Bike-sharing forecasting project

Group - Morges
Group members: Jiaming Huang(21421110), Manunpat Sirijaturaporn (21430814), Ting Yang(21427620)
Date: May 24th, 2022

Introduction

The purpose of this project is to forecast the total number of available bikes for every 10 minutes at 8 bike-sharing stations in Morges, including Medtronic, Moulin, Gracieuse, Préverenges, Sablon, Casino, Dufour, and Temple, on Wednesday, May 25th, 2022. The collected data is from a large Swiss bike-sharing company from March 4th to May 24th over 10 minutes intervals. In this project, several methods, such as Naïve, ETS (Error, Trend, Seasonal), and ARIMA, from Forecasting I are applied. In addition, we also apply ARIMA with the *Fourier* terms to capture complex seasonality.

Exploratory Data Analysis

It is difficult to interpret the time series of each station from the time series plot (Figure 1), so, first, we apply the "gg_season" function to plot the daily and hourly patterns of those stations (Figure 2 and Figure 3). Then, we apply the STL decomposition to examine the components of these time series (Figure 5 - Figure 12). However, it is still not an ideal way to obtain clear trends and seasonalities of every station. Thus, we apply spectrograms to further analyze the seasonality.

- **Trend**
  Regarding the STL decomposition plots (Figure 5 - Figure 12), overall, the trend components fluctuate around the means. However, we observe that Medtronic, Moulin, and Sablon have a strong downward trend after May 7th, but we don't observe such a trend at the other stations.

- **Seasonality**
  Regarding the Moulin and the Medtronic stations, we can observe strong weekly and daily seasonality patterns from the seasonality plots (Figure 2 and Figure 3) and the STL decomposition (Figure 8 - Figure 9). From the seasonality plots, it seems there are high usages at these stations during weekdays, especially around 8 am and 5 pm and the seasonality between these 2 stations seems to be negatively correlated. If we analyze these stations in more detail, we find that Moulin is located close to the train station, while Medtronic is located around 2km further from the city center and there are several offices and schools around the Medtronic station. Thus, we believe that the seasonality that we observe at 8 am and 5 pm during weekdays is because people take bikes from Moulin and commute to Medtronic to work or study in the early morning and ride back to the train station (Moulin) in the late afternoon.

  For those seasonal components observed from STL components of the other stations (from Figure 5 -8 and Figure 11- 12), there seems to be no standard seasonal pattern, but we can observe that they are still affected by a small amount of weekly, daily and even hourly seasonal components. In this case, we apply a spectrogram (periodogram) function to help us to further detect and extract seasonality through *Fourier* Transform. Then, we select the two highest "power" frequencies with a periodogram and add them to the ARIMA model.

- **Residuals**
  For every station (Figure 5 – Figure 12), we can notice that the remainder is still very large. This means the remainder accounts for a large part of the variation of data.

- **Irregularities and outliers**
  From the box plots (Figure 4), we can observe that Gracieuse and Préverenges have many outliers on the larger side. The other stations have a few/no outliers. In addition, except for Préverenges, and Moulin, the medians of most stations deviate from the center of the box. In this case, we can conclude that most of them show a strong level of skewness. Furthermore, Medtronic, Moulin, Sablon and Préverenges have wider boxes. This indicates that they are volatile and have a high degree of fluctuation.

Modelling and Predictions

Abstract

Overall, we build several models, such as Mean, SNaive, auto ARIMA, ARIMA with *Fourier*, and ETS combining with STL decomposition models, for each station. Then, we perform cross-validation and compare accuracy across all models to choose the best model. The Mean model and the SNaive model are used as benchmarks. We also refer to STL decomposition and spectrogram analysis of time series, except for time series plots to build and choose the models.

As the data of most stations have unclear and seemingly complex seasonality patterns, we choose the Dynamic harmonic regression with multiple seasonal periods (ARIMA with *Fourier*) from the fpp3 package. during the process of building models. To use this model, we set D=d=0 and P=Q=0 since non-stationarity and seasonality can be handled by the model through regression terms, but we use the model in conjunction with the periodical signal captured by the spectrogram. In addition, ideally, we should try several K *Fourier* terms, which is the number of cos and sin pairs, and select the model that has the lowest AIC. However, when we increased K, we found that the processing time in R also increased considerably. Thus, in this project, we assign K equal to 1 to make the processing time of the model faster except for the Préverenges station, however; the prediction will be much smoother. This can be improved in further studies.

In addition, we use the ETS combining with STL decomposition model as well, since there are three seasonal patterns shown in STL components for each station. In this case, using the decomposition in forecasting is also a suitable way to handle these stations' data, with each of the seasonal components forecast using a seasonal naïve method, and the seasonally adjusted data forecast using ETS.

We then choose the best models that had the best RMSE. In this step, we might choose more than one model if they have similar performance in terms of RMSE.

After choosing the best models, we add the regression part, including the temperature, rain, and wind speed variables, into the model. The available variables include:

- tre200s0(unit:°C): air temperature at 2 m from the ground; instantaneous value.
- tvi200s0(unit:°C): virtual temperature at 2 m from the ground; instantaneous value
- rre150g0(unit:mm): precipitation; cumulative hourly sum (over 6 intervals of 10 min.)
- rre150z0(unit:mm): precipitation: summation over 10 minutes
- fkl010z0(m/s): scalar wind speed; average over 10 minutes
- fu3010z0(km/h): wind speed; average over 10 minutes

To avoid a multicollinearity problem, we analyze correlations between these variables (Figure 21) and find that the temperature variables (tre200s0 and tvi200s0) are perfectly correlated, so we can choose either of them. We also observe the perfect correlation between the wind speed variables (fkl010z0 and fu3010z0) as well, so we can choose either of them. For the precipitation variables, they are highly correlated with a correlation of 71%, and we decide to choose rre150z0 because we believe that the precipitation over 10 minute-interval should be more suitable for our forecast. To sum up, we choose tre200s0, fu3010z0, and rre150z0 for the regression part.

After adding the regression part into the model, we again compare the accuracy (RMSE) across the chosen models and choose the final best model for the prediction, in case there are more than one chosen models from the previous step,

Results

- **Casino**
  We indeed find that both the RMSEs and the residual characteristics of the auto ARIMA model and the ARIMA with *Fourier* model are very similar (Figure 13a and Figure 13b) during the process of building the model. The best model always shifts between them as the data is constantly updated. However, their forecasts are almost the same. We choose the ARIMA with *Fourier* model with regressors as the best model because it can handle complex seasonality.

- **Dufour**
  For Dufour, the ETS combining STL decomposition model returned high accuracies in the model selection at the very beginning of the analysis. However, the ARIMA with the *Fourier* model takes the place because of the slightly lower RMSE (Figure 14a). Thus, we add all the regressors to the ARIMA with the *Fourier* model. We then perform the residual examination and obtain a fair result of it (Figure 14b). To be precise, the residuals do not have a clear autocorrelation; nevertheless, they are not normally distributed especially on the tails. Considering the nature of this station, without a clear seasonal pattern, we choose the ARIMA with the *Fourier* model as the best model to perform the forecast.

- **Temple**
  The SNaive model performs as well as the auto ARIMA model in terms of the RMSE (Figure 15a) when we build the models for the Temple station. However, we select the auto ARIMA as the best model since the SNaive model returns

an undesirable consequence of its residual (Figure 15b) in the end. There is substantial remaining autocorrelation in the residuals for the SNaive model, which means there is information left in the residuals; nevertheless, the auto ARIMA model provides a better result (Figure 15c) when we examine its residuals. Thus, we apply auto ARIMA with regressors in the forecast. Additionally, the results of a residual examination of both the auto ARIMA model with and without regressors are similar to each other (Figure 15c and Figure 15d).

- **Medtronic**
  From the accuracy table (Figure 16a), the exponential smoothing combined with the STL decomposition model (ETS+STL) has the lowest RMSE, MAE, MASE, and RMSSE. Thus, we choose this model as the best model for our forecast. We then analyze the residuals of this model (Figure 16b) and find that, from the Ljung-Box test, the residuals exhibit serial correlation. In addition, we further test the normality of the residual using the QQ plot and find that the residuals are not normally distributed and have a very fat tail.

  Due to the strong downward trend at the Medtronic station, the forecast trend tends to be downward. Thus, in order to make sure that the forecasts of the total number of available bikes are non-negative, we set the floor of the forecast numbers to zero.

- **Moulin**
  The results are similar to the results of Medtronic. The accuracy table (Figure 18a) shows that the ETS + STL model has the lowest RMSE, MAE, MASE, and RMSSE. Thus, we choose the ETS+STL model as the best model. However, similar to the Medtronic station, the residuals (Figure 18b) exhibit serial correlation and are not normally distributed. Also, due to the strong download trend, we also apply a floor of zero to the forecast numbers to make sure that the forecast numbers will not be below zero.

- **Gracieuse**
  According to the accuracy table (Figure 18a), both the auto ARIMA model and ETS model have the lowest RMSE. However, the auto ARIMA model has a lower MASE. Moreover, it is easier to combine the ARIMA model with regressors. In the end, we choose the ARIMA model and add regressors to perform the forecast. After the model selection, we perform residual diagnostics for it. According to the residuals ACF plot (Figure 18b), the innovation residuals are uncorrelated. It shows there is no information left in the residuals that should be used in computing forecasts. And the innovation residuals have zero mean which shows the forecasts are unbiased.

- **Préverenges**
  In this station, we notice the difference in RMSE between the multi-seasonality ARIMA model and STL+ETS model is very small (Figure 19a). Because we haven't tried many possible *Fourier* terms(k value) to choose the one with the smallest AIC, it may affect the choice of the best model. Thus, we try to make k=1, k=2 and k=3 to see the difference. We find that k=3 gives the smallest RMSE to the ARIMA with a multi-seasonality model. Then we compare the

RMSE of each model and finally choose the ARIMA with the *Fourier* model. For the residual diagnostics of the best model with regressor (Figure 19b), the innovation residuals are uncorrelated and have zero mean.

- **Sablon**
  According to the accuracy table (Figure 20a), the auto ARIMA model, ETS model, and STL+ETS model have the lowest RMSE. Due to the difficulty in combining regression and the ETS model, we set aside it for further study. We only combine regression with the auto ARIMA model. We then test the residuals of these three models(auto ARIMA model with regression, ETS model, and STL+ETS model) (Figure 20b – Figure 20d). For the auto ARIMA model with regression and auto ETS model, the residuals are uncorrelated and have a mean of zero. However, the STL+ETS model has correlated residuals. It shows there is information left in the residual that should be used in computing forecasts. The model can be improved.

Conclusions and limitation

In summary, we reckon that forecasting bike-sharing data is not easy. This is because it covers multiple seasonalities and is influenced by multiple variables. Also, each station has its own unique characteristics, and some stations might be positively or negatively correlated to the others. Regarding the variables used in this study, we observe that they could only improve the models slightly. This might be because the variables don’t fluctuate much during the time horizon of our data set in our study, which covers only the spring season. If we have a longer historical time series (e.g., winter), we anticipate that these variables will play a bigger role in the prediction. In addition to the influence of the weather, we actually need to consider the distance of each station, the attributes of the area to which the station belongs, the number of bikes in transit, the average usage time, etc. Moreover, as mentioned above, there are limitations in terms of building the ARIMA with the *Fourier* model which we are supposed to improve in further steps. Furthermore, in considering the process of building the model, we did more investigation for this type of time series data forecasting finding that there are additional methods, the Prophet + XGBoost model, which might be useful to handle these data in-depth as well.

Reference

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on April 25th, 2022.

Arima with *Fourier* terms, Web traffic time series forecasting. Accessed on May 2nd, 2022, <https://www.kaggle.com/code/kailex/arima-with-fourier-terms/notebook>

XGBoost, ARIMA and Prophet for Time Series, Hourly Energy Consumption. Accessed on May 5th, 2022, <https://www.kaggle.com/code/furiousx7/xgboost-arima-and-prophet-for-time-series/notebook>

Rob J Hyndman 2012, accessed on May 16th, 2022,
<https://robjhyndman.com/hyndsight/ets-regressors/>

Appendix

Remark: All the plots and accuracy tables based on the data up to May 16th, 2022

Figure 1: Time series plot

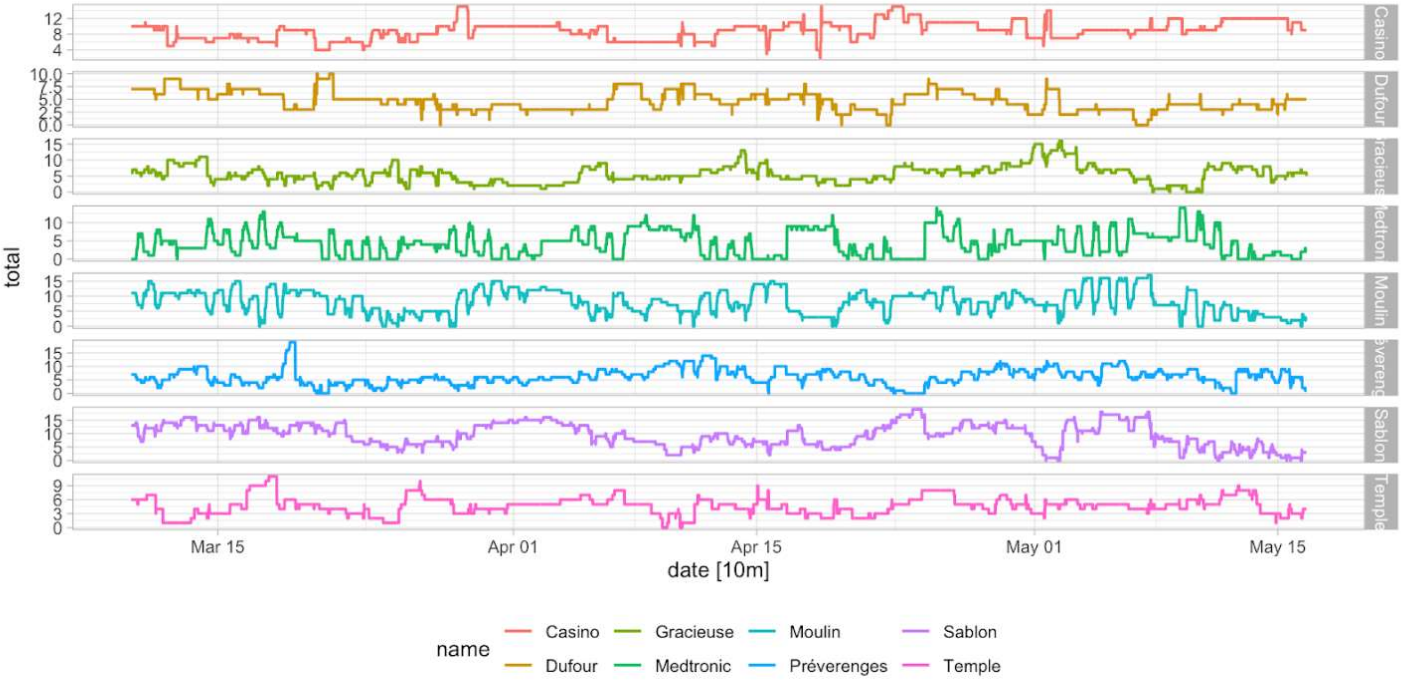


Figure 2: Daily pattern/Daily seasonality

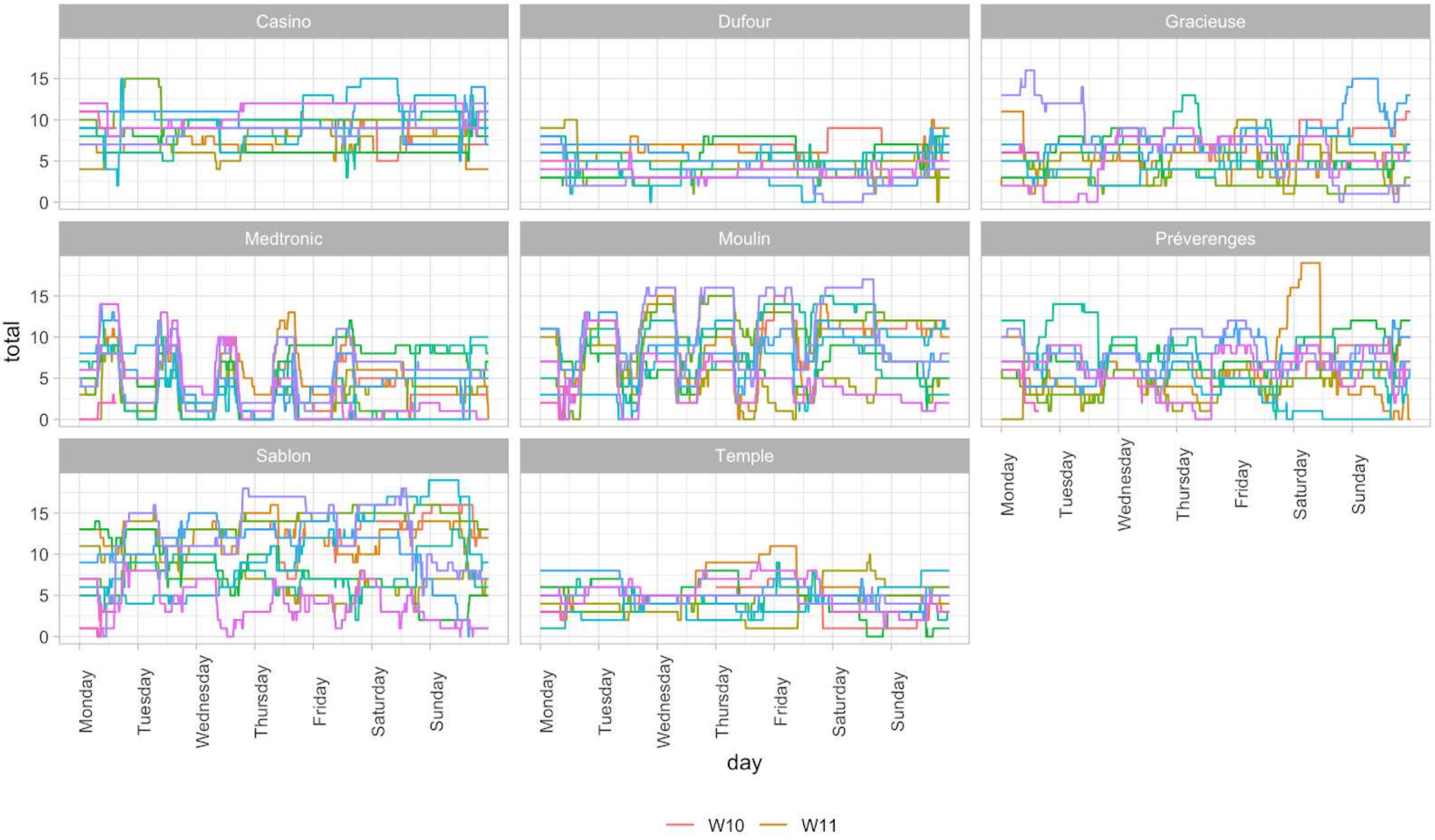


Figure 3: Hourly pattern/hourly seasonality

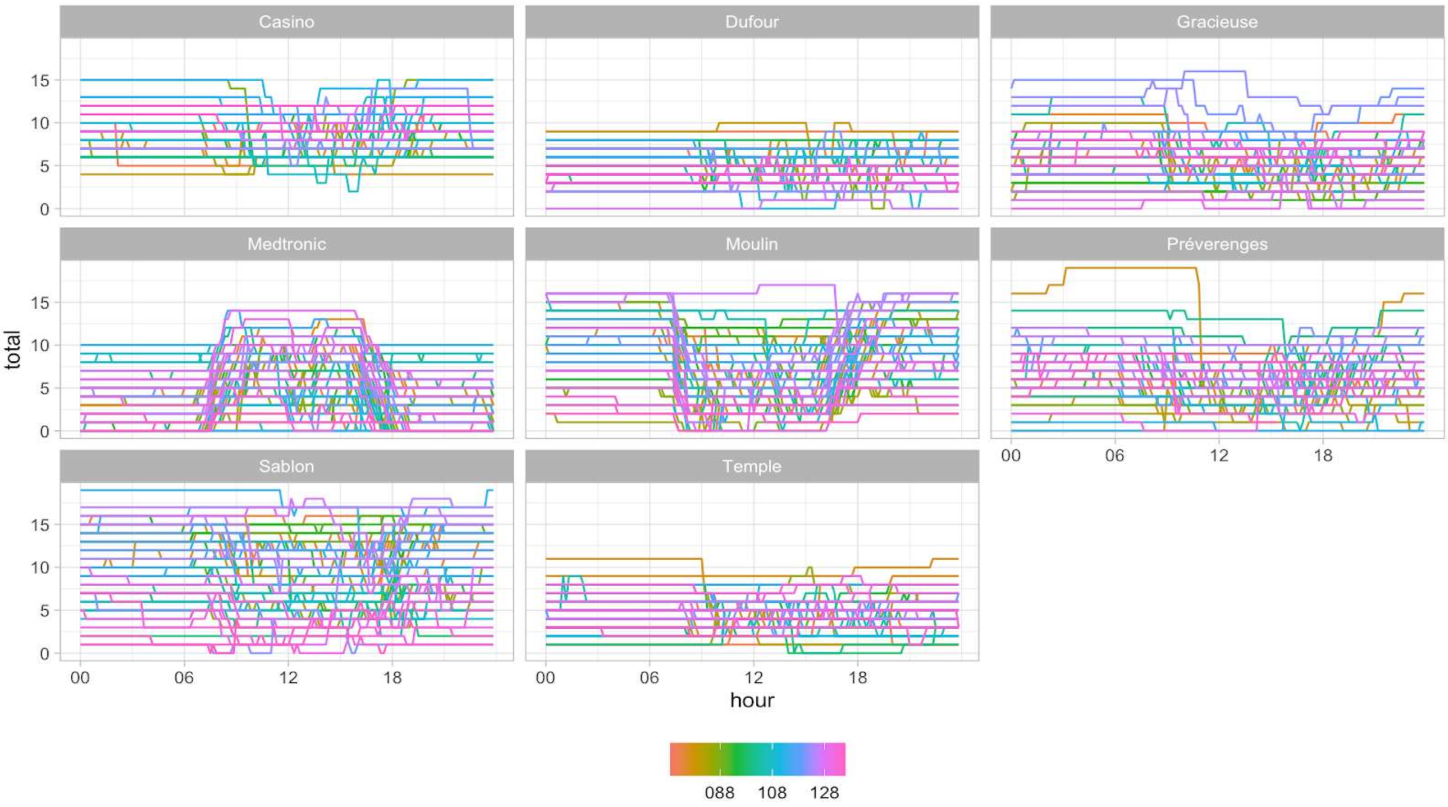


Figure 4: Box plot

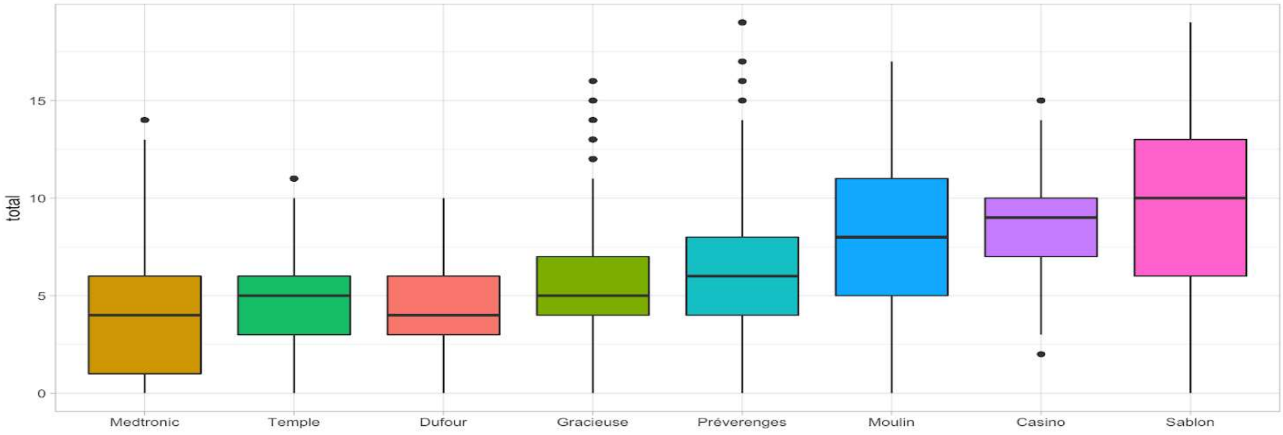


Figure 5: Casino STL component

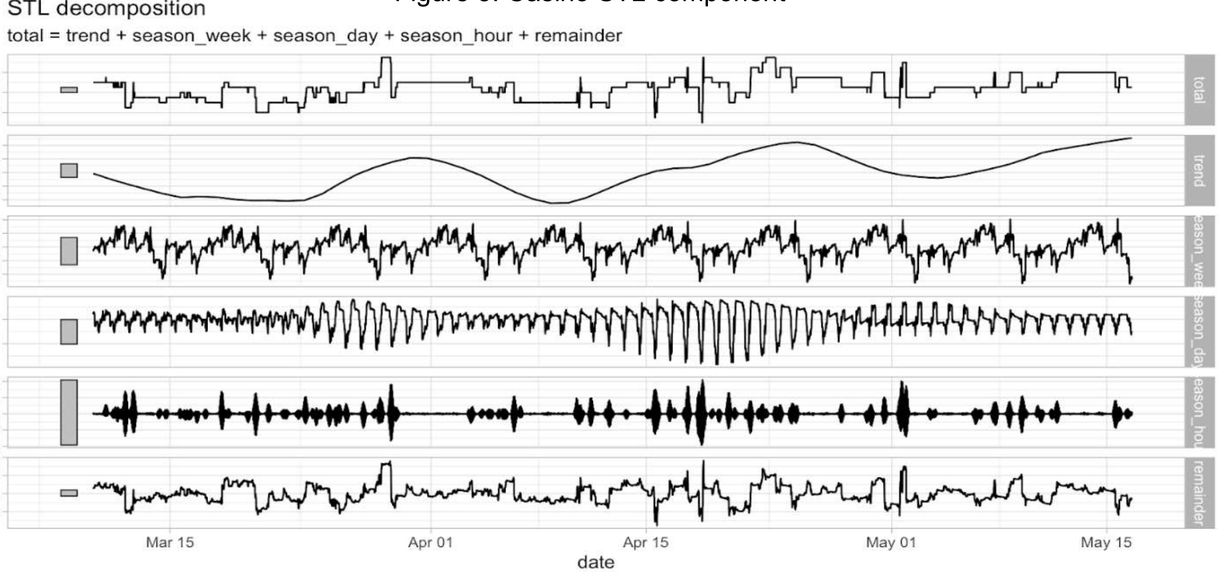


Figure 6: Dufour STL component

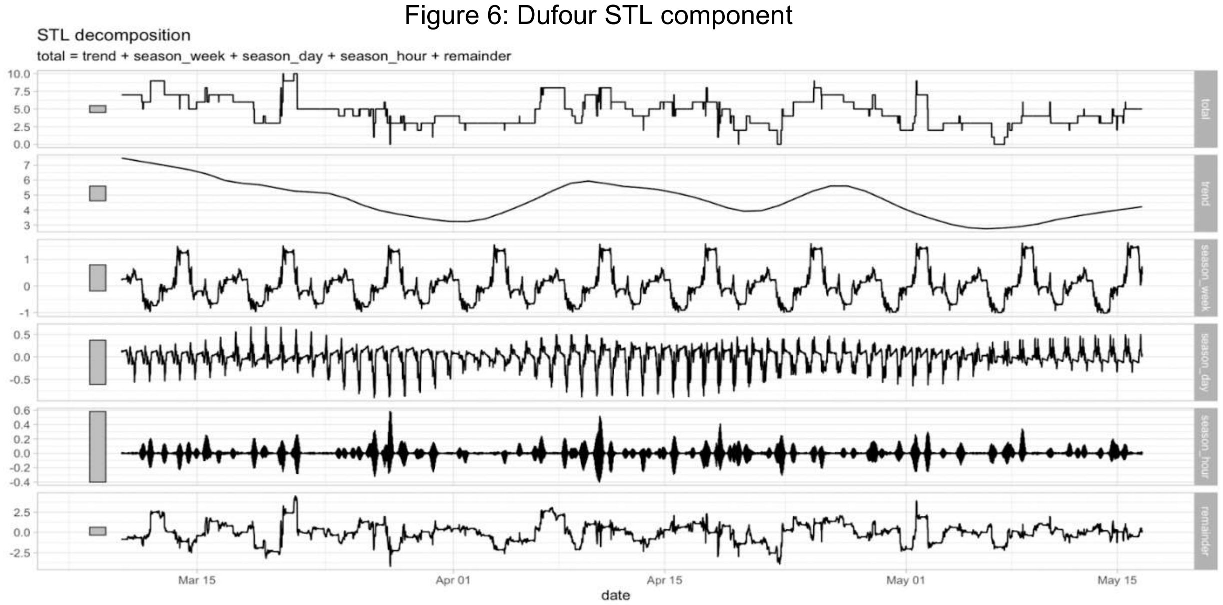


Figure 7: Temple STL component

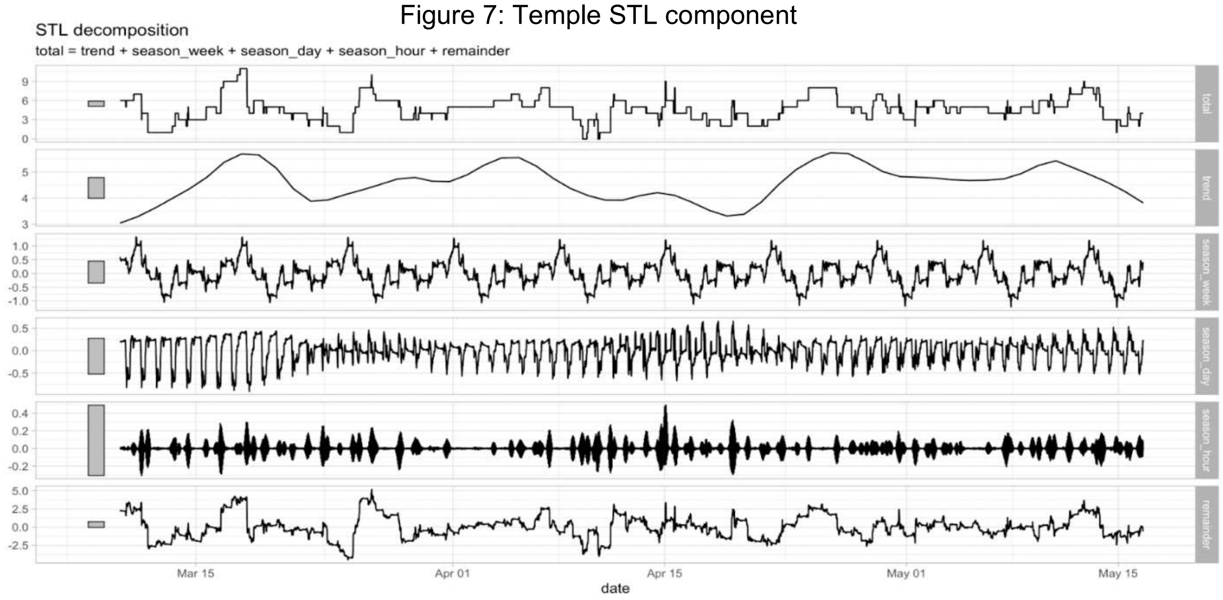


Figure 8: Medtronic STL component

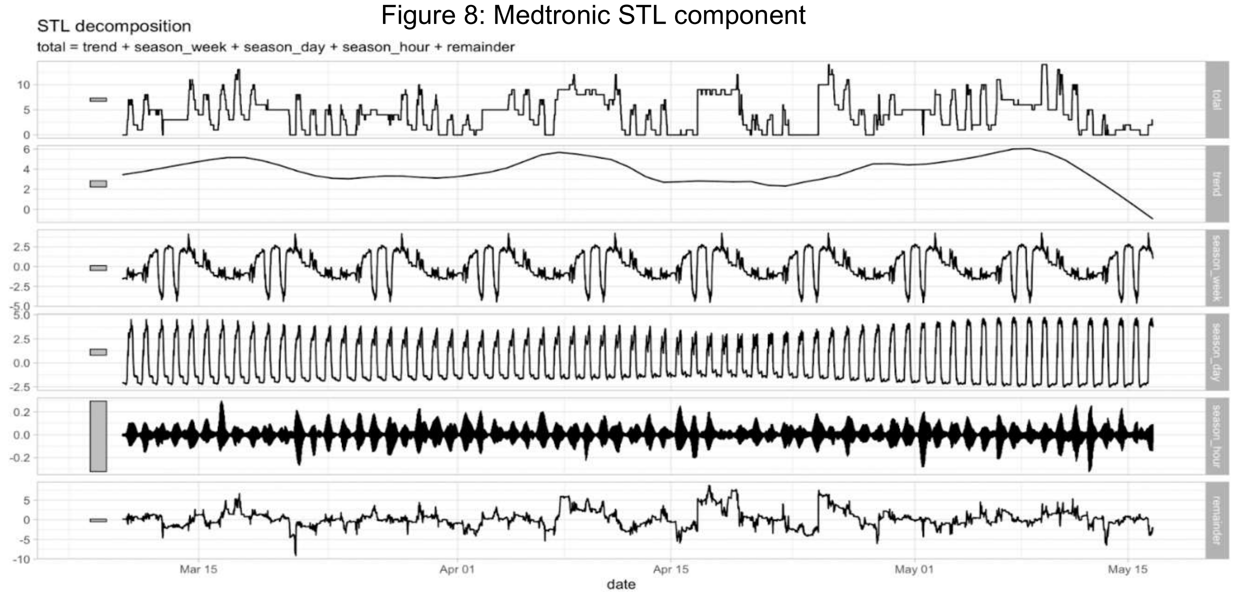


Figure 9: Moulin STL component

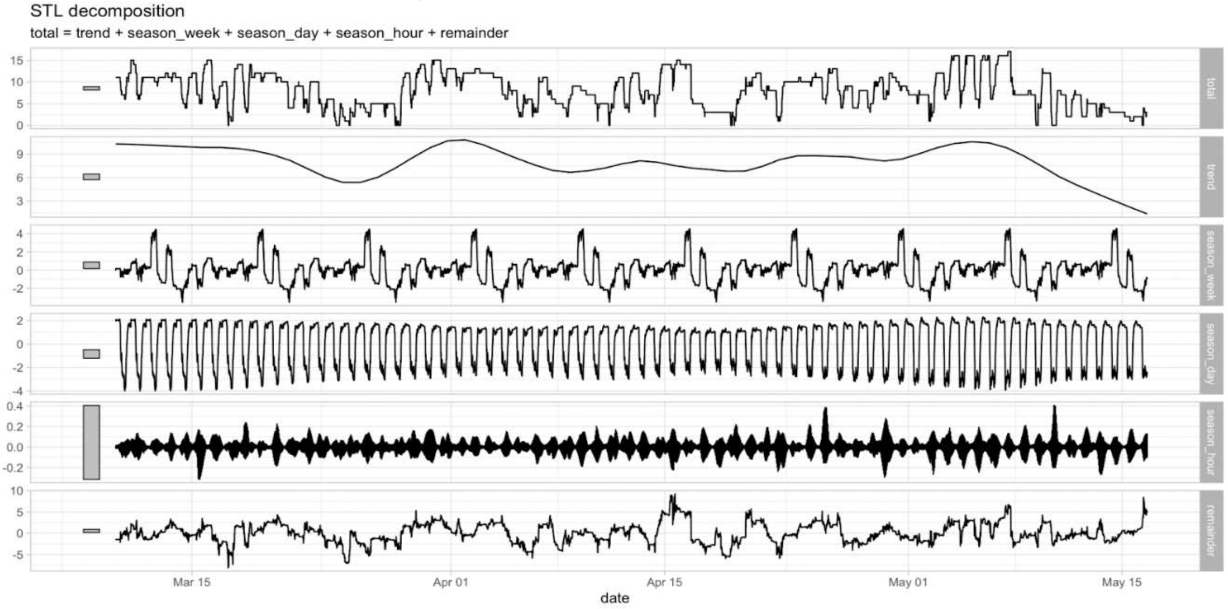


Figure 10: Gracieuse STL component

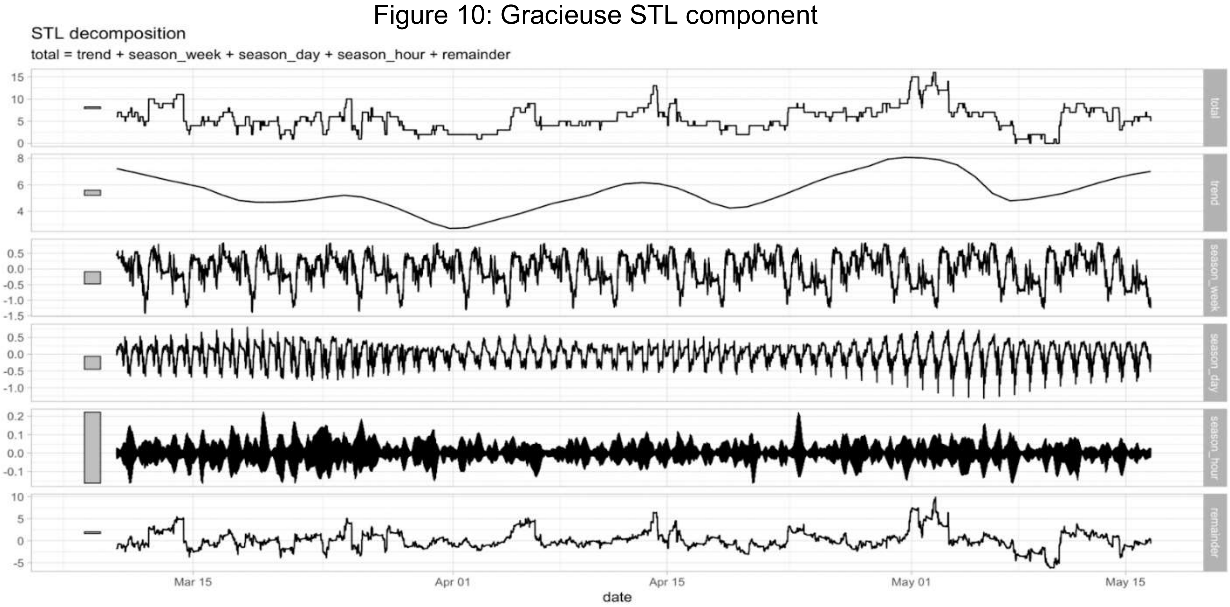


Figure 11: Préverenges STL component

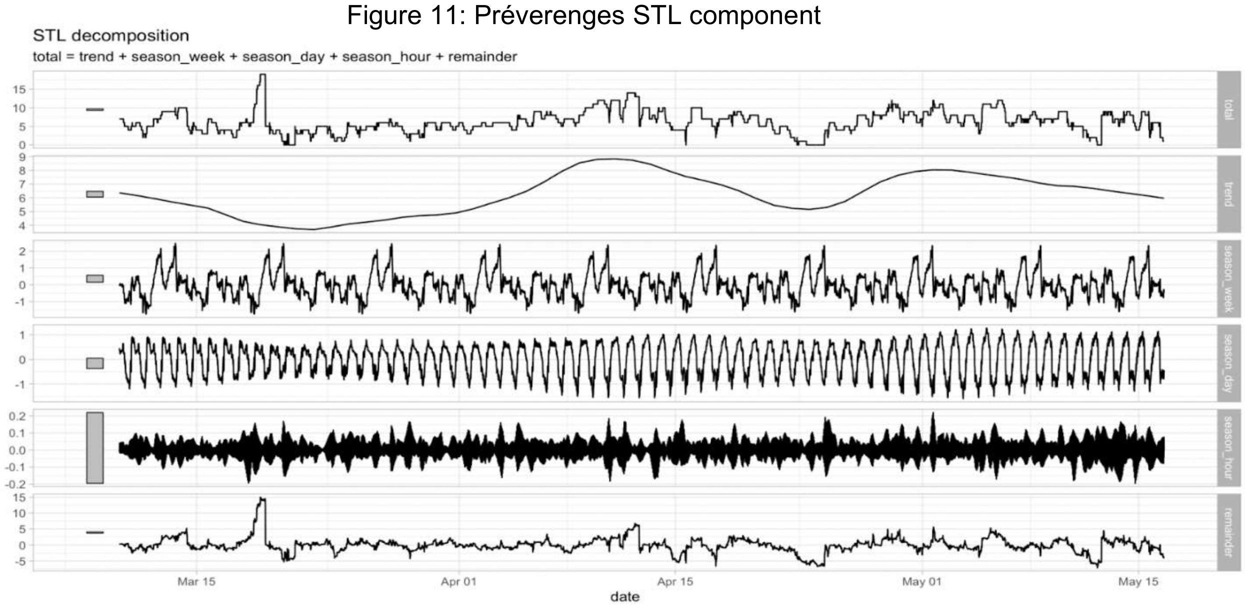


Figure 12: Sablon STL component

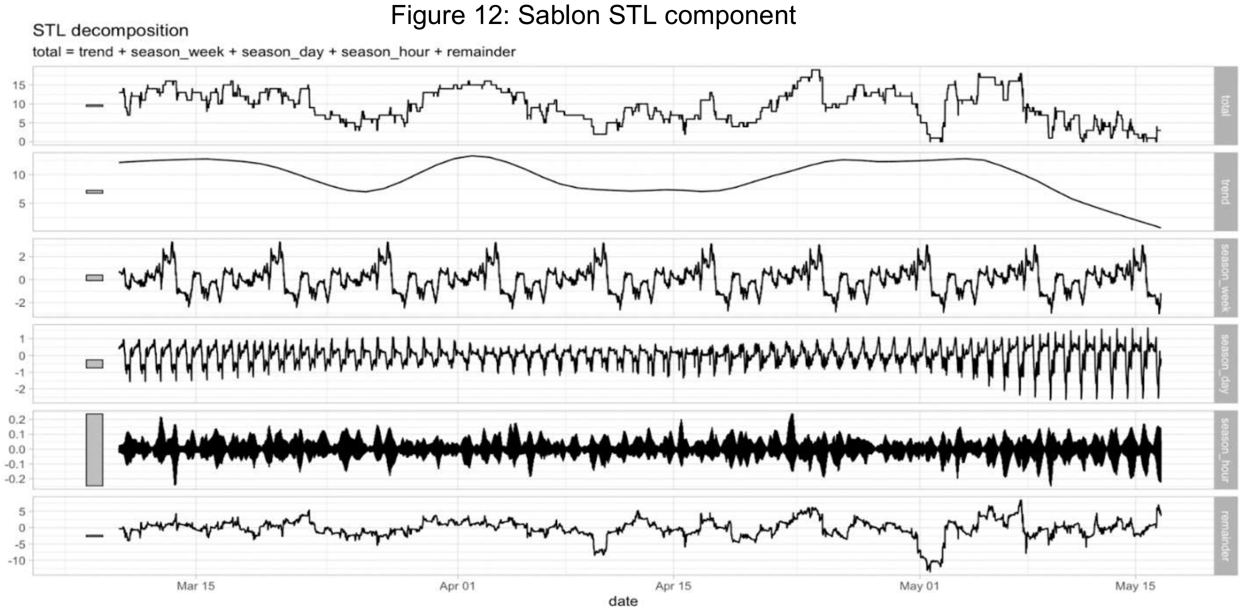
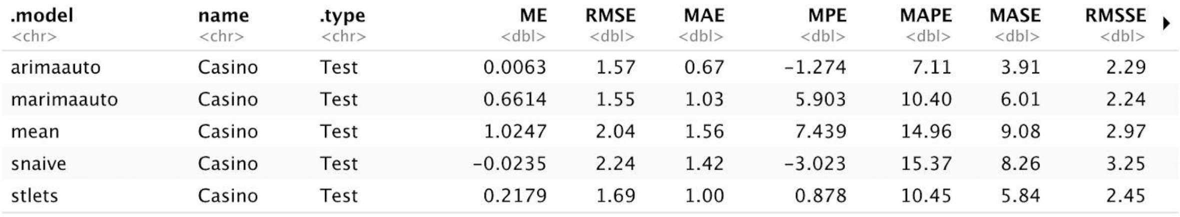


Figure 13: Casino results

Figure 13a: Casino Accuracy table (from the cross-validation)



.model <chr>	name <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>	MASE <dbl>	RMSSE <dbl>
arimaauto	Casino	Test	0.0063	1.57	0.67	-1.274	7.11	3.91	2.29
marimaauto	Casino	Test	0.6614	1.55	1.03	5.903	10.40	6.01	2.24
mean	Casino	Test	1.0247	2.04	1.56	7.439	14.96	9.08	2.97
snaive	Casino	Test	-0.0235	2.24	1.42	-3.023	15.37	8.26	3.25
stlets	Casino	Test	0.2179	1.69	1.00	0.878	10.45	5.84	2.45

Figure 13b: (ACF, Ljung-box and QQ plot) from the ARIMA with Fourier model

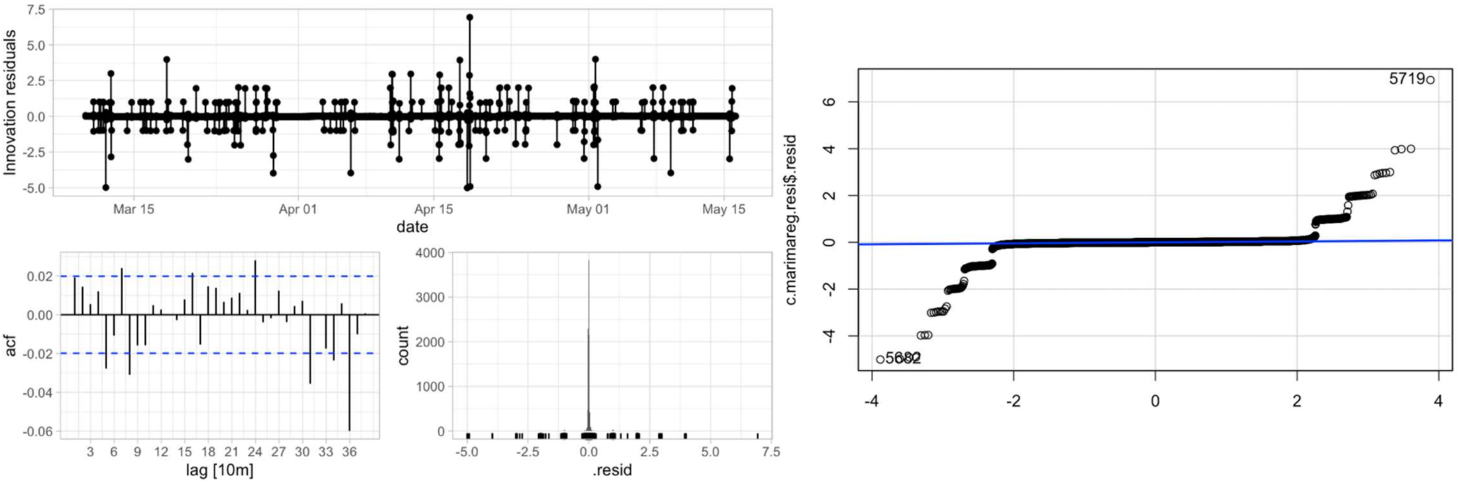
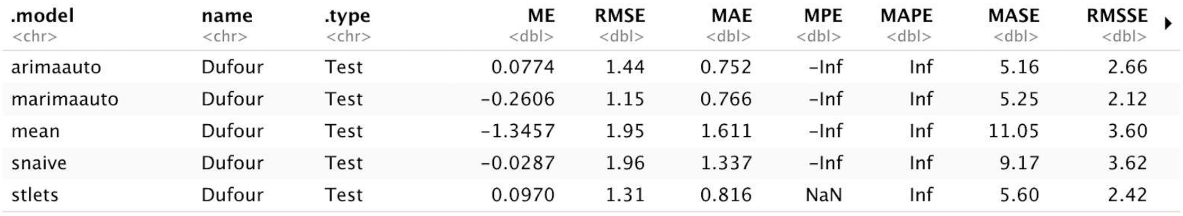


Figure 14: Dufour results

Figure 14a: Dufour Accuracy table (from the cross-validation)



.model <chr>	name <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>	MASE <dbl>	RMSSE <dbl>
arimaauto	Dufour	Test	0.0774	1.44	0.752	-Inf	Inf	5.16	2.66
marimaauto	Dufour	Test	-0.2606	1.15	0.766	-Inf	Inf	5.25	2.12
mean	Dufour	Test	-1.3457	1.95	1.611	-Inf	Inf	11.05	3.60
snaive	Dufour	Test	-0.0287	1.96	1.337	-Inf	Inf	9.17	3.62
stlets	Dufour	Test	0.0970	1.31	0.816	NaN	Inf	5.60	2.42

Figure 14b: (ACF, Ljung-box and QQ plot) from the ARIMA with Fourier model

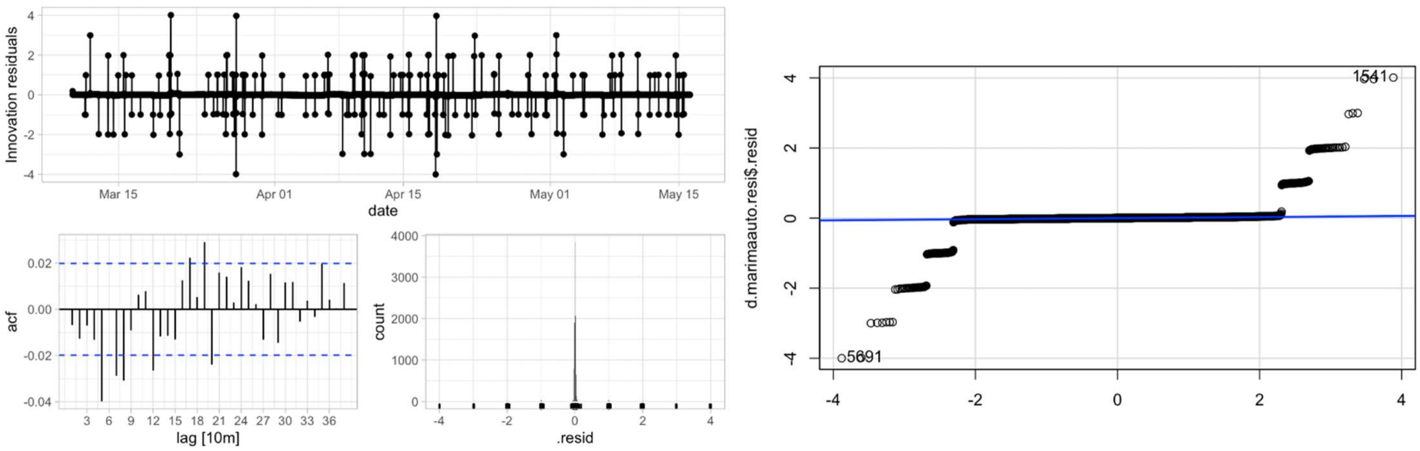
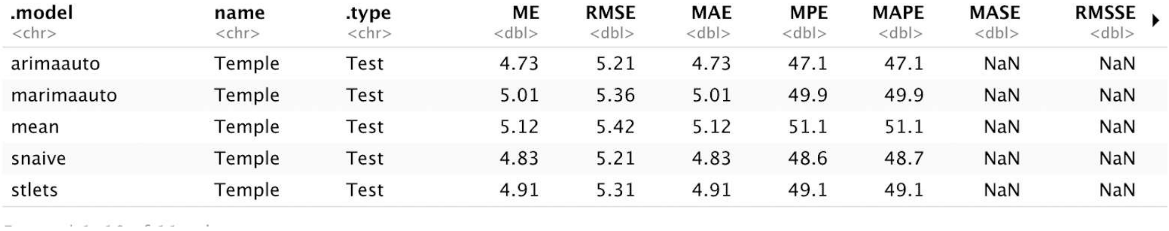


Figure 15 Temple results

Figure 15a: Temple Accuracy table (from the cross-validation)



.model <chr>	name <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>	MASE <dbl>	RMSSE <dbl>
arimaauto	Temple	Test	4.73	5.21	4.73	47.1	47.1	NaN	NaN
marimaauto	Temple	Test	5.01	5.36	5.01	49.9	49.9	NaN	NaN
mean	Temple	Test	5.12	5.42	5.12	51.1	51.1	NaN	NaN
snaive	Temple	Test	4.83	5.21	4.83	48.6	48.7	NaN	NaN
stlets	Temple	Test	4.91	5.31	4.91	49.1	49.1	NaN	NaN

Figure 15b: (ACF, Ljung-box and QQ plot) from the SNaive model

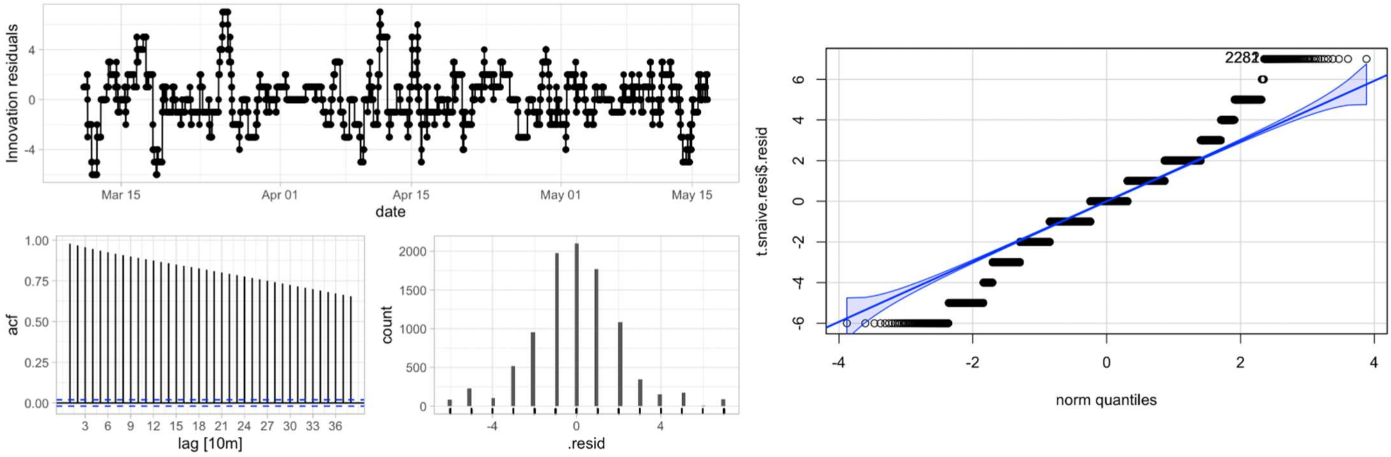


Figure 15c: (ACF, Ljung-box and QQ plot) from the auto ARIMA model without regressors

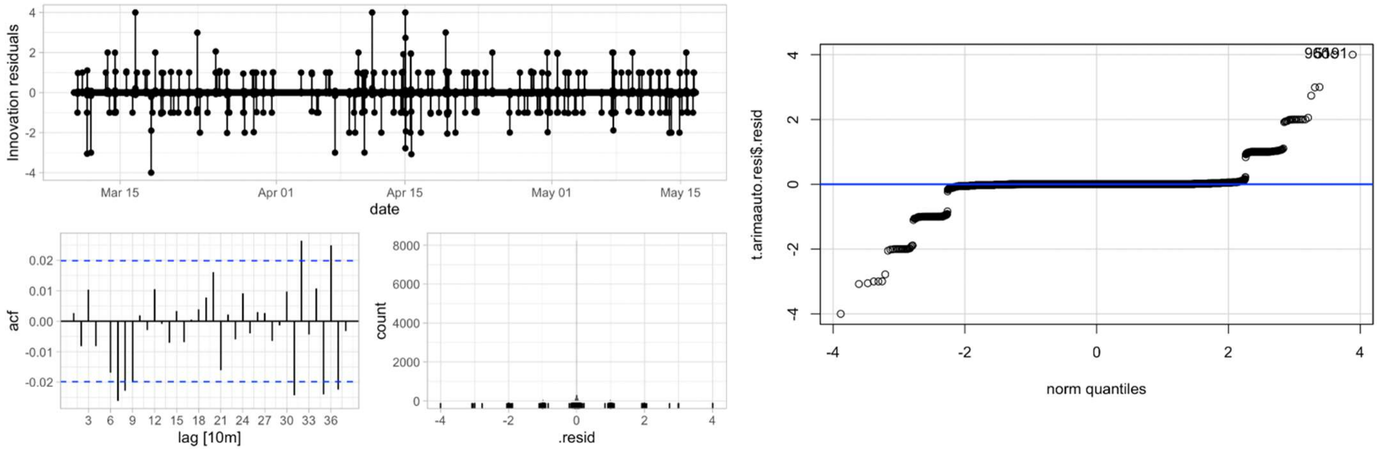


Figure 15d: (ACF, Ljung-box and QQ plot) from the auto ARIMA model with regressors

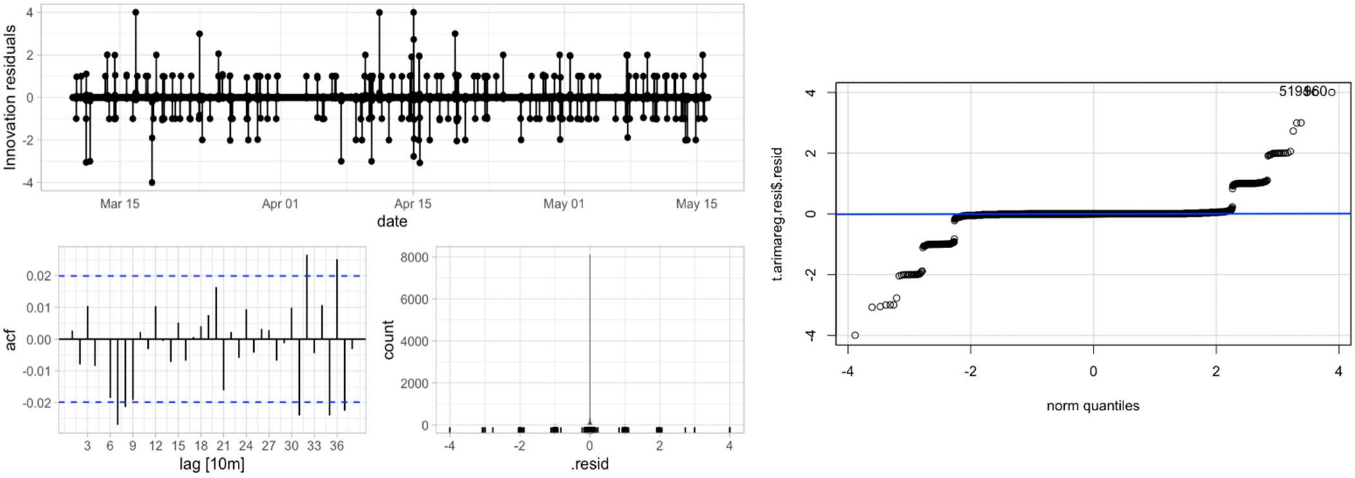
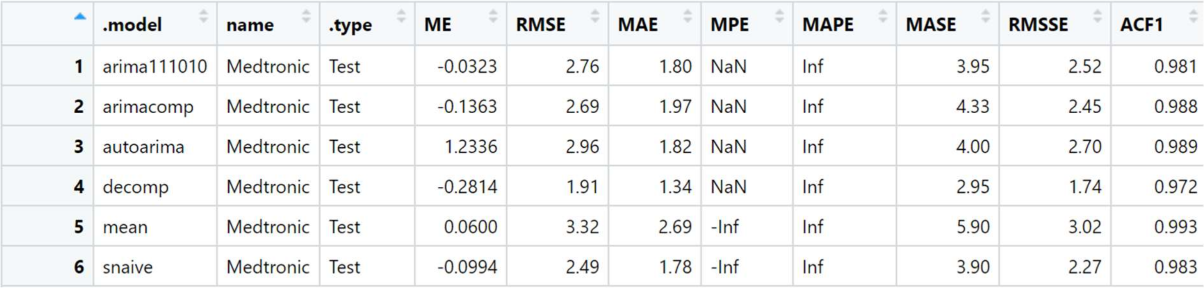


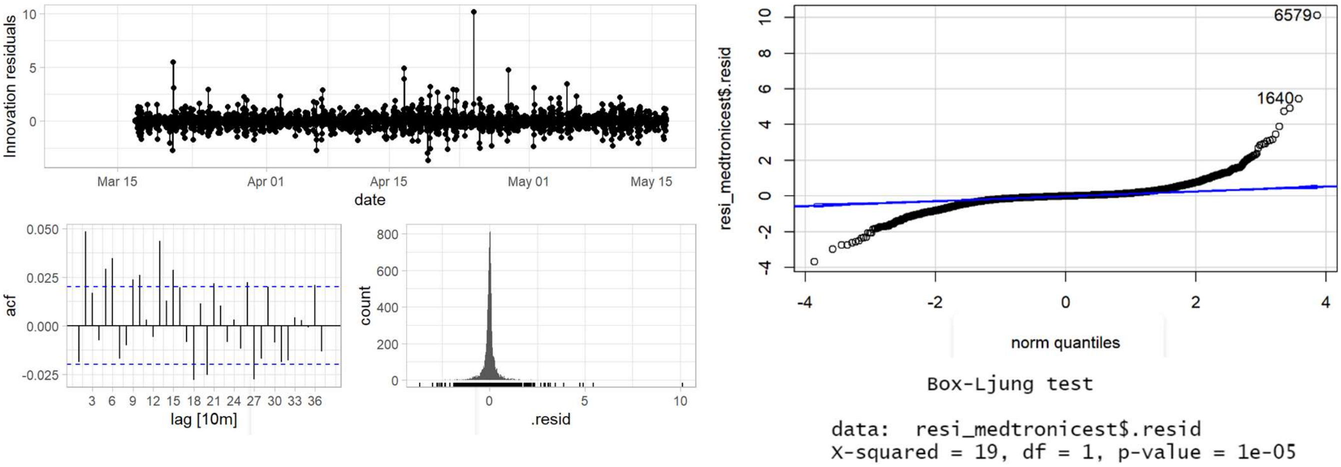
Figure 16 Medtronic results

Figure 16a: Medtronic Accuracy table (from the cross-validation)



	.model	name	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
1	arima111010	Medtronic	Test	-0.0323	2.76	1.80	NaN	Inf	3.95	2.52	0.981
2	arimacomp	Medtronic	Test	-0.1363	2.69	1.97	NaN	Inf	4.33	2.45	0.988
3	autoarima	Medtronic	Test	1.2336	2.96	1.82	NaN	Inf	4.00	2.70	0.989
4	decomp	Medtronic	Test	-0.2814	1.91	1.34	NaN	Inf	2.95	1.74	0.972
5	mean	Medtronic	Test	0.0600	3.32	2.69	-Inf	Inf	5.90	3.02	0.993
6	snaive	Medtronic	Test	-0.0994	2.49	1.78	-Inf	Inf	3.90	2.27	0.983

Figure 16b: Medtronic residuals (ACF, Ljung-box and QQ plot) from the ETS+STL model

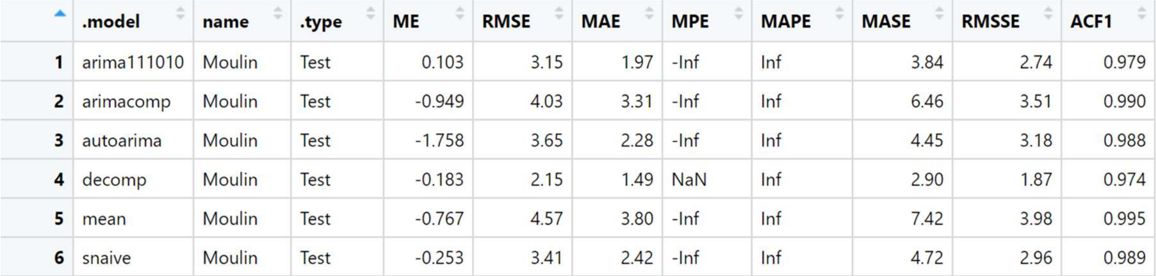


```
        Box-Ljung test

data:  resi_medtronicest$.resid
X-squared = 19, df = 1, p-value = 1e-05
```

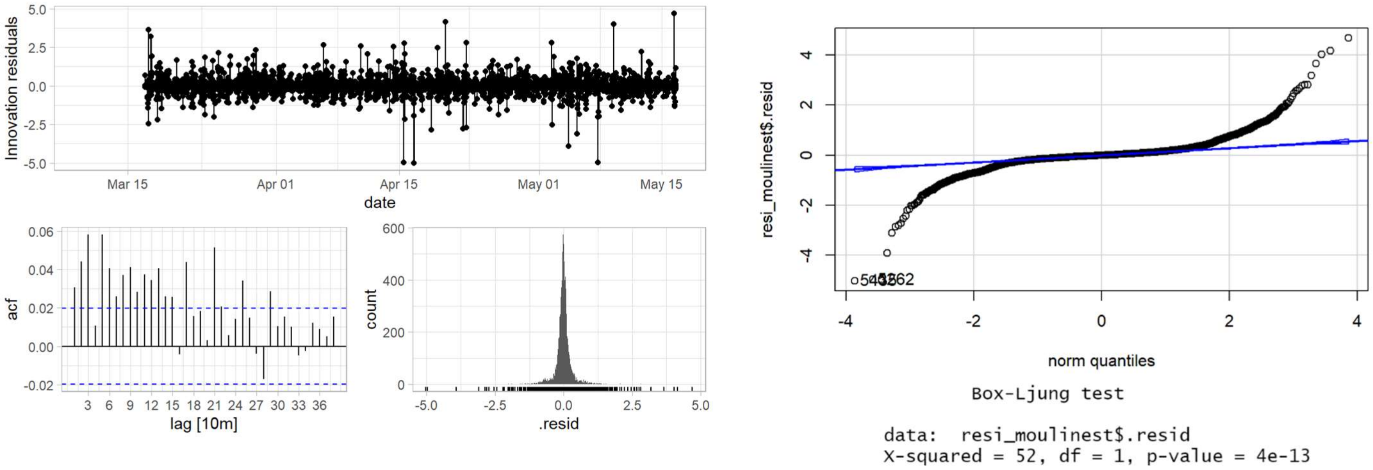
Figure 17 Moulin results

Figure 17a: Moulin Accuracy table (from the cross-validation)



	.model	name	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
1	arima111010	Moulin	Test	0.103	3.15	1.97	-Inf	Inf	3.84	2.74	0.979
2	arimacomp	Moulin	Test	-0.949	4.03	3.31	-Inf	Inf	6.46	3.51	0.990
3	autoarima	Moulin	Test	-1.758	3.65	2.28	-Inf	Inf	4.45	3.18	0.988
4	decomp	Moulin	Test	-0.183	2.15	1.49	NaN	Inf	2.90	1.87	0.974
5	mean	Moulin	Test	-0.767	4.57	3.80	-Inf	Inf	7.42	3.98	0.995
6	snaive	Moulin	Test	-0.253	3.41	2.42	-Inf	Inf	4.72	2.96	0.989

Figure 17b: Moulin residuals (ACF, Ljung-box and QQ plot) from the ETS+STL model

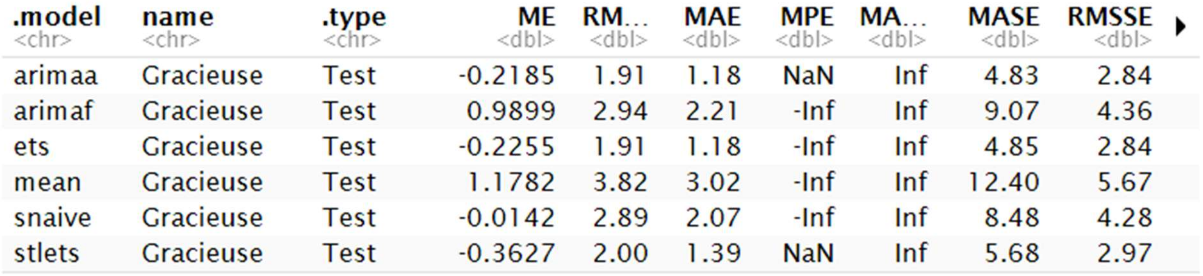


```
        Box-Ljung test

data:  resi_moulinest$.resid
X-squared = 52, df = 1, p-value = 4e-13
```

Figure 18 Gracieuse results

Figure 18a: Gracieuse Accuracy table (from the cross-validation)



.model	name	.type	ME	RM…	MAE	MPE	MA…	MASE	RMSSE
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
arimaa	Gracieuse	Test	-0.2185	1.91	1.18	NaN	Inf	4.83	2.84
arimaf	Gracieuse	Test	0.9899	2.94	2.21	-Inf	Inf	9.07	4.36
ets	Gracieuse	Test	-0.2255	1.91	1.18	-Inf	Inf	4.85	2.84
mean	Gracieuse	Test	1.1782	3.82	3.02	-Inf	Inf	12.40	5.67
snaive	Gracieuse	Test	-0.0142	2.89	2.07	-Inf	Inf	8.48	4.28
stlets	Gracieuse	Test	-0.3627	2.00	1.39	NaN	Inf	5.68	2.97

Figure 18b: Gracieuse residuals (ACF) from the auto ARIMA model with regressors

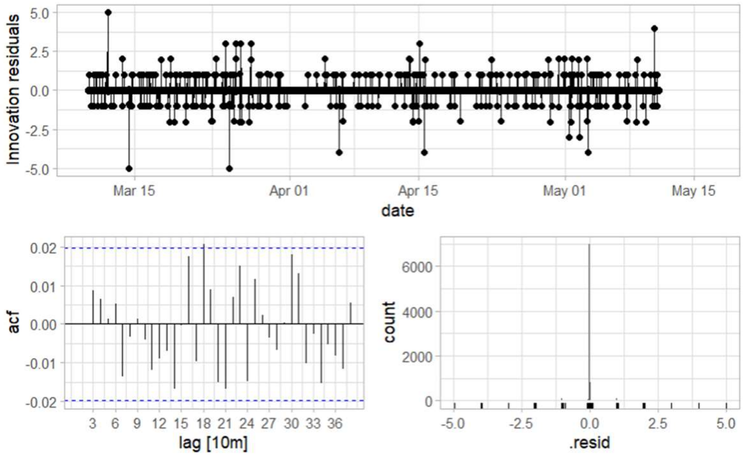
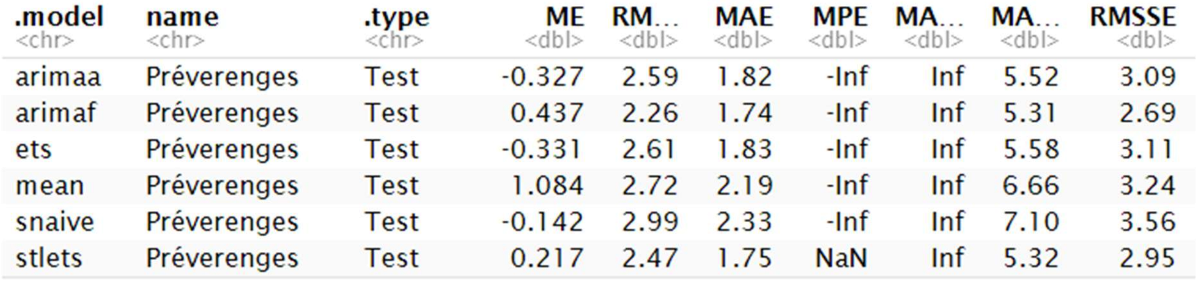


Figure 19 Préverenges results

Figure 19a: Préverenges Accuracy table (from the cross-validation)



.model	name	.type	ME	RM…	MAE	MPE	MA…	MA…	RMSSE
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
arimaa	Préverenges	Test	-0.327	2.59	1.82	-Inf	Inf	5.52	3.09
arimaf	Préverenges	Test	0.437	2.26	1.74	-Inf	Inf	5.31	2.69
ets	Préverenges	Test	-0.331	2.61	1.83	-Inf	Inf	5.58	3.11
mean	Préverenges	Test	1.084	2.72	2.19	-Inf	Inf	6.66	3.24
snaive	Préverenges	Test	-0.142	2.99	2.33	-Inf	Inf	7.10	3.56
stlets	Préverenges	Test	0.217	2.47	1.75	NaN	Inf	5.32	2.95

Figure 19b: Préverenges residuals (ACF) from the ARIMA with the Fourier model

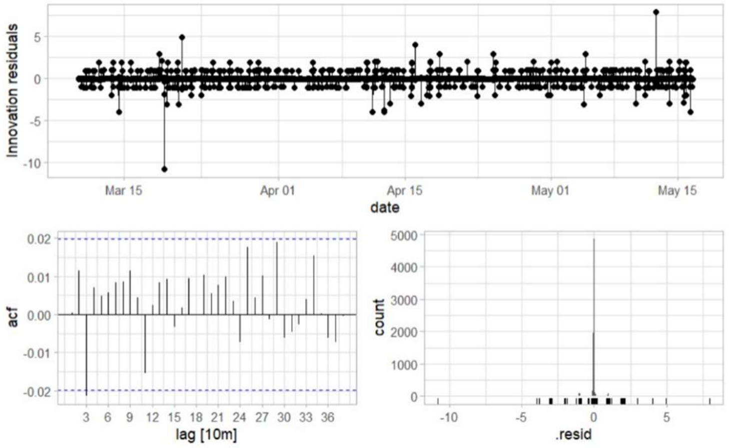
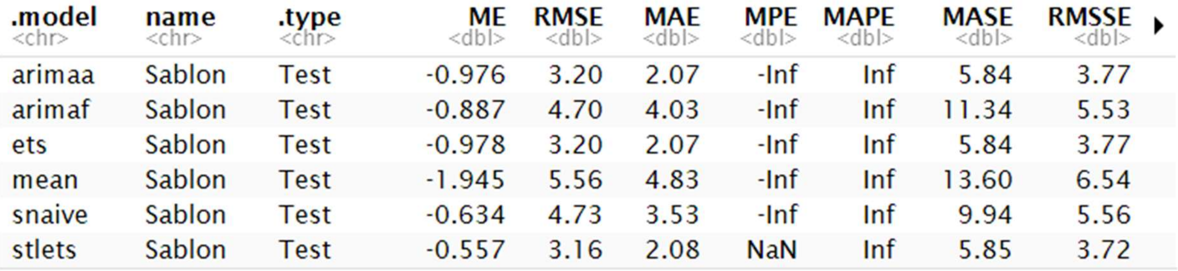


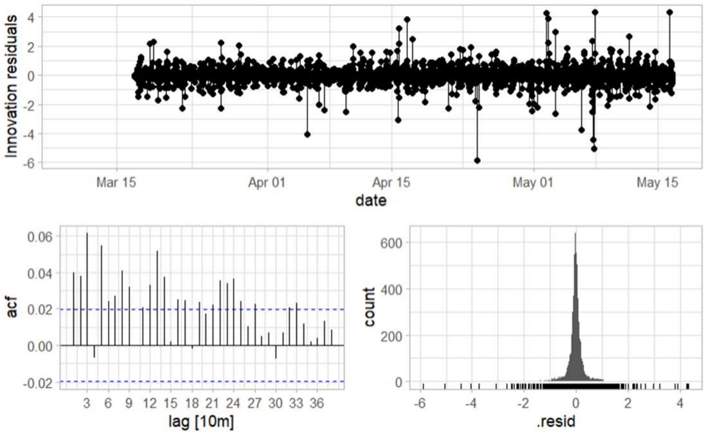
Figure 20 Sablon results

Figure 20a: Sablon Accuracy table (from the cross-validation)



.model	name	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
arimaa	Sablon	Test	-0.976	3.20	2.07	-Inf	Inf	5.84	3.77
arimaf	Sablon	Test	-0.887	4.70	4.03	-Inf	Inf	11.34	5.53
ets	Sablon	Test	-0.978	3.20	2.07	-Inf	Inf	5.84	3.77
mean	Sablon	Test	-1.945	5.56	4.83	-Inf	Inf	13.60	6.54
snaive	Sablon	Test	-0.634	4.73	3.53	-Inf	Inf	9.94	5.56
stlets	Sablon	Test	-0.557	3.16	2.08	NaN	Inf	5.85	3.72

Figure 20b: Sablon residuals (ACF) from the ETS+STL model (left)
Figure 20c: Sablon residuals (ACF) from the auto ARIMA with regression model (right)



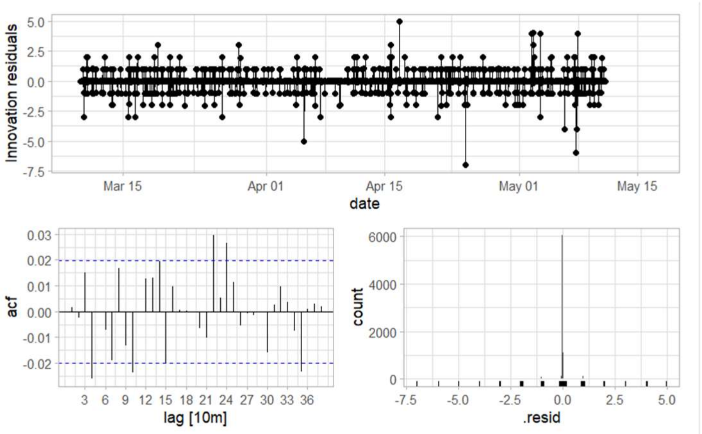


Figure 20d: Sablon residuals (ACF) from the ETS model

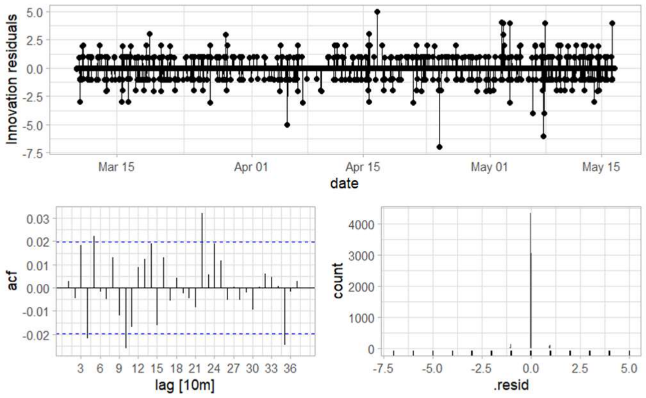
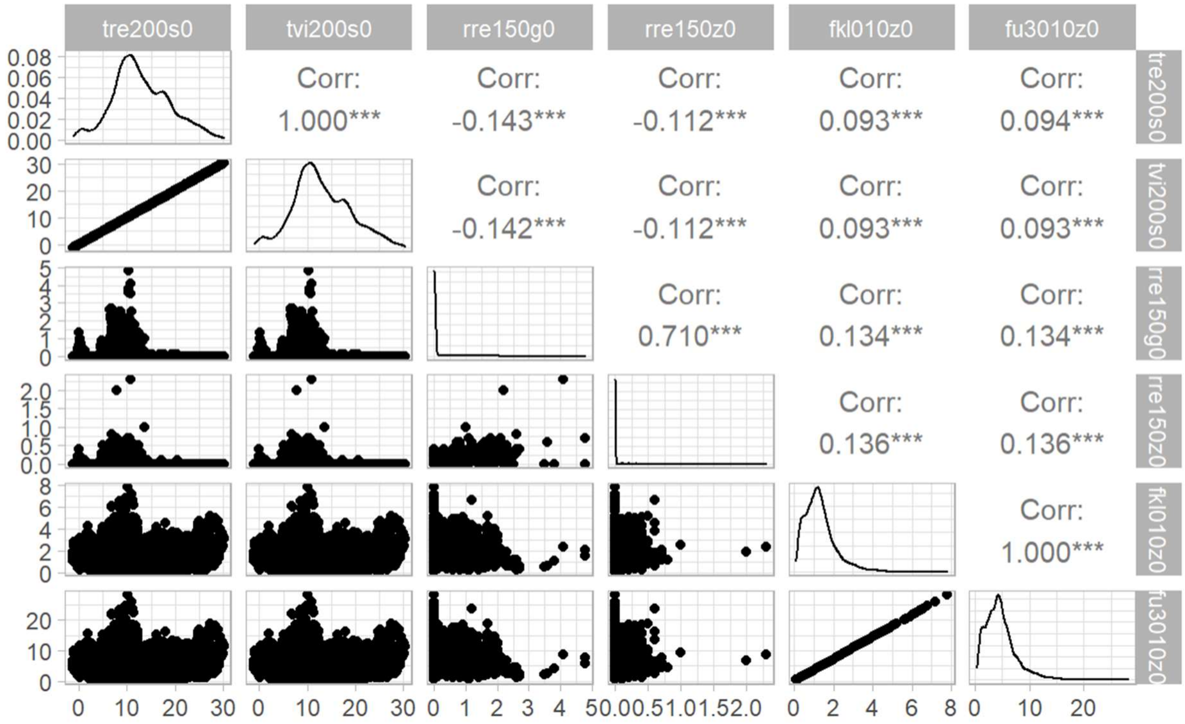


Figure 21 Weather correlation plot


tre200s0
tvi200s0
rre150g0
rre150z0
fkl010z0
fu3010z0
Corr:
1.000***
Corr:
-0.143***
Corr:
-0.112***
Corr:
0.093***
Corr:
0.094***
Corr:
-0.142***
Corr:
-0.112***
Corr:
0.093***
Corr:
0.093***
Corr:
0.710***
Corr:
0.134***
Corr:
0.134***
Corr:
0.136***
Corr:
0.136***
Corr:
1.000***