Time Series project

Patrick Guerin - NOMA: 80541700

2018-05-21

Contents

Introduction **1**

Data exploration **2**

Box-Jenkins analysis **3**
Transforming the data toward stationarity 5
Adressing trend and seasonality with lagged differences 5

Seasonal-ARIMA process **8**
Testing normality and significance 8

Model evaluation **10**

Comparison with ARMA model **10**

Prediction : Comparing SARIMA with Holt-Winters model **10**

Random Forest for time series **12**
Brief description of the model 12
Application for time series 12

Conclusion **15**

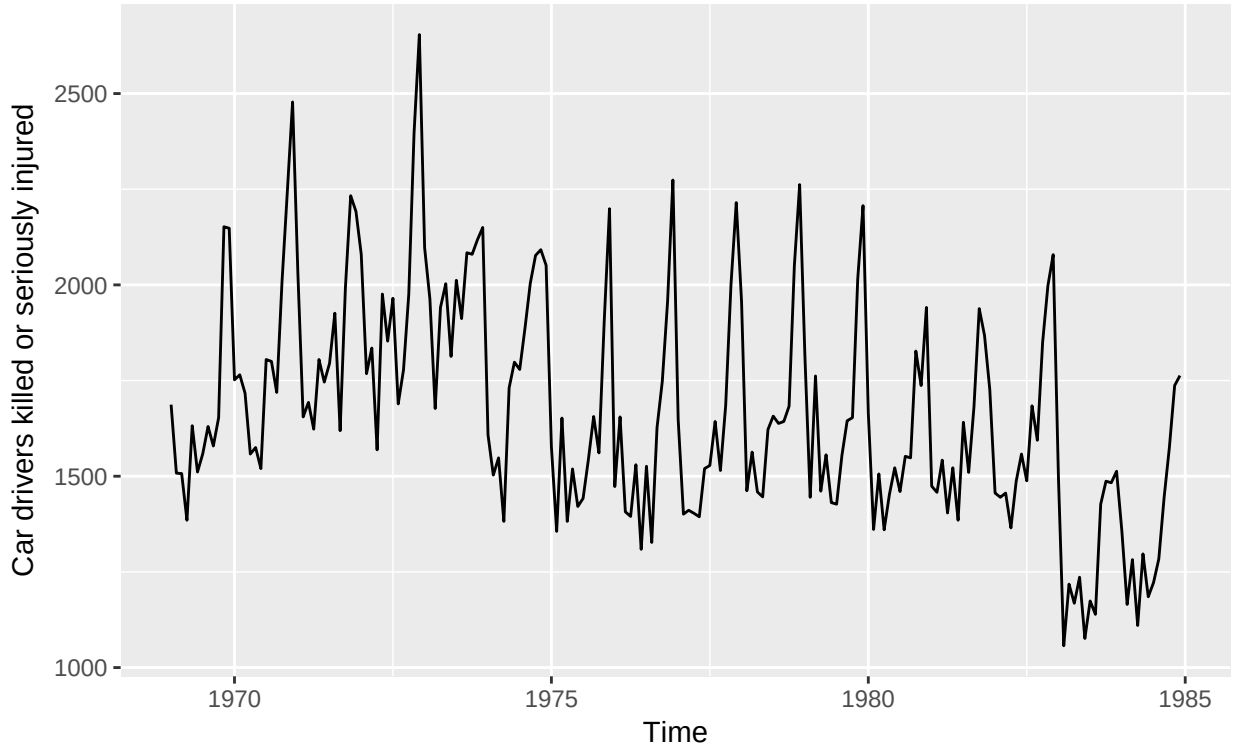
Annexe **15**

Introduction

In this project we study a time series representing the **monthly number of Car drivers killed or seriously injured** in Great Britain, from January 1969 to December 1984. We will explore the data in order to understand it,then choose and fit an appropriate model to forecast the monthly numbers of serious accidents for the year 1985. Finally we will assess the quality of our model.

Data exploration

Firstly, to get an intuition about the data, we simply plot it.



We can observe that our data exhibits obvious seasonality. To dig into it let's examine the average number of accidents associated with each month.

Table 1: average number of car drivers killed or seriously injured for each month of the year

January	February	March	April	May	June	July	August	September	October	November	December
1698	1498	1548	1435	1573	1517	1592	1607	1660	1800	1999	2116

We can see that there is far much accidents during the autumn and winter seasons, until the month of January where the number of accident drops to 1698 in average.We can identify a seasonal component with a period of 12 months.

In the precedent plot trend we also notice a trend: the number of accidents per year seems to decrease in time. we can investigate this by computing the average number of accident per year.

Table 2: average number of car drivers killed or seriously injured per year

1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
1663	1828	1859	1962	1988	1788	1601	1602	1614	1703	1664	1578	1596	1622	1289	1368

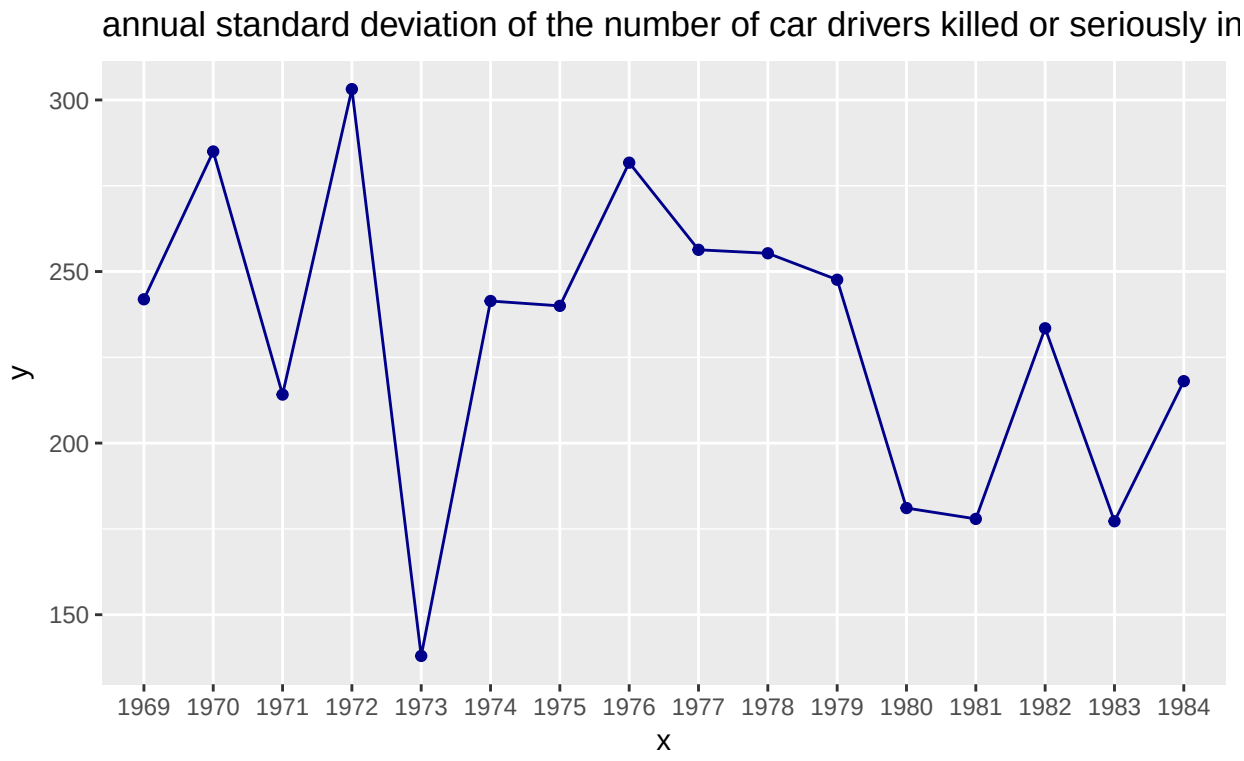
We notice that after a first rise from 1969 to 1973, the number of serious car accidents declined from 1973 to 1977, in 1978 there was a spike, and after that,the number of accident tended to decline. In 1983 It appears than some changes occured since the number of accidents declined abrutly.

Therefore, we will need to adress both issues of trend and seasonality.

Moreover, it can be interesting to see if our process exhibits a variance constant in time or if it varies as time goes.

Table 3: standard deviation of car drivers killed or seriously injured

1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
242	285	214	303	138	241	240	282	256	255	248	181	178	233	177	218



We can see that the variance of our process clearly depends of the time.

Box-Jenkins analysis

In this part,we will follow the Box-Jenkins method:

1-Identifying trend and seasonality in the time serie

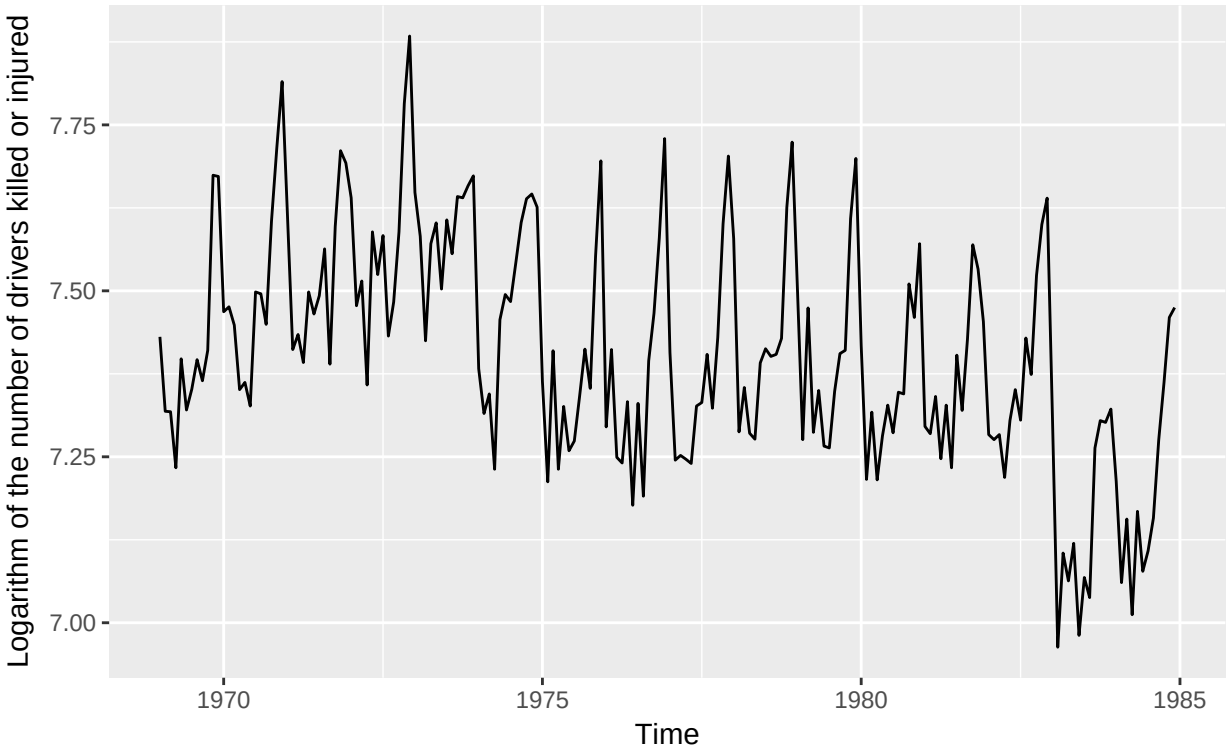
2-Estimation of the parameter of the model by Likelihood maximisation

3-Model assessment with (p)acf plots and Ljung-Box test

Transforming the data toward stationarity

Evaluating if the serie if stationary is crucial since Box-Jenkins models rely on the hypothesis of a weakly stationary process. We have seen with the preceding exploration that both the mean and the variance are not constant over time, hence we have to make the data stationary.

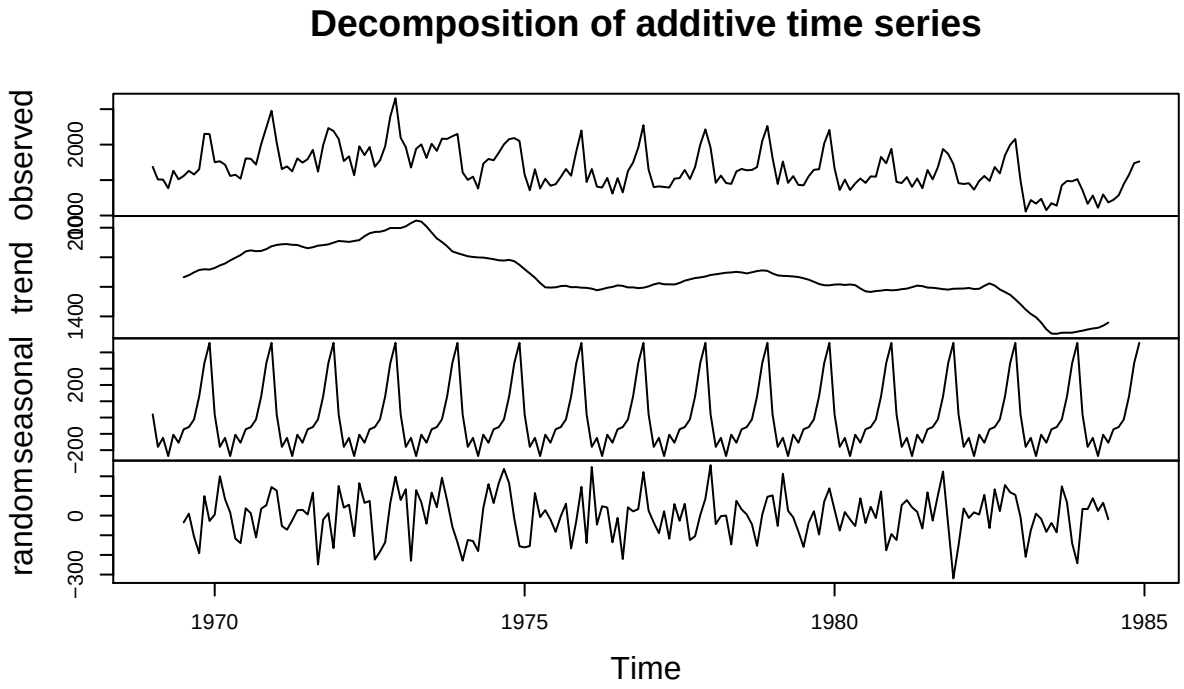
Firstly,we stabilize its variance with a logarithm transformation to reduce the largest fluctuations.



We can observe that if the variance has been reduced, the data is not stationary yet, and we need to remove its trend and seasonality.

Adressing trend and seasonality with lagged differences

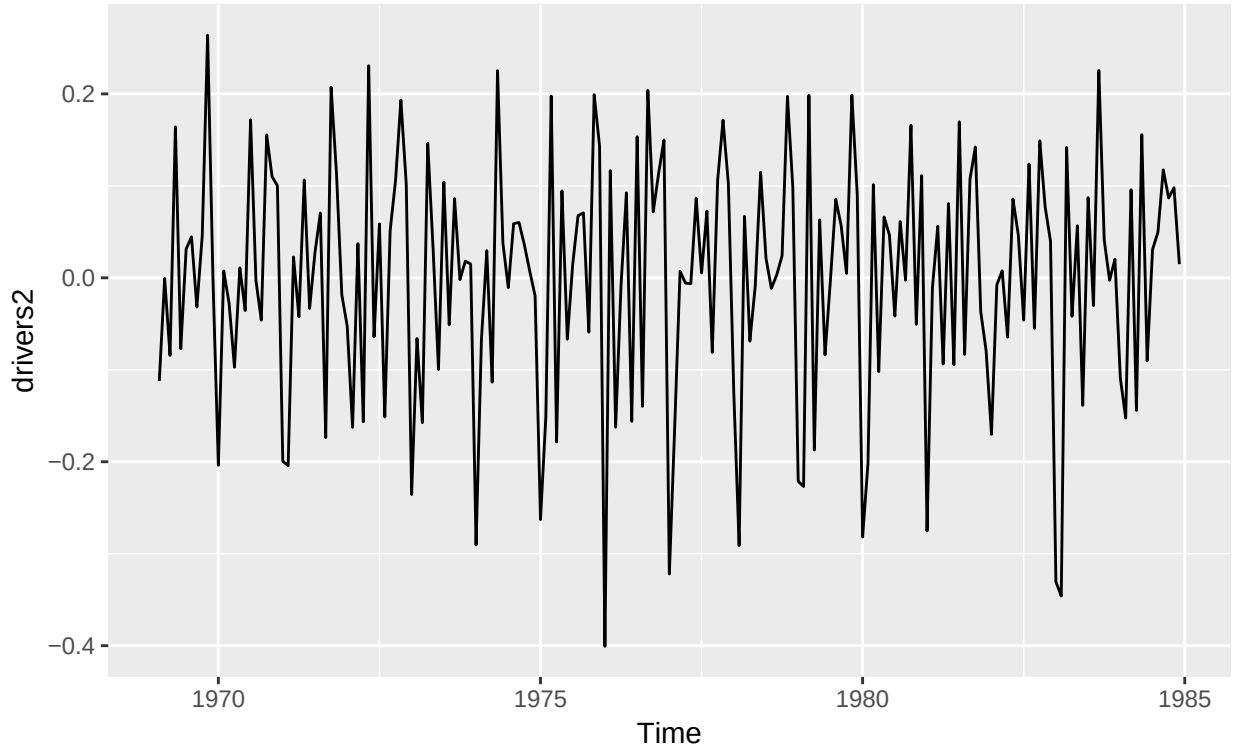
We can decompose the time serie to separate its trend, seasonality and random component with the function **decompose** which use moving averages to estimate each component of the additive model $Y_t = T_t + S_t + e_t$



Detrending

We can remove the trend by differencing the time serie to stabilize the mean:

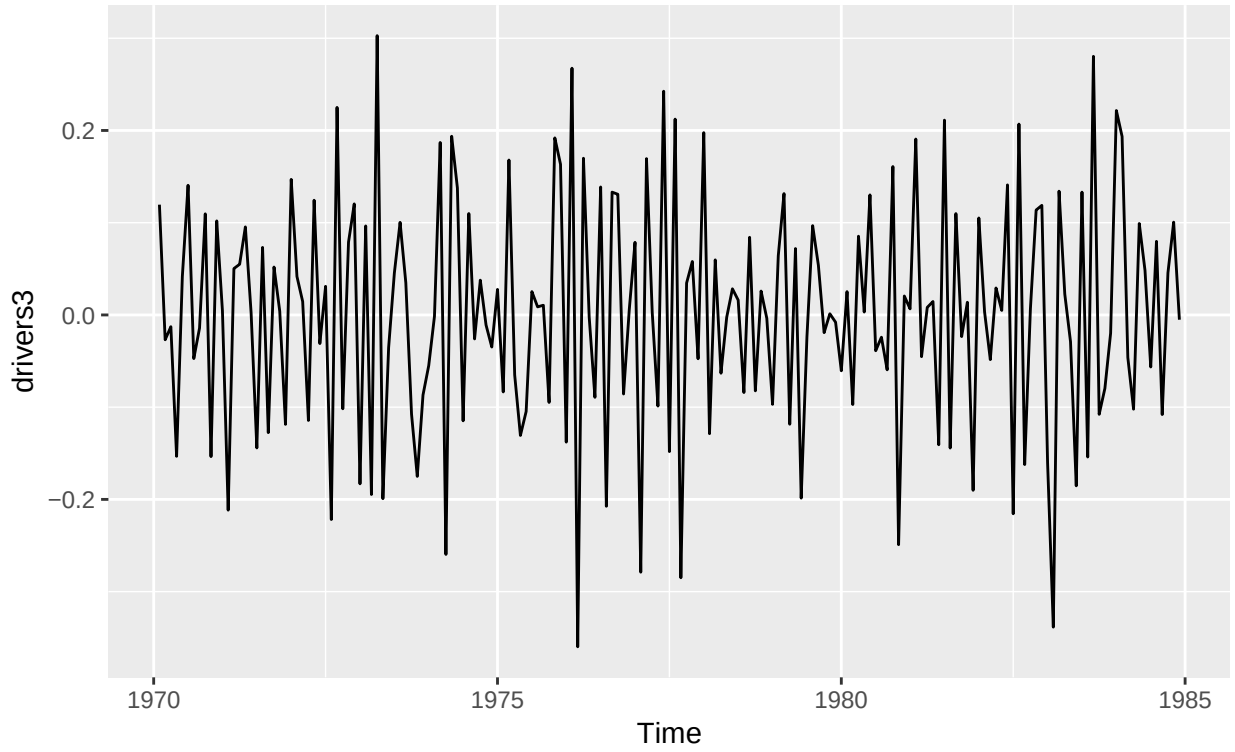
The transformation is $\Delta_1 X_t$.



After this treatment,the mean now seems to be constant.

Deseasonalizing

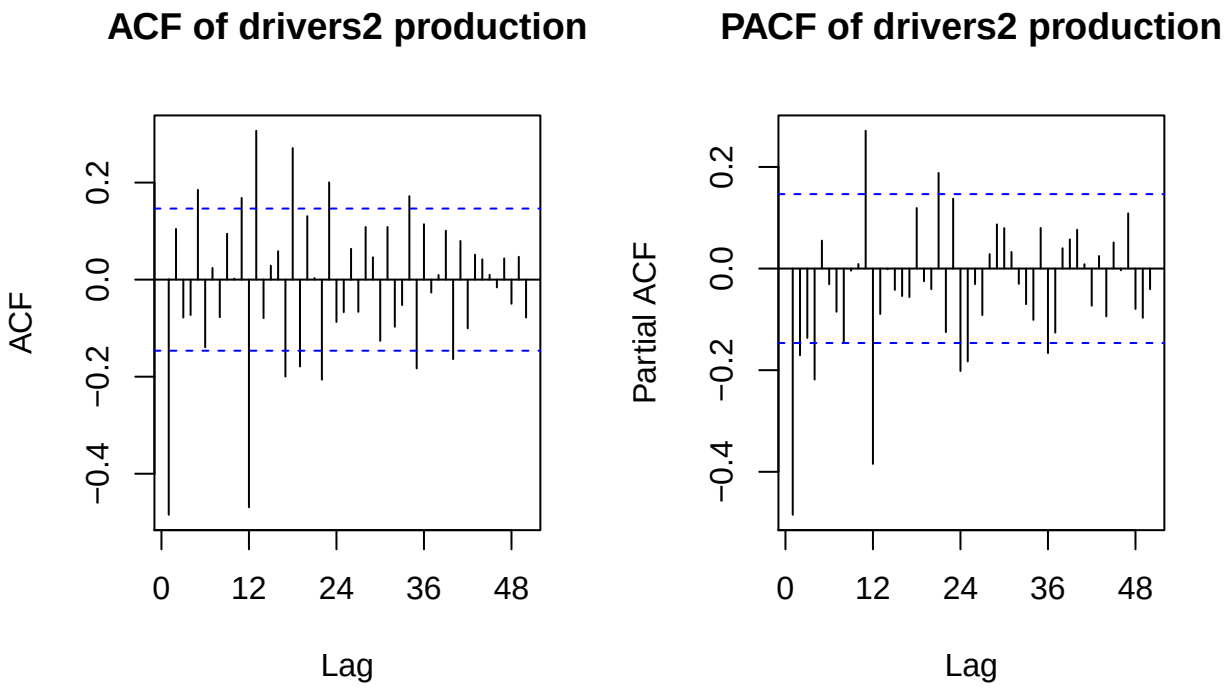
To remove the effect of seasonality, we can take the lagged difference of order 12. The transformation is $\Delta_{12} X_t$



We can see the effect of removing this periodic component.

We are going to assume the stationarity of this serie to fit a Seasonal ARIMA model

In order to know if there is any correlation left we plot one more time the autocorrelation and partial autocorrelation functions:



We can see that the data is still correlated after those corrections, this can suggest that a stochastic seasonal component may remain: a Seasonal ARIMA model is adapted to adress this.

From the Autocorrelation and partial autocorrelation function, we can identify an autoregressive part of order 1 and a moving average part of order 1, respectively corresponding to the first spike in ACF and PACF. We can identify a seasonal autogressive process of order 1 (other seasonal peaks are barely significant), and a seasonal moving average process of order 1 too.

Seasonal-ARIMA process

To fit our SARIMA model, we use the Akaike criterion information to compare several SARIMA(p,1,q)(P,1,Q)12 with $p, q, P, Q \in [0, 1]$:

The best model selected by AIC is the SARIMA(1,1,1)x(0,1,1)12, which can be written

$(1 - \Phi B)(1 - B)(1 - B_{12})Y_t = (1 + \theta B)(1 + \Theta B^{12})\epsilon_t$

Next,we estimate the parameters of the model by maximising the likelihood of the data conditionally to the parameters:

Testing normality and significance

Normality

We can test the significativity of the model coefficients, based on the assumption that the data is normally distributed, we use the Shapiro-wilk test to verify this hypothesis.

p-value
0.3539371

For an alpha level of 5% we cannot reject the hypothesis of normality.We can use this fact to test the significance of the SARIMA coefficients.

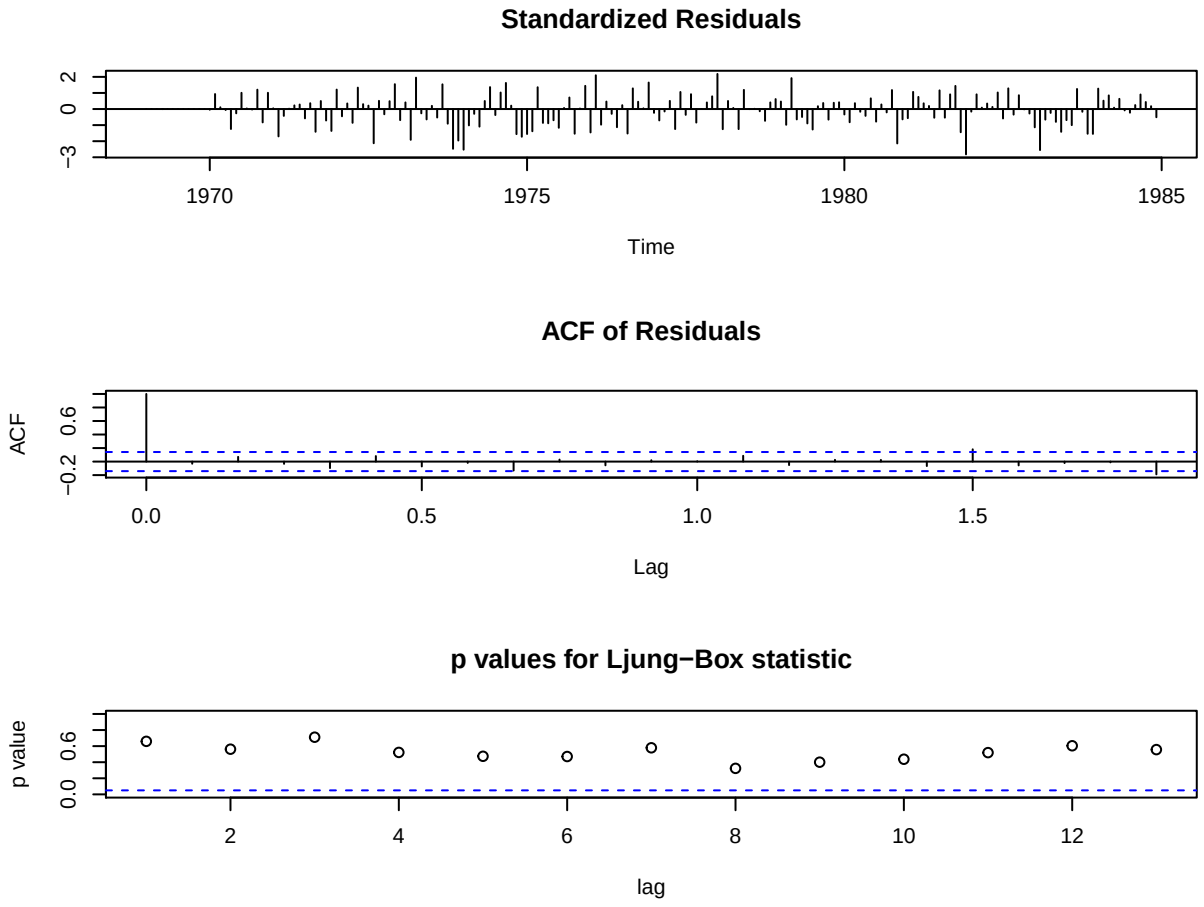
$$(\text{H0} : \text{coeff}_i = 0, \text{H1} : \text{coeff}_i \neq 0.)$$

Statistical significance

ar1	ma1	sma1
0.0298	4.046e-24	4.691e-16

At a risk of 5% we can reject H0,all the coefficients are significantly different from zero.

To see if there is any correlation left in the residuals, we can look at the autocorrelation function of the residuals.More formally,we can also use the Ljung-Box test to test the independance of the residuals.



All the p-values are far higher than the 5% treshold and we cannot reject the hypothesis of independance.

Moreover, the ACF plot shows that there is no autocorrelation in the residuals: the presence of autocorrelation in the residuals woud have suggested that there was information that had not been accounted for in the model.

Model evaluation

One way to assess the predictive power of our model is to fit it on 80% of the data, make a one-ahead prediction, compute the error $\epsilon_t = \hat{X}_{t+1} - X_{t+1}$, and recursively fit the model with the new information obtained each time. At the end we obtains the root mean square error of all of one-ahead predictions: $\sqrt{\sum_{i=1}^{N}(\hat{X}_{t+i} - X_{t+i})^2}$.

RMSE with one step-ahead prediction
140.31

Comparison with ARMA model

It can be interesting to compare the performance of the SARIMA with ARMA and ARIMA models to highlight the necessity of having a seasonal model . After trials and error we chose to use BIC to do the model selection since AIC tended to favour less parcimonious (and not more accurate) models.

```
## Best model according to Bayesian information criterion : ARMA( 1 , 1 )
## BIC: 10.68152
## Best model according to Bayesian information criterion : ARIMA( 1 , 1 , 1 )
## BIC: 10.65587
```

The best ARMA model selected has the form : $\Phi(B)Y_t = c + \theta(B)\epsilon_t \iff Y_t = c + \Phi_1 Y_{t-1} + \epsilon_t + \theta_1\epsilon_{t-1}$

The best ARIMA model selected has the form: $y_t = c + \Phi_1 y_{t-1} + \epsilon_t + \theta_1\epsilon_{t-1}$,

with $y_t = Y_t - Y_{t-1}$

Table 7: RMSE with one step-ahead prediction

ARMA	ARIMA	SARIMA
190	181.27	140.31

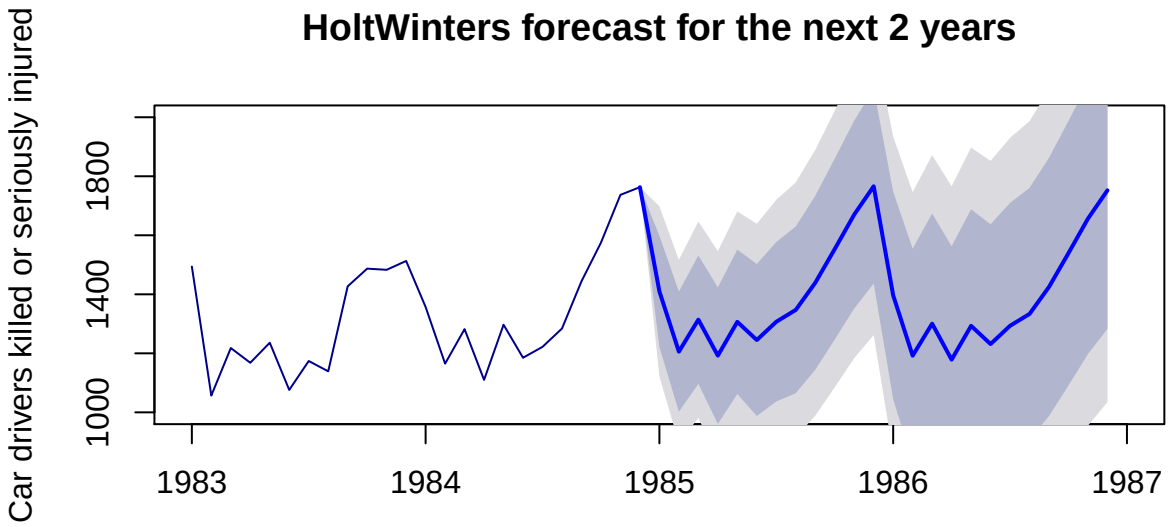
Without surprise, the SARIMA model have a much better performance than the ARMA and ARIMA models since the data is strongly seasonal.

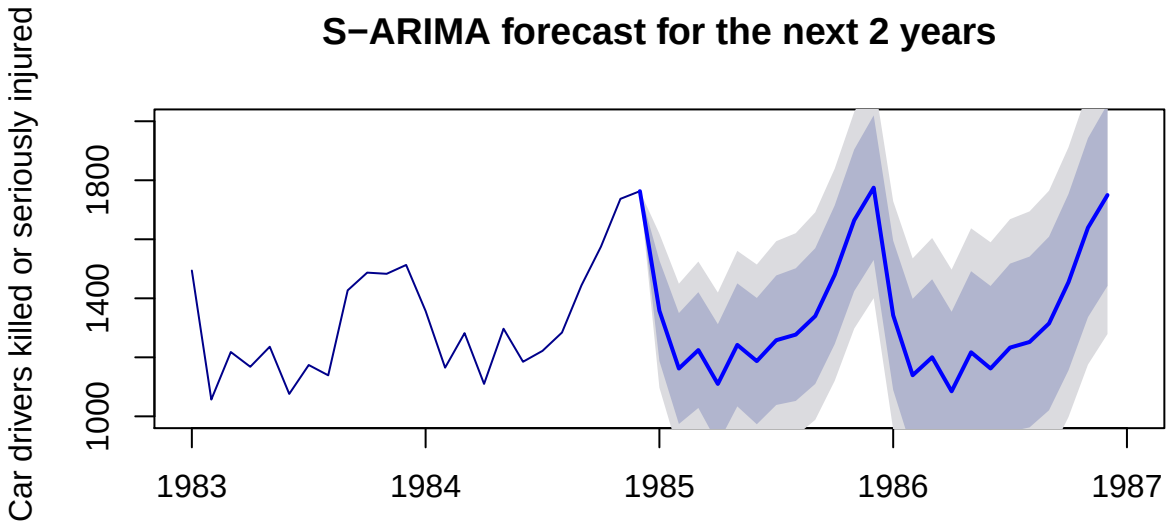
Prediction : Comparing SARIMA with Holt-Winters model

The Holt-Winters modem also have a seasonal component and can be an interesting alternative to the SARIMA model.

Below we Compute the one-ahead prediction error of the Holt-Winter algorithm. After that, we plot the predictions of the two models for the next two years, with 80% and 95% confidence intervals.

Holt-Winter one step-ahead prediction error
144.57





We can see that the forecast obtained with the Holt-Winter algorithm is smoother than the Seasonal ARIMA forecast while the confidence intervals are wider for the Holt-Winter method. Moreover, the one step-ahead prediction error is sligthly worse.

This results can be explained by the fact that the Holt-Winter model is a more general model than the S-ARIMA, consequently it has also less risk of overfitting and could be prefered in case of long term prediction.

Random Forest for time series

In this part, we tried to use the Random Forest model to assess its capacity to compete with the traditional approach of time series.

Brief description of the model

In one hand Decision trees suffers from low accuracy when they are not grown deep enough, in the other hand if they are deep they tend to overfit the dataset. Hence, the bootstraping procedure is particularly adapted to the decisions trees. This is the idea of the random forest model.

Each decision tree outputs a mean prediction and the random forest model is simply an average of an ensemble of decision trees, using bootstrap to decorelate the tree and further reduce the variance of the final prediction.

The Boostrap aggregation is used in two ways for each tree:

1. We sample n observations from the training set
2. We sample m features from the feature space to be used in the trees.

The second sampling is necessary because if one or a few features are very strong predictors for the response variable, these features will be selected in many of the trees, causing them to become correlated.

Application for time series

The random forest algorithm is not natively made to deal with temporal data and we had to do create new variables in order to obtain good results.

First we created a dataframe, with each row representing one observation containing the monthly number of severe accidents and several other features that we had to create:

1. Creation of two features from the date: **year** and **month**, month was encoded as a categorical variable ranging from 1 to 12, year could not be encoded as categorical variable since some categories would not present in the training set (e.g 1985 and futur years) and we would not be able to predict using unknown categories, hence it was encoded as a numerical variable.
2. Creation of a lot of features created from the past value of car accidents: $X_{t-1}, ..., X_{t-N}$. Each observations is composed of both its value at time t but also of all the values of the past.

The number of features is considered as a parameter to optimize,in this case the best results were found the maximum number of features possible:‘191. Indeed the more features X_{t-i} we input in the model, the more information about the past it will have. Therefore, our dataset consists of 216 rows (for the 216 months from january 1969 to december 1986) and 194 columns. You can find in annexe a screenshot of a part of the dataset.

We could afford to have so much features because the random Forest will itself perform a feature selection.

From that dataset,we could estimate the generalization error by contructing a function to fit the model on 80% of the data, do successive one step ahead predictions, and update the dataset with the new row generated by each new prediction.

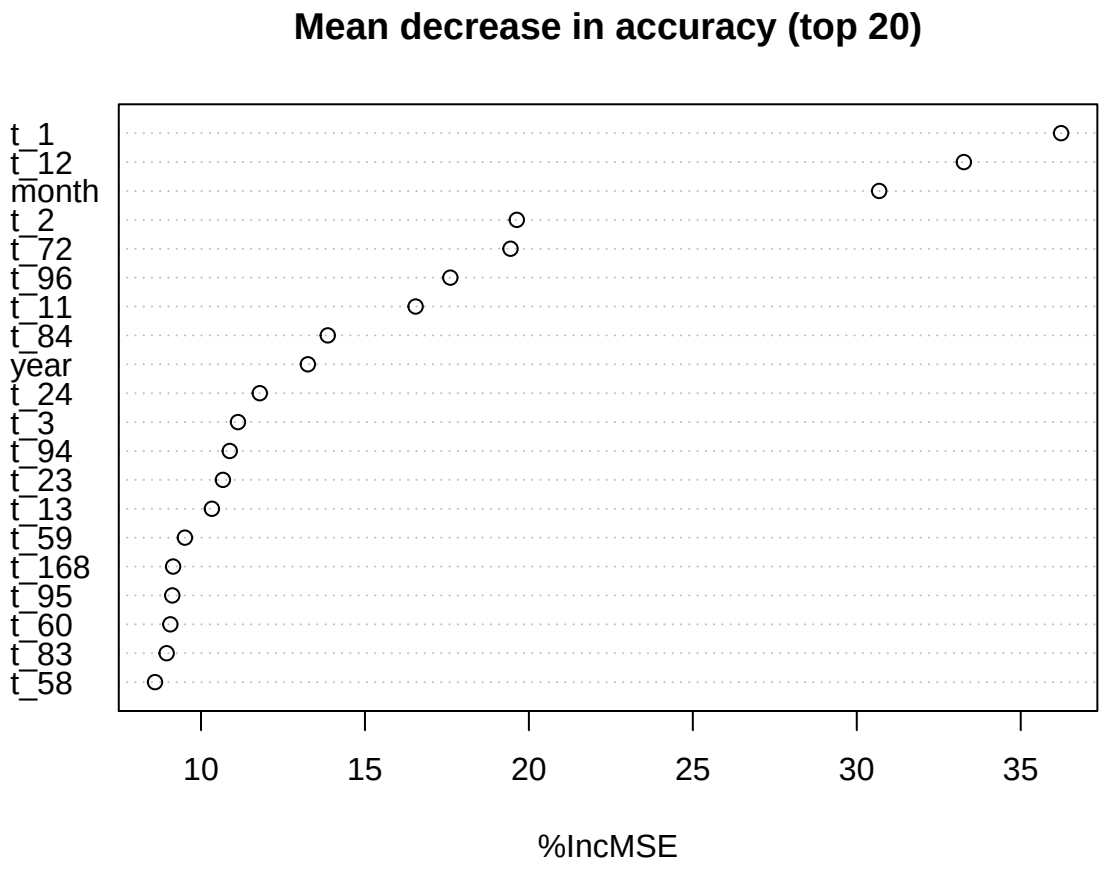
This measure was used to fit the main parameters of the random forest (number of features sampled,number of trees used) and hyperparameters (number of features in the dataset).

After this, we used the random forest model to get an hint about the important variables of the dataset. To quantify the importance of a variable we used *mean decrease in accuracy* criterion. [1]

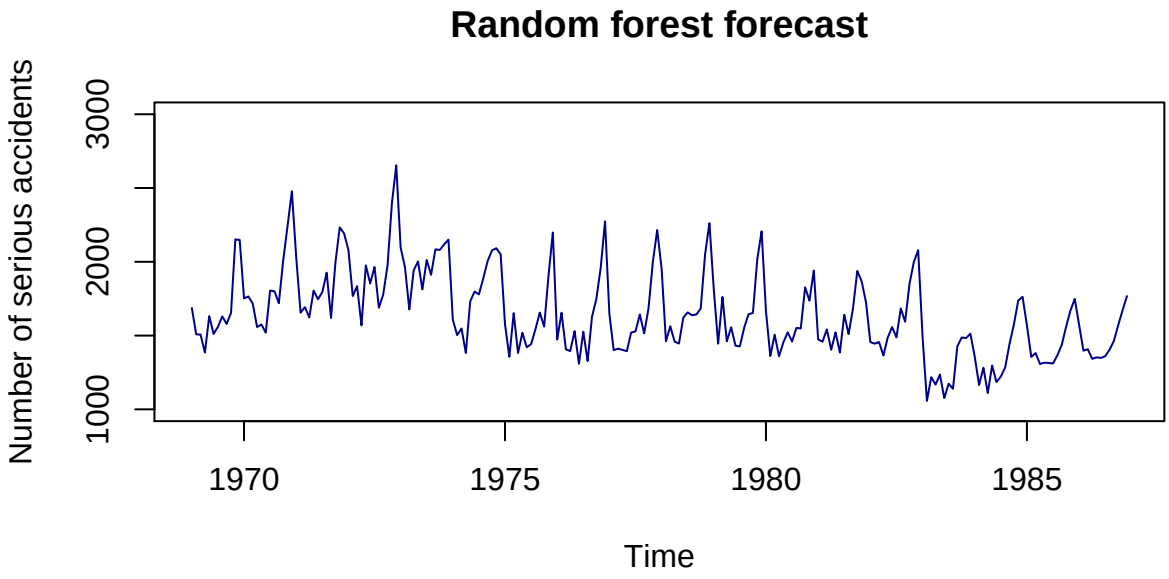
Then, we sorted our features by decreasing order of importance and ploted them.

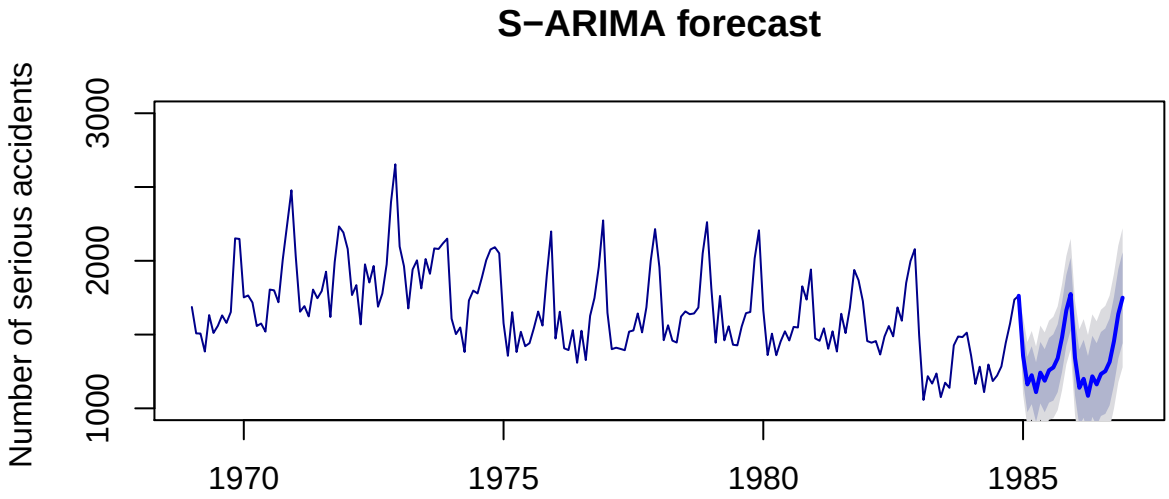
Finally, we used the fitted model to forecast the next 24 months of accidents.we also plotted the SARIMA forecast to facilitate the comparison.

Random Forest one step-ahead prediction error
205.92



[1] To measure the "prediction strength" of a variable j,the accuracy(RMSE) of the model is recorded then the values of the variable j are permuted and the accuracy is again measured.The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.





The RMSE of our model is higher than those of the previous models, but this is not very surprising since we only created basic features.

The measure of the variables importance is more interesting. The 3 more important variables are X_{t-1}, X_{t-12} and month, and had already been found with the Box-Jenkins approach. However other variables are also found important, in a lesser extent, like X_{t-11} who indicate that the seasonality observed in the could be partly explained by a seasonal term of order 11.

Finally, the ploted forecast is more optimistic than the SARIMA forecast.

Conclusion

In this project we first applied the traditional time series models in order to predict the number of severe accendents over a two years period. Our forecasts (SARIMA) suggests that the variable of interest will have a neutral trend for the next two years.
In order to improve the forecast it would be interesting to use multivariate techniques such as regSARIMA and integrate other variables such as the maximum speed authorized or the weather in a given month.
Besides this traditional approach we tried to see how decision trees could be used,considering the simplicity of the approach the results were reasonably good and showed a good potential, particularly if we take into account that such methods does not require any assumption of stationarity.

Annexe

Figure 1: Dataset screenshot

	value	year	month	t_1	t_2	t_3	t_4	t_5	t_6	t_7
198	1316.394	1985	6	1324.502	1306.796	1381.281	1350.021	1550.994	1763.000	173
199	1322.105	1985	7	1316.394	1324.502	1306.796	1381.281	1350.021	1550.994	176
200	1373.876	1985	8	1322.105	1316.394	1324.502	1306.796	1381.281	1350.021	155
201	1430.827	1985	9	1373.876	1322.105	1316.394	1324.502	1306.796	1381.281	135
202	1556.596	1985	10	1430.827	1373.876	1322.105	1316.394	1324.502	1306.796	138
203	1657.051	1985	11	1556.596	1430.827	1373.876	1322.105	1316.394	1324.502	130
204	1734.998	1985	12	1657.051	1556.596	1430.827	1373.876	1322.105	1316.394	132
205	1581.204	1986	1	1734.998	1657.051	1556.596	1430.827	1373.876	1322.105	131
206	1401.427	1986	2	1581.204	1734.998	1657.051	1556.596	1430.827	1373.876	132
207	1413.697	1986	3	1401.427	1581.204	1734.998	1657.051	1556.596	1430.827	137
208	1345.492	1986	4	1413.697	1401.427	1581.204	1734.998	1657.051	1556.596	143
209	1362.373	1986	5	1345.492	1413.697	1401.427	1581.204	1734.998	1657.051	155
210	1357.395	1986	6	1362.373	1345.492	1413.697	1401.427	1581.204	1734.998	165
211	1367.232	1986	7	1357.395	1362.373	1345.492	1413.697	1401.427	1581.204	173
212	1409.831	1986	8	1367.232	1357.395	1362.373	1345.492	1413.697	1401.427	158
213	1465.585	1986	9	1409.831	1367.232	1357.395	1362.373	1345.492	1413.697	140
214	1570.783	1986	10	1465.585	1409.831	1367.232	1357.395	1362.373	1345.492	141
215	1667.329	1986	11	1570.783	1465.585	1409.831	1367.232	1357.395	1362.373	134
216	1753.398	1986	12	1667.329	1570.783	1465.585	1409.831	1367.232	1357.395	136