Time Series Analysis: Car Sales In Italy

Lorenzo Leoni

***Abstract*—In this project was analyzed a database composed by almost all the new registrations of car manufacturers in Italy in each month since January 2012. With this database it was possible to calculate for each month the total registrations in the country and obtain the time series ready to be analyzed. The first step was the analysis of it, checking its characteristics such as seasonality and trend. Afterwards it was done model selection between seven models such as two Holt-Winter's exponential models, three SARIMA and two SARIMAX. The best ones found were a SARIMA and the SARIMAX model with unemployment rate as exogenous regressor. These two were tested from September 2021 until March 2022 and the best was SARIMAX.**

I. Introduction

With the advent of the pandemic, every market has suffered a major backlash. This is also the case of the automotive industry, since the lockdown in March and April caused a slump in sales. In Italy the market seems to be more stable but predictions about it are difficult to do since forecasting models now have to be able to deal with the slump caused by the COVID-19 that, obviously, influences the quality of the predictions generated by them for these years. In this paper it was reported all the steps followed in order to build models able to forecast with good results also with the months of lockdown, that can be thought as outliers.

II. Methodology

In this section will be reported all the analysis done on the time series and all the models built excepted the best two that will be reported in the *Conclusion* section.

A. Time Series Data Preparation and Analysis

Once loaded all the useful library it can be possible to import the main database *SalesITA.csv*. The data belong to the *Unione Nazionale Rappresentanti Autoveicoli Esteri* website [8]. The dataset has 40 car manufactures and 123 months from 2012 until 2022. All the csv from the website were merged using Excel and it was kept only the car manufactures which were present in all the files. This check can be possible using the Excel function *VLOOKUP* that control if an element is in a table and links it with its value.

Now The preparation phase can start. To initialize a time series object in R, it was necessary to transpose all the database in order to have car manufactures as columns and in the rows the registrations for each month. It was possible to add in row every monthly registration for each car manufactures in order to obtain the monthly registrations in Italy. The time series as named

TABLE I

Summary Statistics					
Min	1st Qu	Median	Mean	3rd Qu	Max
4119	109093	133365	132127	155404	225448

Summary Statistics of car sales in Italy time series from January 2012 until March 2022

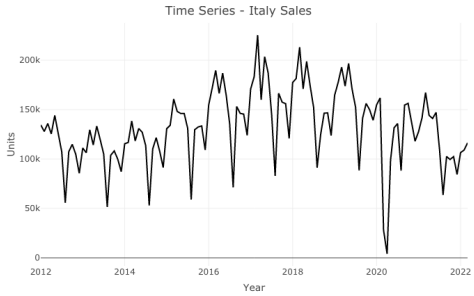

Fig. 1: Car sales in Italy time series from January 2012 until March 2022

ITA and in Table I are reported its summary statistics. In Figure 1 can be seen the evolution of the sales in Italy, where in March and April 2020 there is the slump caused by the lockdown. In Figure 2 and Figure 3 is clear that in August there are the lower sales in general.

Time series X are composed by three type of components: *Trend* T, *Seasonality* S and *Noise* N. The different components are commonly considered to be either additive or multiplicative [6]. A time series with additive components can be written as:

$$X = T + S + N$$

A time series with multiplicative components can be written as:

$$X = T \cdot S \cdot N$$

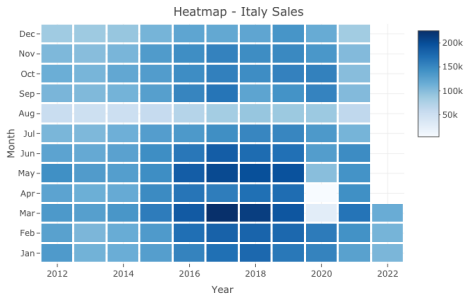

Fig. 2: Car sales in Italy time series heatmap for each year and month from January 2012 until March 2022

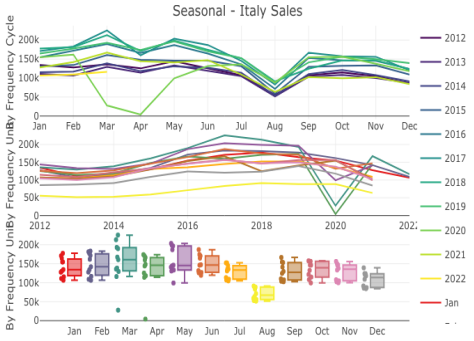

Fig. 3: Car sales in Italy time series for each year and month from January 2012 until March 2022

For seasonal patterns, two common approximations are additive seasonality (where values in different seasons vary by a constant amount) and multiplicative seasonality (where values in different seasons vary by a percentage).

The time series in analysis seems to be multiplicative since the seasonality has an increasing volatility, in particular before the lockdown period. To be sure about that it was used a method of coefficient of variation of seasonal differences and quotients [3]. The seasonal differences D was computed by taking the difference between a certain season of a year and the same season from the year before while the seasonal quotient Q was computed as the quotient of a certain season of a year and the same season from the year before. In particular the formulas are the following:

$$\begin{cases} D_{i,j} = X_{i,j} - X_{i-1,j} \\ Q_{i,j} = \dfrac{X_{i,j}}{X_{i-1,j}} \end{cases}$$

where i is the year and j the month. Thereafter the coefficient of variation of the seasonal differences $CV(D)$ and the coefficient of variation of the seasonal quotients $CV(Q)$ are computed as:

$$\begin{cases} CV(D) = \frac{sd(D)}{mean(D)} \\ CV(Q) = \frac{sd(Q)}{mean(Q)} \end{cases}$$

where sd is the standard deviation. The decision

rule that aid the choice of model was define as:

$$\begin{cases} |CV(D)| < |CV(Q)| & \text{additive} \\ |CV(D)| \geq |CV(Q)| & \text{multiplicative} \end{cases}$$

This approach return that the time series of the car sales in Italy is multiplicative like said before.

Known that it can be possible to decompose the time series with the function *decompose*, which uses the moving average method in order to extract the trend from the series. Dividing the time series (since it is a multiplicative one) with the trend component, the series is detrended. The next step is to estimate the corresponding seasonal component for each frequency unit. This simple calculation is done by grouping the observations by their frequency unit and then averaging each group. The output of this process is a new series with a length that is equal to the series frequency and is ordered accordingly. This series represents the seasonal component. This method is not problematic when applying to an additive series, as the magnitude of the seasonal oscillation remains the same (or close to the same) over time. On the other hand, this not the case for a series with multiplicative growth, as the magnitude of the seasonal oscillation grows over time [6]. The last step is to divide the time series with the two components found, in order to obtain the noise. It's important to check if the remaining noise is similar to a white noise, also called as "purely random process" if the observations are uncorrelated random variables [7]. In Figure 4 is reported the decomposition of the series. In this case the noise component has an important variation caused by the pandemic.

Another characteristic to check before start building forecasting model is the stationarity. Stationarity means that statistical parameters of a time series do not change over time. In other words, basic properties of the time series data distribution, like the mean and variance, remain constant over time. Therefore, stationary time series processes are easier to analyze and model because the basic assumption is that their properties are not dependent on time and will be the same in the future as they have been in the previous historical period of time. Alternatively, time series that exhibit changes in the values of their data, such as a trend or seasonality, are clearly not stationary, and as a consequence, they are more difficult to predict and model [5]. The white noise is the example of a calssic stationary time series. In this case it has evidence about the seasonality while the trend component is not linear or monotonic, hence it'll be necessary an in-depth study for it that it was done during the building of the Holt-Winter exponential models.

The non-stationarity can be checked with the autocorrelation and partial autocorrelation plots. The autocorrelation is the correlation of the variable with itself at different times (lags) while the

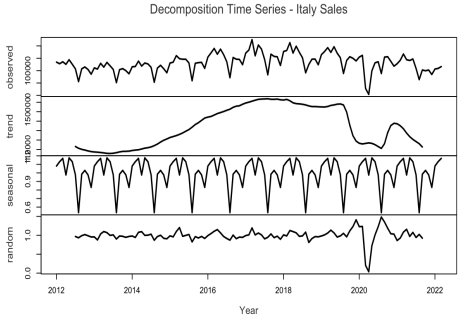

Fig. 4: Car sales in Italy time series decomposition in each component: Trend, Seasonality and Noise, from January 2012 until March 2022

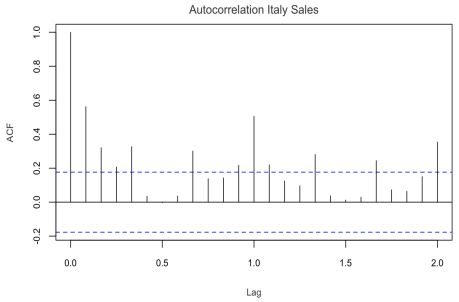

Fig. 5: Car sales in Italy time series autocorrelation plot for the first 24 lags

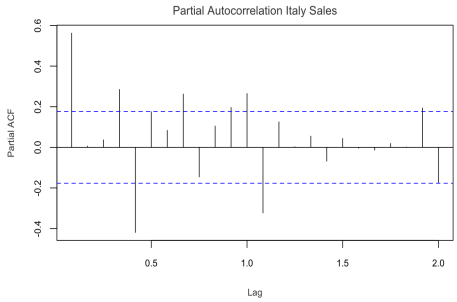

Fig. 6: Car sales in Italy time series partial autocorrelation plot for the first 24 lags

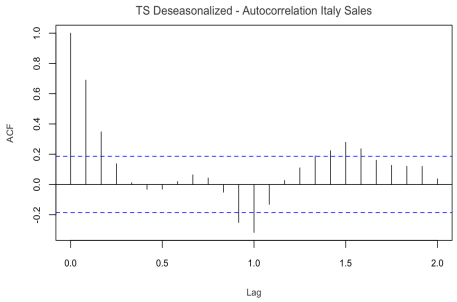

Fig. 8: Car sales in Italy time series deseasonalized autocorrelation plot for the first 24 lags

partial autocorrelation is the correlation that results in a fixed lag after removing the effect of any correlations due to the terms at shorter lags [1]. In Figure 6 and Figure 5 are reported respectively the autocorrelation and the partial autocorrelation plots. In the ACF plot can be seen the influence of the seasonality component, indeed there are several peaks in different lags and the peak at 1 year indicates seasonal variation.. The trend component is not clear, since the gradual decay, typical of a time series containing a trend, isn't evident [1].

One of the last steps is to analyze how the series is without seasonality and trend component. This is useful for isolating the components not removed and understanding if the assumptions made before with the ACF/PACF are correct. In Figure 7 is reported the time series deseasonlized (thanks to a lag-12 differencing [6]), which seems to be non-stationary since the variance change with the time. This is confirmed also with the ACF and PACF in Figure 8 and 9 respectively, since the ACF plot is not as a ACF plot of a white noise (where the significant lag is only the first one). These two plots are important when it has to choose the order of some parameters in the next models.

In Figure 10 is reported the time series detrendalized (thanks to a lag-1 differencing [6]), which seems to be non-stationary since the variance change with the time, not like the deseasonlized time series but still not stable enough. This is confirmed also with the ACF and PACF in Figure 11 and 12 respectively, since the ACF plot is

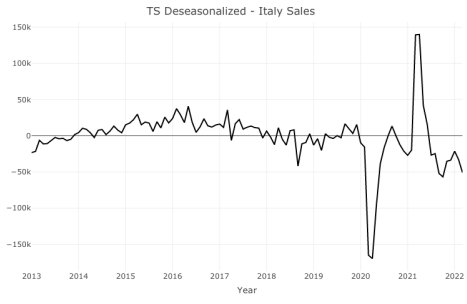

Fig. 7: Car sales in Italy time series deseasonlized from January 2012 until March 2022

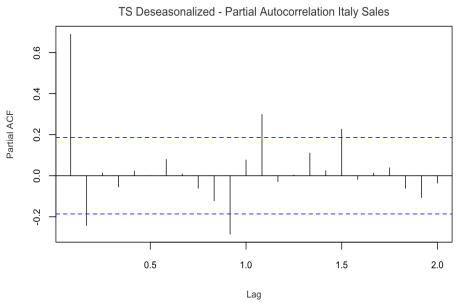

Fig. 9: Car sales in Italy time series deseasonalized partial autocorrelation plot for the first 24 lags

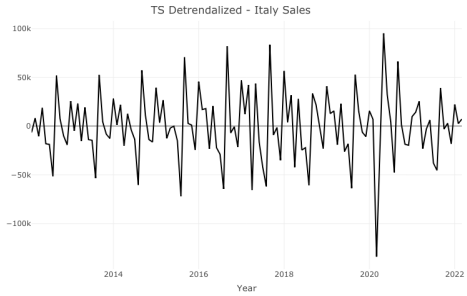

Fig. 10: Car sales in Italytime series detrendalized from January 2012 until March 2022

not as a ACF plot of a white noise (where the significant lag is only the first one). The last plot reported in Figure 13 shows the relationship between the time series and its lags. It's clear how strong is the linear relation with the annual lag (lag 12), that confirms the ACF/PACF values with this lag reported in the previous plots.

B. Time Series Model Selection

Now that the data analysis is finished it can be possible to proceed with the first model analyzed: **HWES** (Holt-Winter's Exponential Smoothing Model). It is a generalization of the simple exponential smoothing, that is similar to forecasting with a moving average, except that instead of taking a simple average over the k most recent values, we take a weighted average of all past

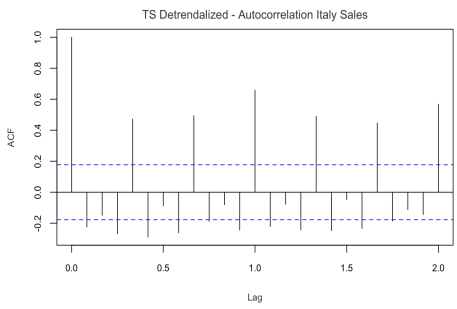

Fig. 11: Car sales in Italy time series detrendalized autocorrelation plot for the first 24 lags

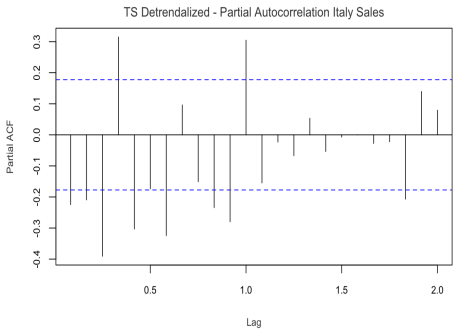

Fig. 12: Car sales in Italy time series detrendalized partial autocorrelation plot for the first 24 lags

values, so that the weights decrease exponentially into the past.

The idea is to give more weight to recent information, yet not to completely ignore older information. Like the moving average, simple exponential smoothing should only be used for forecasting series that have no trend or seasonality. Series that have only a level and noise. One solution for forecasting series with trend and/or seasonality is first to remove those components (via differencing). Another solution is to use a more sophisticated version of exponential smoothing, which can capture trend and/or seasonality.

For series that contain both trend and season-

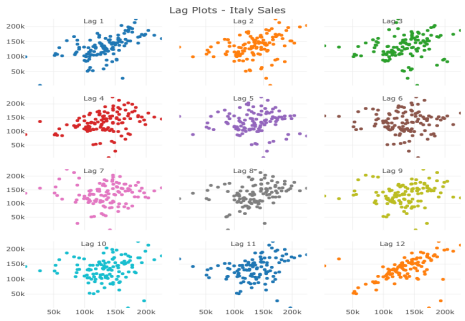

Fig. 13: Car sales in Italy time series lags plot for the first 12 lags

ality like the car sales in Italy, the Holt-Winter's exponential smoothing method can be used. It can modeled the trend/seasonality with three type of configuration: additive, multiplicative or null (this when the component is not detected as influence/present in the series). There are also two types of errors that an exponential smoothing model can capture: the additive error case, the errors are assumed to have a fixed magnitude, irrespective of the current level and trend of the series. If, however, it is more reasonable that the size of the error grows as the level of the series increases, then we will want to use a multiplicative error [6].

It was used the automated model selection for the HWES in order to find which type of error, trend and seasonality the algorithm finds. It uses the AIC (Akaike's Information Criterion), which combines fit to the training data with a penalty for the number of smoothing parameters and initial values included in a model. The result was a model with additive error, null trend and additive seasonality. The most particular was the trend result, since it seemed that a slight trend was present. For validating this result it was used the WAVK statistical test that has as null hypothesis that the series doesn't have a non-monotonic trend. This test is based on the calculation of autoregression (AR) parameters, and the chosen method to do it was the Burg one. This because in some special cases the Yule-Walker estimation method leads to poor parameter estimates, even for moderately sized data samples. Least squares should not be used either, as it may lead to an unstable model. Burg's method is preferable [2]. The standard method used by the function *notrend_test()* is the Hall and Van Keilegom one, which used a different type of formula in order to calculate autocovariates but still using the Yule–Walker method to derive autoregression coefficients [9].

The test function gives the possibility to tune parameters such as the order of the AR. It can be found thanks to the PACF plots of the car sales in Italy (Figure 6). The order p of the AR corresponding to the lags of the PACF plots that are cutting off [4]. In this case. the appropriate order is two and the p-value obtained was equal to 0.137 hence the null hypothesis (it's not present a non-monotonic trend) can't be reject and the result of the auto hwes model is less strange. Furthermore, a simple stochastic model can "explain" the correlation and trends in some series [1], hence this case can be an example.

Another unexpected result was the additive seasonality, since from the plots of the time series (Figure 1) and the method of coefficient of variation of seasonal differences and quotients the series has at least a multiplicative component. Since the trend has tested not evident, the multiplicative contribute has to be of the seasonality,

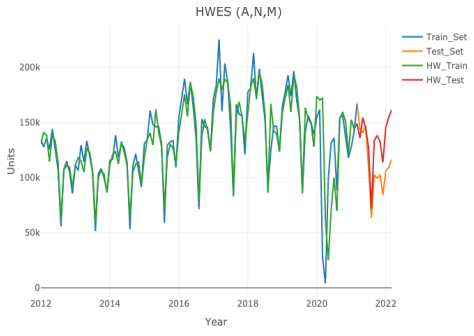

Fig. 14: HWES train and forecasting with additive error, null trend and multiplicative seasonality

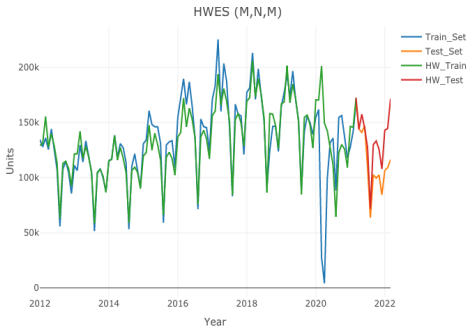

Fig. 15: HWES train and forecasting with multiplicative error, null trend and multiplicative seasonality

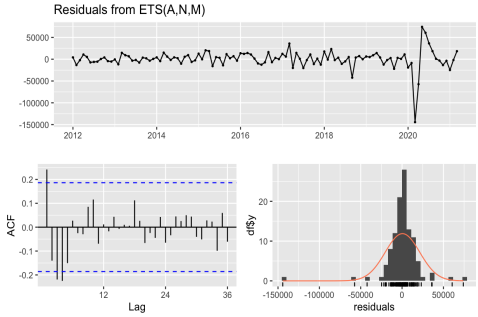

Fig. 16: HWES residuals plots with additive error, null trend and multiplicative seasonality

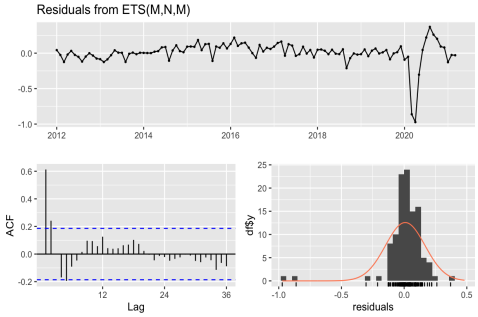

Fig. 17: HWES residuals plots with multiplicative error, null trend and multiplicative seasonality

hence it was fixed as multiplicative. The check was done between a HWES model with additive and multiplicative error.

The train set is from January 2012 until March 2021 while the test set the remaining year. In Figure 14 the results of the train and the forecasting on the HWES with additive error are reported, while in Figure 15 the ones of the HWES with the multiplicative error. In order to validate time series forecasting models it's important to analyze the residuals and do two important statistical tests: the Ljung-Box (for the independence of the series observations) and Shapiro-Wilk (for the normality of the residuals).

In Figure 16 and Figure 17 the residuals of the HWES with additive error and the HWES multiplicative error respectively are reported. It can be seen that both the normality (from the histogram) and the independence (from the correlogram) seem to be not respected. Indeed, in Table II the results of the tests and the metrics

TABLE II

HWES (A,N,M)				
P-Value L-B	P-Value S-W	AIC	BIC	AICc
0.0007189	$2.663 \cdot 10^{-13}$	2759.94	2800.58	2764.99
HWES (M,N,M)				
P-Value L-B	P-Value S-W	AIC	BIC	AICc
$1.226 \cdot 10^{-10}$	$3.154 \cdot 10^{-14}$	2755.71	2796.35	2760.76

P-Value of the Ljung-Box and Shapiro-Wilk tests and AIC, BIC, and AICc of the HWES model with additive/multiplicative error, null trend and multiplicative seasonality

such as AIC, BIC and AICc show how these models aren't good for this series. It's important to note that the model with multiplicative error has better metrics, even if the Ljung-Box p-value is much higher then the model with additive error.

The second model analyzed is the **SARIMA** (Seasonal Autoregressive Integrated Moving Average). It's a generalization of the standard ARIMA model, where there are: an autoregressive (AR) process (establish a relationship between the series and its past p lags with the use of a regression model), a moving average (MA) process (Similar to the AR process, the MA process establishes the relationship with the error term at time t and the past q error terms, with the use of regression between the two components) and an integrated (I) process (the process of differencing the series with its d lags to transform the series into a stationary state. In order to build an ARIMA model, the times series has to be with a trend component but without seasonality. For this it was used the extension SARIMA, that in addition to the non-seasonal parameters for the ARIMA model (p,d,q) requires other three seasonal parameters (P,D,Q): the SAR(P) process that is a seasonal AR process of the series with its past P seasonal lags, th SMA(Q) process that is a seasonal MA process of the series with its past Q seasonal error terms and a seasonal differencing of the series with its past D seasonal lags [4].

Three SARIMA models were built, one with the help of *auto.arima()* function which finds

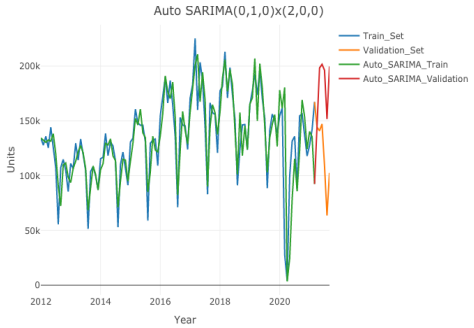

Fig. 18: Train and validation of the auto SARIMA model with parameters (0,1,0)x(2,0,0)

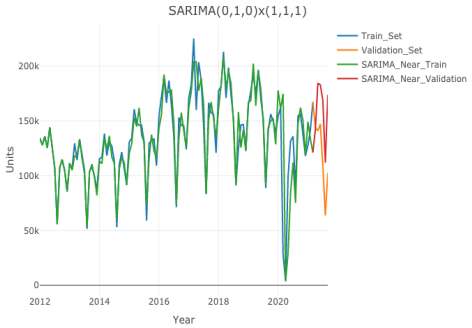

Fig. 20: Train and validation of the SARIMA model with parameters (0,1,0)x(1,1,1)

the best non-season and season parameters for the AIC, BIC and AICc metrics. The second one is a SARIMA model where the parameters have been choosen near to the auto ones in order to find better metrics values. The last one is based on the ACF/PACF plots of the original and deseasonalized series, as the theory wants. This one was the best one, hence here will be reported the other two models. The train set was fixed from January 2012 until March 2021, the validation set goes from March 2021 until September 2021 and the test set from September 2021 until March 2022.

The auto SARIMA model finds the (0,1,0) non-seasonal parameters and the (2,0,0) seasonal parameters. The non-seasonal model is equivalent to a random walk [4], while the seasonal model is a simple seasonal AR with two seasonal lags. The SARIMA model near to the auto one has the (0,1,0) non-seasonal parameters and the (1,1,1) seasonal parameters. These choices have been made based on the values of the metrics AIC and BIC.

The train and validation results are reported in Figure 18 for the auto SARIMA model, while in Figure 20 for the SARIMA model near to the auto one. It can already see that the forecasting of the auto SARIMA is much farther from the real observations then the other model. The check of the residuals is in Figure 19 for the auto SARIMA model and in Figure 21 for the SARIMA model near to the auto one. For both models the lag plot shows with a good probability the independence of the observations in the series.

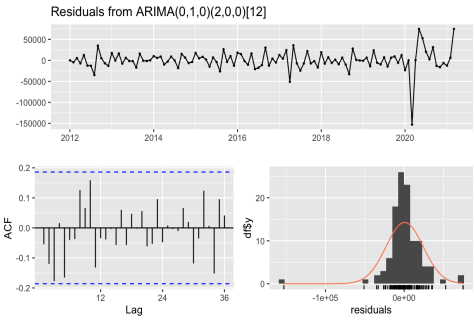

Fig. 19: Residuals plots of the auto SARIMA model with parameters (0,1,0)x(2,0,0)

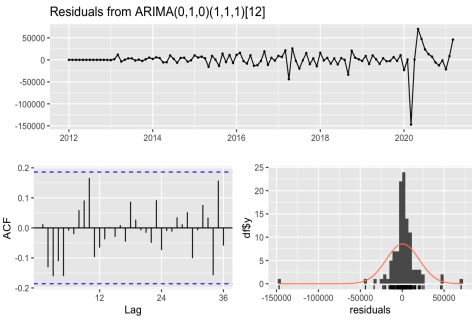

Fig. 21: Residuals plots of the SARIMA model with parameters (0,1,0)x(1,1,1)

TABLE III

Auto SARIMA (0,1,0)x(2,0,0)			
P-Value L-B	P-Value S-W	AIC	BIC
0.4593	$2.855 \cdot 10^{-11}$	2544.37	2552.47
SARIMA Near (0,1,0)x(1,1,1)			
P-Value L-B	P-Value S-W	AIC	BIC
0.5667	$1.027 \cdot 10^{-14}$	2256.41	2264.16

P-Value of the Ljung-Box and Shapiro-Wilk tests, AIC and BIC of the auto SARIMA model and the SARIMA model near to the auto one

TABLE IV

Auto SARIMA (0,1,0)x(2,0,0)			
Set	MAE	RMSE	MAPE
Train	14247.40	23462.75	15.48494
Validation	65053.21	72025.55	65.81618
SARIMA Near (0,1,0)x(1,1,1)			
Set	MAE	RMSE	MAPE
Train	9789.521	19638.63	11.12877
Validation	43417.129	48579.44	42.75743

Mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) of the auto SARIMA model and the SARIMA model near to the auto one

Histograms, ideed, are not normal. In order to confirm these conclusions the statistical tests have been made. In Table III can be seen the p-values for the two tests, and for both the models the the Ljung-Box and Shapiro-Wilk null hypothesis are maintained and refused respectively, as previously mentioned. AIC and BIC are better for the SARIMA model near the auto one. In Table IV are reported the scores used to valuate the two models: the mean absolute error (MAE) which gives the magnitude of the average absolute error, the root mean squared error (RMSE) which measure has the same units as the data series and the mean absolute percentage error (MAPE) which gives a percentage score of how forecasts deviate (on average) from the actual values [6].

In both the train and validation set the SARIMA model near to the auto one is the best, even if its scores are low in general. The model has a percentage error (on average) about 42.76% on the validation set. It's too high, such as the RMSE and the MAE, and this is a signal of a likely overfitting.

The third and the last model in analysis is the **SARIMAX**. A further generalization of the ARIMA base model where the seasonality and an exogenous regressor can be taken into consideration. The main SARIMA model used is the best of the three built before (the one with (2,1,3) non-seasonal parameters and (1,1,1) seasonal parameters). Afterwards, two regressors have been tested, the FTSE MIB (acronym for Financial Times Stock Exchange Milano Indice di Borsa), that is the most significant stock index of the Italian Stock Exchange, and the unemployment rate in Italy. Both are analyzed from January 2012 until March 2022, the MIB data have been

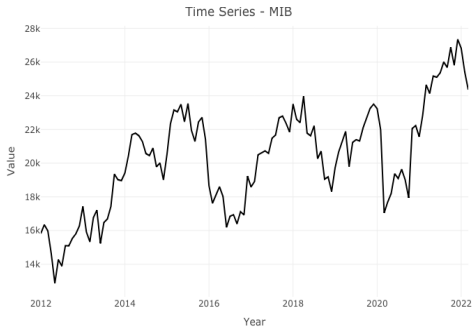

Fig. 22: The FTSE MIB (Financial Times Stock Exchange Milano Indice di Borsa) from January 2012 until March 2022

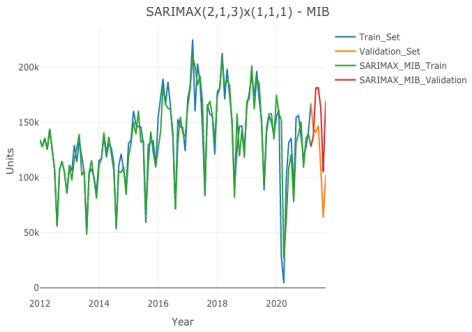

Fig. 23: Train and validation of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the MIB as exogenous regressor

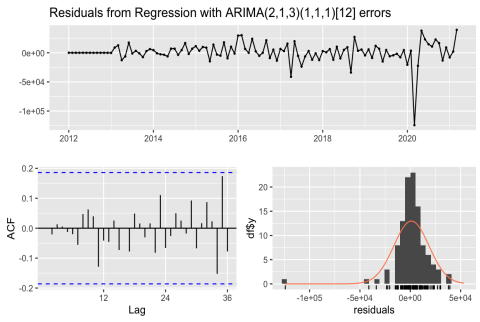

Fig. 24: Residuals plots of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the MIB as exogenous regressor

taken from *Marketwatch* website [10] while the unemployment rate from the *Istat* website [11]. In particular, the last one is the rate of the population between 15-74 years old.

From this type of models is found the second best model of the study, the one with the unemployment rate as exogenous regressor, hence here will be reported the model with the MIB as regressor. In Figure 22 is shown the MIB time series, while in Figure 23 is reported the train and forecasting results on the training and validation set as defined in the previous SARIMA models.

The results seems to be similar to the SARIMA (0,1,0)x(1,1,1), indeed in Table V the scores are slightly better but near to the SARIMA model. In Table VI AIC value are lower while the BIC slightly higher then the SARIMA "near" model. The value of BIC was predictable since the SARIMAX has a lot of paramaters more then the SARIMA "near", indeed a part of the exogenous regressor there are also the seasonal and non-

TABLE V

SARIMAX(2,1,3)x(1,1,1) with MIB			
Set	MAE	RMSE	MAPE
Train	10165.97	17405.24	15.35067
Validation	40622.73	44762.82	39.44043

Mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the MIB as exogenous regressor

TABLE VI

SARIMAX(2,1,3)x(1,1,1) with MIB			
P-Value L-B	P-Value S-W	AIC	BIC
0.8935	$1.86 \cdot 10^{-12}$	2243.7	2266.96

P-Value of the Ljung-Box and Shapiro-Wilk tests, AIC and BIC of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the MIB as exogenous regressor

seasonal parameters of the best model used as baseline for the SARIMAX. Known that, the BIC (which penalize more then the AIC the use of several parameters) still near to the SARIMA model. This is clear a prove that this model is already better then the previous one. Also the scores are slightly better then the SARIMA, how it can see in Table V. The scores are still too high in general since the MAPE is about 39.44% on the validation set. The check of the residuals in Figure 24 has the same results of the previous model, that is the independence can be confirmed with a higher probability level then the SARIMA but again the residuals are not normal.

III. Results and Discussions

The two chosen model as the best ones are the SARIMA theory based model and the SARIMAX with the unemployment rate (Unr) as exogenous regressor. In Figure 25 is reported the unemployment rate between 15-74 years old in Italy. The SARIMA model is called theory based since its parameters have been chosen following the

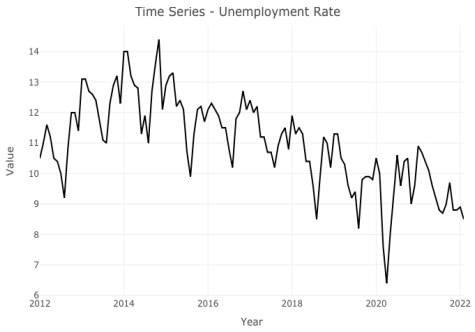

Fig. 25: Unemployment rate between 15-74 years old in Italy

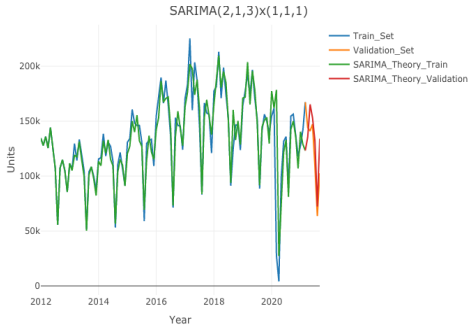

Fig. 26: Train and validation of the SARIMA model with parameters (2,1,3)x(1,1,1)

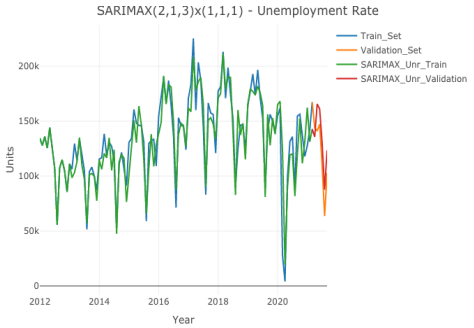

Fig. 27: Train and validation of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the unemployment rate as exogenous regressor

standard indications to find them. In particular, the order 2 of the AR corresponding to the lags of the PACF plot of the time series deseasonalized (Figure 9) that are cutting off, the order 1 of the I corresponding to the first order lag-1 differencing of the non-seasonal series, in order to make it stationary, and the order 3 of the MA corresponding to the lags of the ACF plot of the time series deseasonalized (Figure 8) that are cutting off [4]. The seasonal parameters follow the same rules, and looking to the ACF/PACF of the seasonal series (Figure 5 and Figure 6 respectively) both SAR and SMA have order 1. The order 1 of the SI is the lag-12 differencing of the seasonal series, in order to make it stationary, indeed with two lags difference the series is stationary.

In Figure 26 and in Figure 27 are reported the train and validation of the SARIMA theory based and SARIMAX with unemployment rate models respectively. The plots show how well these model predict the validation set compared to their competitors models. This is clearly confirmed by the scores in Table VII, where all the values in the validation set are the best, in particular the MAPE of both models is about the 15.46% for the SARIMA and 18.32% for the SARIMAX, an huge improvement of the performances. The plots of residuals (Figure 28 for the SARIMA and Figure 29 for the SARIMAX) still show a non-normal distribution, confirmed by the p-values of the tests in Table VIII. More precisely, the Ljung-Box null hypothesis is confirmed with an higher probability level for the SARIMA then the SARIMAX, at the same time the Shapiro-Wilk

TABLE VII

SARIMA(2,1,3)x(1,1,1)			
Set	MAE	RMSE	MAPE
Train	8995.061	18151.56	15.26784
Validation	17037.117	19587.28	15.46243
SARIMAX(2,1,3)x(1,1,1) with Unr			
Set	MAE	RMSE	MAPE
Train	10607.78	15945.15	13.61187
Validation	18798.90	19667.10	18.31904

Mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) of the SARIMA model with parameters (2,1,3)x(1,1,1) and the SARIMAX model with the same parameters and the Unr as exogenous regressor

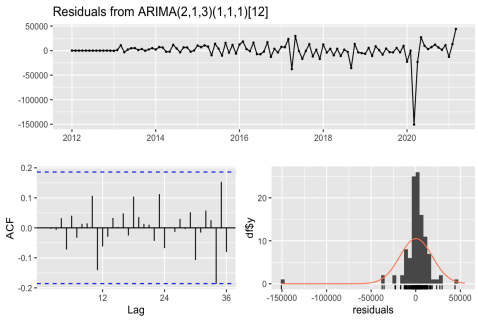

Fig. 28: Residuals plots of the SARIMA model with parameters (2,1,3)x(1,1,1)

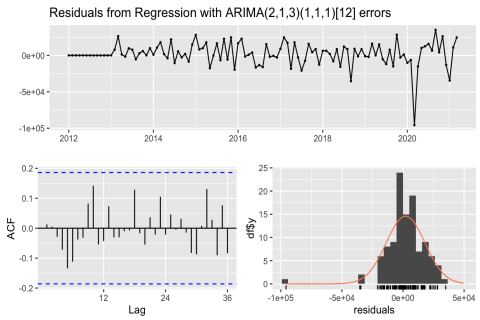

Fig. 29: Residuals plots of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the unemployment rate as exogenous regressor

null hypothesis is rejected by both models but the SARIMA one has a much lower p-value then the SARIMAX. The AIC and BIC metrics are both better for the SARIMAX model, even if it has an extra parameter (the regressor). For the validation set seems that the SARIMAX is the best model, since has two better scores and 3 out of 4 better values between tests and metrics.

Hence it has been proofed that these two models are the best compared to the others, now they will be valuated with the test set from September 2021 until March 2022 while the previous validation set will be a part of the new train set that start in January 2012 and end in September 2021. In Figure 30 and in Figure 31 are reported the results for the SARIMA and SARIMAX models

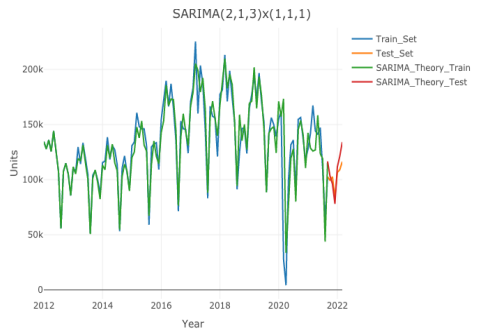

Fig. 30: Train and validation of the SARIMA model with parameters (2,1,3)x(1,1,1)

TABLE VIII

SARIMA(2,1,3)x(1,1,1)			
P-Value L-B	P-Value S-W	AIC	BIC
0.8782	$5.915 \cdot 10^{-16}$	2252.2	2272.88
SARIMAX(2,1,3)x(1,1,1) with Unr			
P-Value L-B	P-Value S-W	AIC	BIC
0.5908	$4.905 \cdot 10^{-9}$	2233.05	2256.31

P-Value of the Ljung-Box and Shapiro-Wilk tests, AIC and BIC of the SARIMA model with parameters (2,1,3)x(1,1,1) and the d the SARIMAX model with the same parameters and the Unr as exogenous regressor

respectively. Both the models seem to have improved their performances, indeed in Table IX all the values of the SARIMAX models in both the set are better then the results with the previous train and validation sets (except for the RMSE in the previous train set). Also the SARIMA model in the test set has better results then the same model but in the previous validation set, while it has better performances with the previous train set rather than the new one. In particular the SARIMA has a MAPE about 8.14% while the SARIMAX about 6.45%, these are very good results. The check of the residuals is in Figure 32 for SARIMA and in FIgure 33 for SARIMAX and even in this case the independence seems to be respected but the normality not. This is confirmed by the values in Table X.

It can note that the p-values and AIC/BIC of the SARIMAX are better then the SARIMA

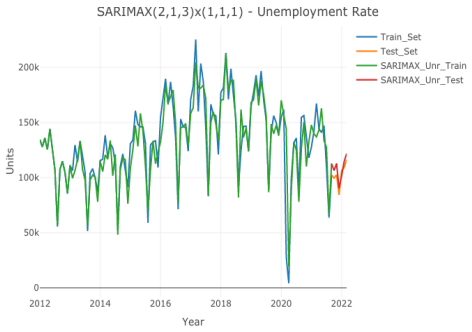

Fig. 31: Train and test of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the unemployment rate as exogenous regressor

TABLE IX

SARIMA(2,1,3)x(1,1,1)			
Set	MAE	RMSE	MAPE
Train	9329.957	17628.486	16.367974
Test	8604.020	9882.253	8.143556
SARIMAX(2,1,3)x(1,1,1) with Unr			
Set	MAE	RMSE	MAPE
Train	10250.30	16486.468	13.522537
Test	6551.21	6780.918	6.447388

Mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) of the SARIMA model with parameters (2,1,3)x(1,1,1) and the SARIMAX model with the same parameters and the Unr as exogenous regressor

model hence, after all this checks, it can conclude that the SARIMAX is in general the best model to forecast the car sales in Italy.

IV. Conclusion

In the end, this study gives back two good models for forecasting the car sales in Italy. The SARIMA model theory based has an huge improvement of the performances compared to the others SARIMA models and, adding an appropriate exogenous regressor, the SARIMAX was able to reach higher performances. The problem of the non-normality of the residuals remains, but is clear that the slump caused by the pandemic has an important contribution about that. Indeed, checking the residual time series for each of the 7 models built it's evident that the 2/3 months

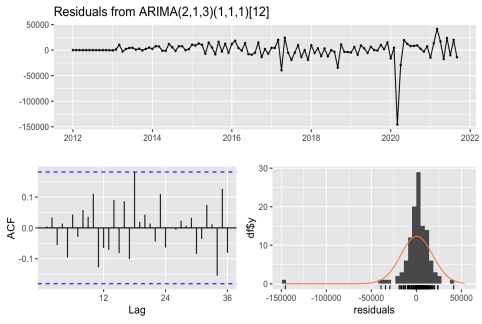

Fig. 32: Residuals plots of the SARIMA model with parameters (2,1,3)x(1,1,1)

TABLE X

SARIMA(2,1,3)x(1,1,1)			
P-Value L-B	P-Value S-W	AIC	BIC
0.2916	$1.37 \cdot 10^{-15}$	2385.19	2406.35
SARIMAX(2,1,3)x(1,1,1) with Unr			
P-Value L-B	P-Value S-W	AIC	BIC
0.3949	$3.891 \cdot 10^{-12}$	2368.93	2392.73

P-Value of the Ljung-Box and Shapiro-Wilk tests, AIC and BIC of the SARIMA model with parameters (2,1,3)x(1,1,1) and the d the SARIMAX model with the same parameters and the Unr as exogenous regressor

during the lockdown aren't detected well. However, it must be considered that if a model that was too precise was built in the train set, there was probably overfitting in the validation/test set. For this reason the Shapiro-Wilk test would be difficult to be passed (i.e. the null hypothesis has to be rejected).

The use of the unemployment rate is often used when the time series in analysis describes a sale/purchase of tangible property, since the rate can be a good representation of the economical situation of the population in the country. Several books propose this approach in order to improve the performances of forecasting models. An idea could be to use the number of hospitalizations from 2012 but, excepted for the pandemic period, this can't be good explanatory regressor since the automotive market had ups and downs caused not by the sanitary conditions of the population but

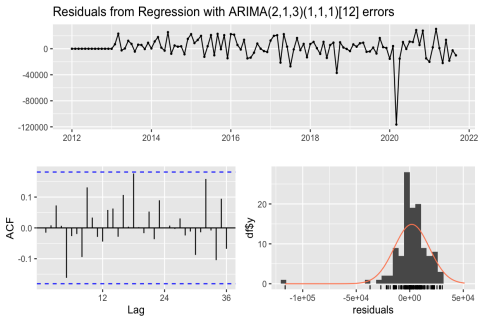

Fig. 33: Residuals plots of the SARIMAX model with parameters (2,1,3)x(1,1,1) and the unemployment rate as exogenous regressor

e.g. the tax increase. This regressor, probably, would cause overfitting in the pandemic period and would increase the error in the previuos period.

REFERENCES

[1] P. S. P. Cowpertwait and A. V. Metcalfe. *Introductory Time Series with R*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[2] M. de Hoon, T. van der Hagen, H. Schoonewelle, and H. van Dam. Why yule-walker should not be used for autoregressive modelling. *Annals of Nuclear Energy*, 23:1219–1228, 1996.

[3] E. Dennis, I. Iwueze, M. Ijomah, and T. Owolabi. Methods for choice of model in descriptive time series : A review with example. *International Journal of Advanced Statistics and Probability*, 6:10, 12 2017.

[4] R. Krispin. *Hands-On Time Series Analysis with R: Perform Time Series Analysis and Forecasting Using R*. Packt Publishing, 2019.

[5] F. Lazzeri. *Machine Learning for Time Series Forecasting with Python*. Wiley, 2020.

[6] G. Shmueli and K. Lichtendahl. *Practical Time Series Forecasting with R: A Hands-On Guide [2nd Edition]*. Practical Analytics. Axelrod Schnall Publishers, 2016.

[7] W. Woodward, H. Gray, and A. Elliott. *Applied Time Series Analysis with R*. CRC Press, 2017.

[8] http://www.unrae.it/dati-statistici/immatricolazioni.

[9] https://cran.r-project.org/web/packages/funtimes/funtimes.pdf.

[10] https://www.marketwatch.com/investing/index/i945/download-data?countrycode=it.

[11] http://dati.istat.it/Index.aspx?DataSetCode=DCCV_TAXDISOCCUMENS1#.