. While fish consumption significantly improved the predictive accuracy for Baccalà Mantecato, it had the opposite effect for Baccalà Vicentina. These findings highlight the nuanced relationship between external factors and product demand, emphasizing the need for tailored modeling approaches in time series forecasting.

## 5.4 GAM models

### 5.4.1 Baccalà Mantecato

In this analysis, we aimed to explore the potential benefits of non-linear models, using Generalized Additive Models (GAM), and compared them to the previously selected Multiple Linear Regression model. Initially, we fitted the MLR model and then examined the linear effects through GAM to assess the

possibility of improving the fit with non-linear relationships.

Upon visual inspection of the plots (Fig. 18), it was evident that the relationship between the predictors and the response variable was clearly linear.

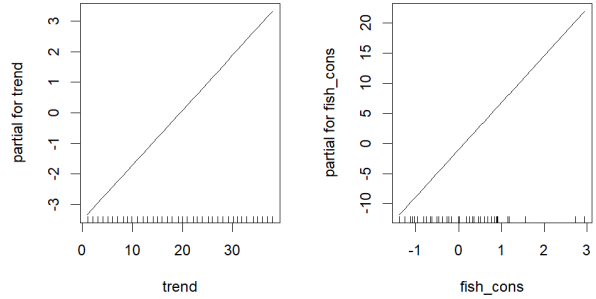


Figure 18

This observation was further supported by the statistical results from the 'Anova for Nonparametric Effects', which showed that the nonparametric  $F_{values}$  for the smoothing terms  $s(trend)$  and  $s(fish\_cons)$  were not statistically significant, with  $P_{values}$  of 0.19 and 0.08, respectively. These results strongly indicate that these variables do not exhibit non-linear relationships with the response variable as initially hypothesized.

Therefore, for parsimony, interpretability, and simplicity reasons, we concluded that the GAM model was unnecessary and opted to exclude it in favour of the multiple linear regression model. This model not only aligns better with the data but also provides more meaningful and statistically sound results, consistent with the initial assumption of linearity.

### 5.4.2 Baccalà Vicentina

For Baccalà Vicentina, we followed a similar approach by fitting a GAM model to check for non-linear relationships. However, as with Baccalà Mantecato, the plots clearly demonstrated that the relationship between the explanatory variables and the

response variable remained linear. This again suggested that the non-linear modeling approach was not needed.

Moreover, in the section dedicated to linear regression, we had already determined that the best model for Baccalà Vicentina was the reduced model, where only the Month variable was included as a predictor. Given that this model was the most effective, we decided to retain it.

In conclusion, both analyses—Baccalà Mantecato and Baccalà Vicentina—strongly suggest that non-linear models, such as GAMs, are not necessary in this case. The evidence indicates that linear models are sufficient for capturing the relationships in the data, and choosing them over GAMs offers a more parsimonious, interpretable, and statistically sound solution.

## 5.5 Prophet model

For the Prophet model, two distinct models are created for each dish: one with logistic growth and one with linear growth. The choice of model depends on the nature of the data, but both models include yearly seasonality, which is crucial to capturing the seasonal cycles typical of food sales, especially for traditional dishes like Baccalà. The parameter  $n.changepoints = 5$ , that is selected after several tries, trying to avoid overfitting but also to reach a good fit, specifies the number of points in time where the series can experience structural changes. Finally a multiplicative seasonality is selected for both dishes.

### 5.5.1 Baccalà Mantecato

We adapted the model for the Mantecato sales data, and the residuals revealed some issues, particularly highlighted by both the  $PACF$  and  $ACF$  plots, especially with the spike at lag 12. Despite this, attempts to improve the model by increasing the number of change points or experimenting with different specifications led to a deterioration in performance on the test set.

We also tried to incorporate a model estimated on the residuals using the ‘*auto.arima*’ function,

but this did not improve the model’s performance. Given these results, we decided to retain the original Prophet model for its simplicity, as further adjustments did not yield better predictive accuracy.

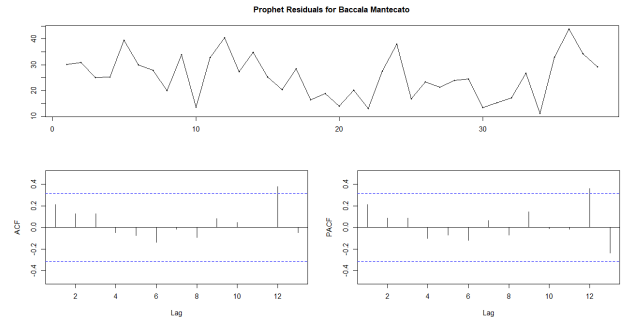


Figure 19

Finally, in the graph below (Fig. 20), we show the predicted values versus the actual values. From the graph, we can observe a good fit to the training data, but the same issues as with the other models are evident on the test set.

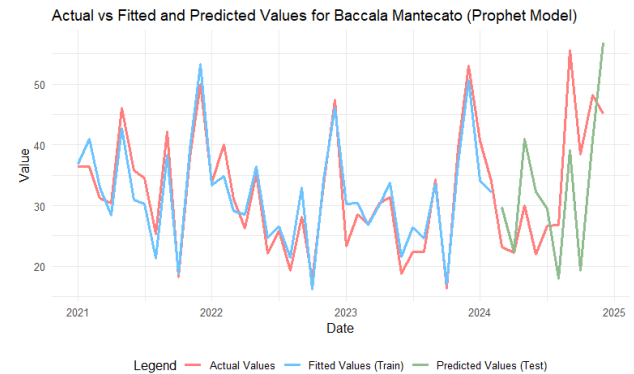


Figure 20

### 5.5.2 Baccalà Vicentina

The same behavior is observed for Baccalà alla Vicentina, where the residuals (Fig. ??) show a spike at lag 12. However, we do not want to dwell too much on repeating the same observations, so for complete-

ness, we are including the graph of the actual values versus the predicted values (Fig. 21).

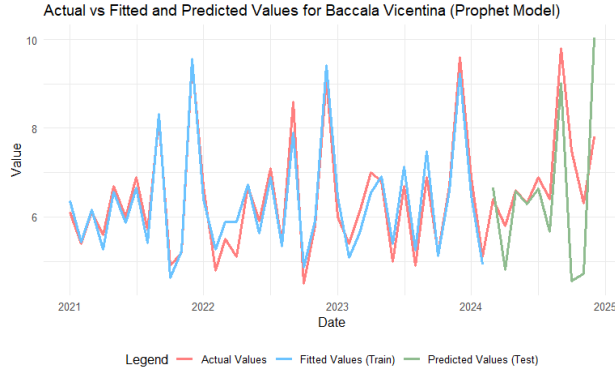


Figure 21

## 5.6 Exponential smoothing Models

After exploring other approaches, we decided to test Exponential Smoothing (ES) models to analyze and forecast the time series data for Baccalà Mantecato and Baccalà Vicentina. These models are particularly valued for their simplicity and their ability to assign exponentially decreasing weights to past observations, emphasizing the most recent data. They also offer flexibility in handling trend and seasonality, making them ideal for time series with recurring patterns. For both datasets, we began by testing the classic ETS model (Error, Trend, Seasonality), an approach that automatically selects the optimal configuration based on the data.

### 5.6.1 Baccalà Mantecato

For the Baccalà Mantecato series, the ETS model effectively captured the key dynamics of the data, providing a strong baseline for further analysis. The model demonstrated its ability to adapt to the trends and seasonality present in the series, which were confirmed by both visual and statistical evaluations. Residual analysis (Fig. 22) revealed that the residuals were randomly distributed around *zero*, with no significant patterns or autocorrelation. This indicates

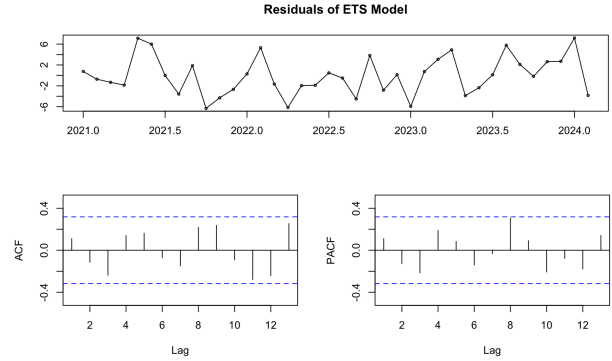


Figure 22

that the model successfully accounted for the underlying trends and seasonality, leaving minimal systematic errors. Such behavior is a positive indication of the effectiveness of the ETS model for this time series. The quality of the forecast was further validated by a

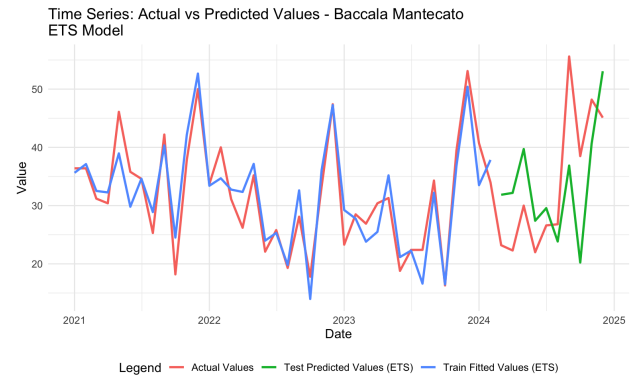


Figure 23

visual comparison between the actual and predicted values (Fig. 23). During the training phase, the forecast line closely followed the actual data, reflecting the model's ability to capture key patterns. During testing, the predictions remained near the observed values, with only minor deviations in the most extreme peaks. This consistency highlights the model's capacity to generalize and handle unseen data effectively.

We then explored more specific variants of the ETS model by introducing additive and multiplicative seasonal components through the Holt-Winters method. Using the additive approach resulted in slightly worse outcomes compared to the standard ETS model, with increases in both the  $MSE$  and the  $AIC$ . These metrics indicated that the added complexity of the additive approach did not yield significant benefits. The multiplicative approach for seasonality provided even less satisfactory results, with higher errors and increased complexity, confirming that the original ETS model was the best option for this series as we can see from (Tab. 2).

Model	MSE	AIC
ETS	111.7995	266.4093
Holt-Winters Additive	114.8030	274.4056
Holt-Winters Multiplicative	177.2098	274.7737

Table 2: Comparison of Models based on MSE and AIC — Baccalà Mantecato

### 5.6.2 Baccalà Vicentina

Turning to the Baccalà Vicentina series, the behavior of the models was similar, but presented an interesting difference. Once again, the ETS model produced strong initial results, with a low  $MSE$  and a favorable  $AIC$ . However, incorporating multiplicative seasonality through Holt-Winters led to a slight reduction in  $MSE$ , albeit at the cost of a higher  $AIC$  (Tab. 3).

Model	MSE	AIC
ETS	1.475781	105.6035
Holt-Winters Additive	1.512441	111.4078
Holt-Winters Multiplicative	1.434768	109.9921

Table 3: Comparison of Models based on MSE and AIC — Baccalà Vicentina

This trade-off between forecast accuracy and model complexity led us to select the multiplicative seasonal model as the most suitable for this series. While the increase in  $AIC$  reflected higher complexity, our primary goal of minimizing forecast errors justified this

choice. By prioritizing  $MSE$ , we were able to achieve a more accurate forecast, albeit at a slightly higher computational cost. Residual analysis (Fig. 24) con-

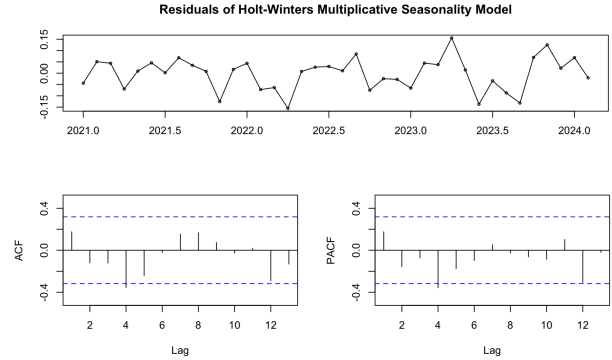


Figure 24

firmed that the selected models effectively captured the main characteristics of the time series. The residuals were randomly distributed around  $zero$ , showing no repeated patterns and indicating no systematic errors. The absence of significant spikes in the residuals'  $ACF$  and  $PACF$  plots further reinforced this conclusion. This is a strong indication that the chosen models are robust and reliable for forecasting purposes. Finally, the forecast plots (Fig. 25) clearly

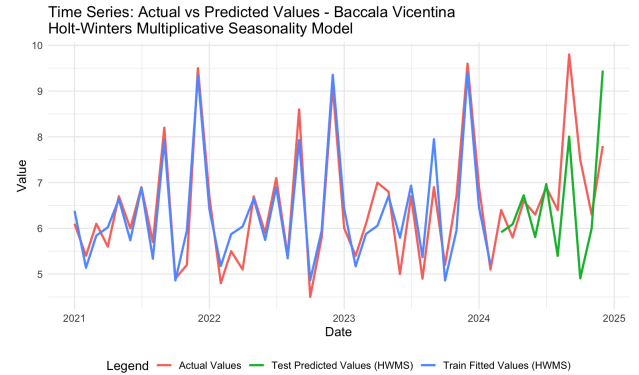


Figure 25

illustrated the models' ability to follow the trends and seasonality of the data in both the training and

testing phases.

## 6 Conclusions

In this last section we are going to compare the results of the different models on the training and test sets. To compare and properly discuss these results we created a table (Tab. 4) with the *MSE* values for each model on the training and test sets.

Model	Train_M	Train_V	Test_M	Test_V
<b>LR</b>	6.7755	0.1879	45.5953	1.5124
<b>SARIMA</b>	29.7579	0.3605	104.5050	2.2650
<b>Prophet</b>	9.2290	0.1130	119.0730	1.8423
<b>SARIMAX</b>	14.2973	0.3683	64.0557	2.3494
<b>ETS</b>	13.2462	0.1936	111.7995	1.4348

Table 4: Performance metrics for different models (MSE)

From the table 4 and subsequent plots, several insights can be drawn regarding the performance of the models.

The results show that Linear Regression provided the best performance, achieving the lowest mean squared error on the test set. However, at the same time, the discrepancy between the results obtained on the training set and those on the test set suggests a possible case of overfitting. This observation applies to all analyzed models and has been the subject of numerous investigations to understand the reasons behind it, which we aim to address in future analyses.

The best alternative to Linear Regression, in our assessment, is the SARIMAX model. By incorporating the effect of the exogenous variable, SARIMAX outperforms models that only account for seasonality and trend.

An interesting observation regarding the test set performance, as previously mentioned, is the significant discrepancy between training and test results. Analyzing the historical series, we observed that the peak for 2024 occurred in September, whereas in previous years, it consistently occurred in December. However, all models predicted the peak for December 2024, adhering to historical trends that were not followed in the most recent year. This anomaly remains

unexplained.

A practical limitation of the Linear Regression model is its dependence on the exogenous variable (fish\_cons), which is necessary for making predictions. For instance, to forecast the value for January 2025, the fish\_cons data for November 2024 is required. Since this data is sourced online, there is no guarantee it will be updated frequently enough, potentially compromising the timeliness of the forecasts. This limitation also applies to SARIMAX, which is why a model like SARIMA could be considered as an alternative.

For Baccalà Vicentina, the issues are similar: the test set is not representative of the training set, and evaluating the models' performance on the test set distorts the results. In this case, the best-performing model on the training set is Prophet, but it is outperformed by the Holt-Winters Multiplicative model on the test set. This result suggests, as evidenced by the p-values in the regression model, that the relationship with the fish\_cons variable is not statistically significant.

Although the results for Baccalà Vicentina initially appear better than those for Baccalà Mantecato, the scale of the data must be considered. The quantities in kilograms sold for Baccalà Vicentina are roughly between 1/4 and 1/3 of those for Baccalà Mantecato. During our analyses, we tested additional models, for both the response variables, such as introducing non-linear trends (look at the code), but, as expected, none of these produced significant improvements for our objective.

Overall, the analysis demonstrates that certain models are preferable to others and that the most complex model is not always the best-performing one. For example, adding transformations of the trend and fish\_cons variables in the code did not yield better results. However, due to the ambiguous behavior of the time series in the test set and the lack of a clear explanation for it, we cannot be certain that using the model for business purposes will be valid and achieve the same level of accuracy observed in the training set.