Introduction to Forecasting Models

Exercise 3: Trend, Seasons and Cycles

Adam Cihlář[1]

20. February 2022

Abstract: The goal of this exercise is to estimate models for predicting the time series of sales of new cars in the United States of America. We employ linear models to decompose the original time series and construct several models from the obtained components. The models' performance is compared based on properties of their prediction errors and prediction metrics. The model consisting from all of the components - trend, season and cycle - surpassed the other simpler models.

Keywords: Linear Regression, Time Series, Automotive Market, Decomposition

Introduction

Purpose of this exercise is to construct various regression models to predict sales of new cars in the United States of America (TOTALNSA) using only the time series itself. Firstly, we visualize the data and split them to training and testing sets to save a portion for the out-of-sample predictions and evaluation of the models. Thirdly, we define and estimate the models and compare them based on how they can fit the data, based on visual analysis and properties of their prediction errors and of course based on their ability to predict future values - forecast metrics. The whole analysis is implemented in statistical software R.

[1] Masaryk University, Faculty of Economics and Administration, Field: Mathematical and Statistical Methods in Economics, 468087@mail.muni.cz

1 Data

In our modelling we will use only the time series of interest for its prediction, respectively its lagged values, and dummy variables constructed with regards to the graphical analysis and our domain knowledge. The observed time series is sales of new cars in the United States of America (TOTALNSA) as reported by Federal Reserve Bank of St. Louis.

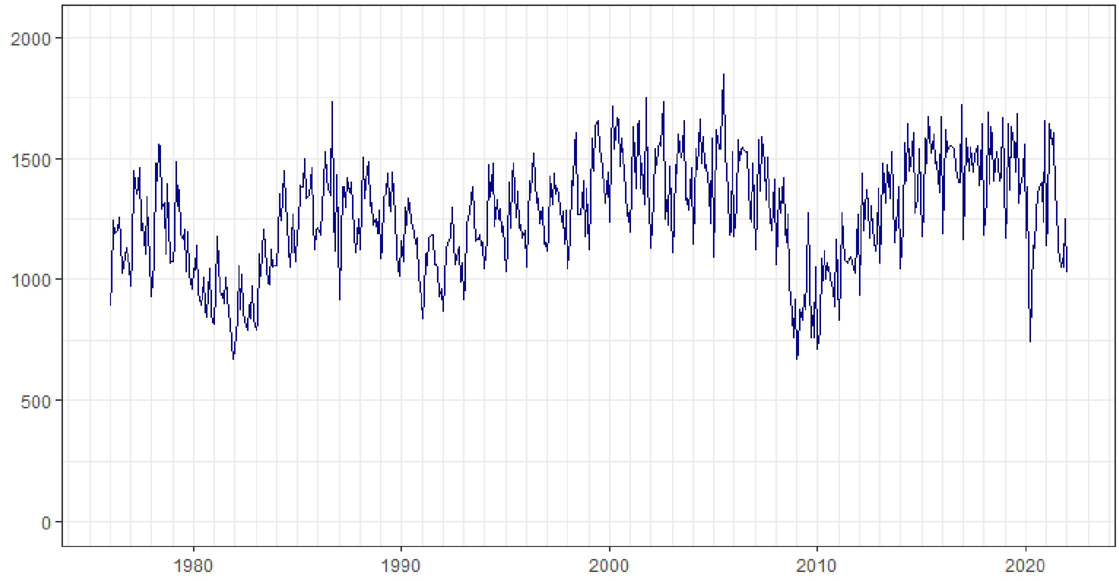

Figure 1: Count of cars monthly sold in the US (TOTALNSA)

The data are available from 1. 1. 1976 to 1. 1. 2022 in monthly frequency. We will omit the first twelve observations as we will use up to twelve lags in our models and use the period from 1. 1. 1977 to 1. 12. 2014 to estimate the models to set up same conditions for in-sample evaluation of the models. The rest of the data - from 1. 1. 2015 to 1. 1. 2022 will be saved to evaluate the models out-of-sample.

2 Models and estimates

We define and estimate following four models to predict the TOTALNSA time series:

1. Trend model:

$$TOTALNSA_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t \tag{2.1}$$

2. Season model:

$$TOTALNSA_t = \beta_0 + \beta_1 Feb + \beta_2 Mar + \beta_3 Apr + ... + \beta_{11} Dec + \varepsilon_t \tag{2.2}$$

3. Trend and season model:

$$TOTALNSA_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 Feb + \beta_4 Mar + \beta_5 Apr + ... + \beta_{13} Dec + \varepsilon_t \tag{2.3}$$

4. Trend, season and cycle model:

$$\begin{aligned} TOTALNSA_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 Feb + \beta_4 Mar + \beta_5 Apr + ... + \beta_{13} Dec + \\ + \beta_{14} TOTALNSA_{t-1} + \beta_{15} TOTALNSA_{t-2} + \beta_{16} TOTALNSA_{t-3} + \\ + \beta_{17} TOTALNSA_{t-4} + \beta_{18} TOTALNSA_{t-12} + \varepsilon_t \end{aligned} \tag{2.4}$$

The trend is modeled by polynomial function of second order, the t and t^2 variables are created as an index from 1 to T, from 1 to T^2 respectively, where T is the number of observations.

We introduce a dummy variable for each month to capture seasonality in the data, except January that serves as a base level to avoid perfect collinearity.

Finally, the cycle is modeled by lagged variables of the original time series. Based on the autocorrelation of residuals of the trend and season model, we use 1, 2, 3, 4, 12 and 13 lags.

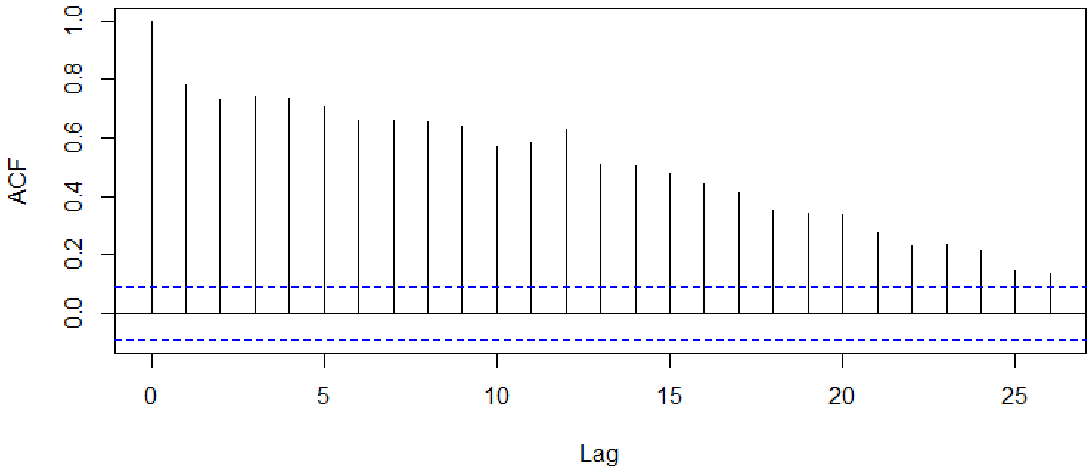

Figure 2: Count of cars monthly sold in the US (TOTALNSA)

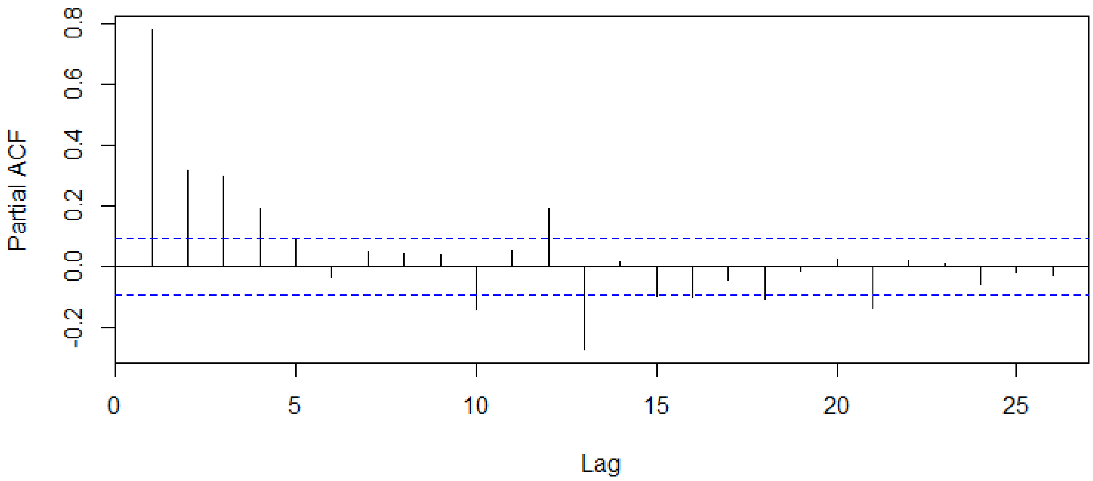

Figure 3: Count of cars monthly sold in the US (TOTALNSA)

The table 2 shows the estimates of all models. With trend only we can explain just 10.7 % of the series' variability, considering seasonality in the data the model captures 24 % of variability of TOTALNSA. The R^2 improves if we capture more components of the time series in the models. For trend and season model we get 0.351 and when considering all of the components - trend, season and cycle - we can explain 83 %. The "full" model clearly outperforms the simpler ones in fitting the training data.

This is also supported by Akaike information criterion and Bayesian information criterion (table 1) and plots of the estimated and true values (figure 4).

	T	S	T+S	T+S+C
AIC	6,160.7	6,105.5	6,037.4	5,440.3
BIC	6,177.2	6,159.1	6,099.2	5,526.9

Table 1: Estimation metrics

	Dependent variable:			
	TOTALNSA			
	T	S	T + S	T + S + C
t	1.986^{***}		2.026^{***}	0.109
t^2	-0.003^{***}		-0.003^{***}	-0.0001
$TOTALNSA_{t-1}$				0.418^{***}
$TOTALNSA_{t-2}$				0.154^{***}
$TOTALNSA_{t-3}$				0.134^{***}
$TOTALNSA_{t-4}$				0.185^{***}
$TOTALNSA_{t-12}$				0.328^{***}
$TOTALNSA_{t-13}$				-0.288^{***}
February		119.141^{***}	124.257^{***}	104.468^{***}
March		358.847^{***}	363.533^{***}	283.024^{***}
April		260.042^{***}	264.306^{***}	172.318^{***}
May		367.988^{***}	371.837^{***}	239.822^{***}
June		337.270^{***}	340.710^{***}	162.188^{***}
July		261.954^{***}	264.993^{***}	70.635^{***}
August		286.814^{***}	289.458^{***}	107.131^{***}
September		175.983^{***}	178.238^{***}	19.937
October		185.970^{***}	187.843^{***}	55.484^{**}
November		92.250^{**}	93.749^{**}	19.568
December		187.228^{***}	188.359^{***}	105.902^{***}
Constant	$1{,}009.534^{***}$	$1{,}019.051^{***}$	782.348^{***}	-39.895
Observations	455	455	455	455
R^2	0.107	0.240	0.351	0.830
Adjusted R^2	0.103	0.221	0.332	0.822
Residual Std. Error	209.6	195.4	180.9	93.3
	(df = 452)	(df = 443)	(df = 441)	(df = 435)
F Statistic	27.1^{***}	12.7^{***}	18.3^{***}	111.7^{***}
	(df = 2; 452)	(df = 11; 443)	(df = 13; 441)	(df = 19; 435)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Estimated models

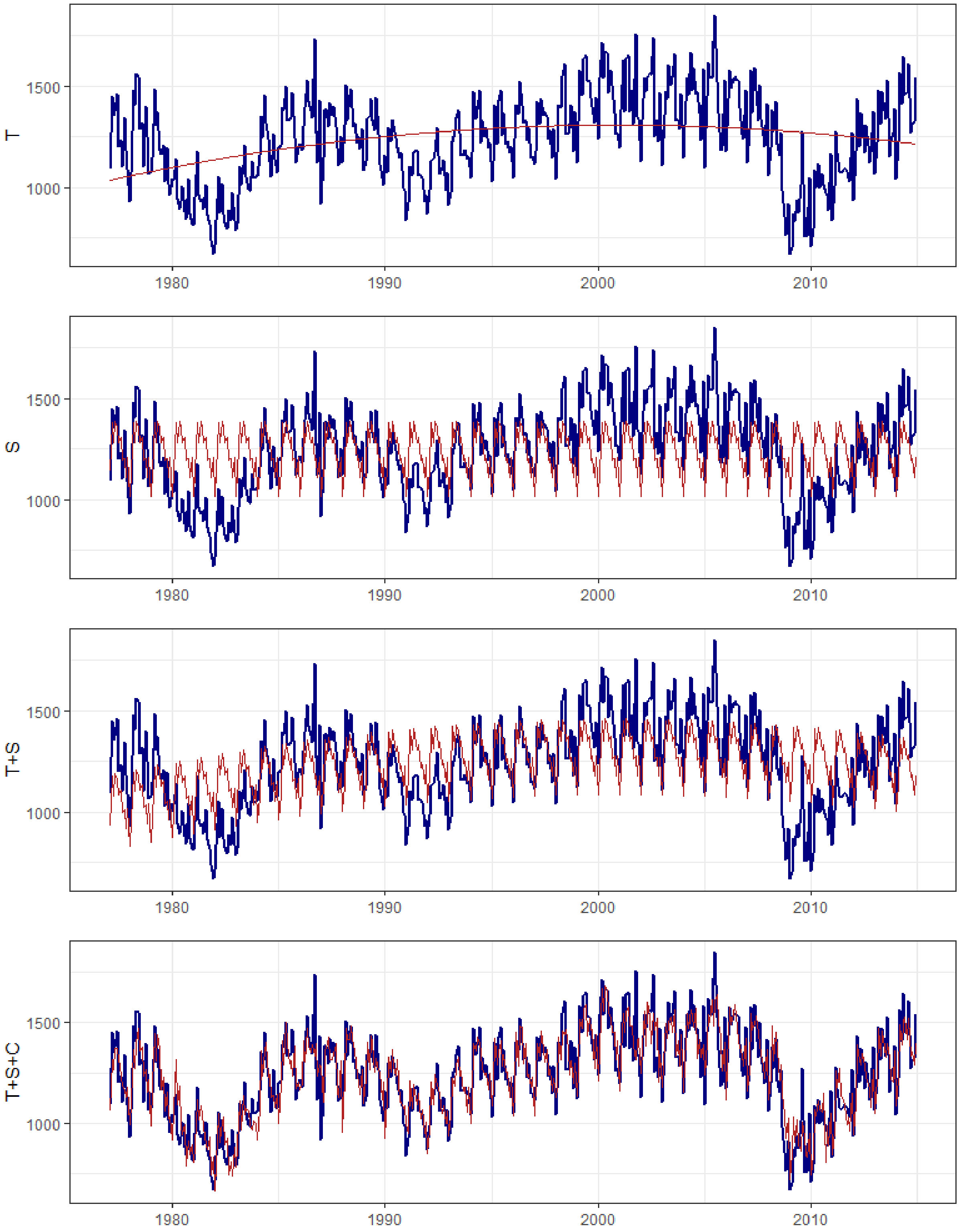

Figure 4: Fitted and true (bold) values

3 Predictions

In this section we will predict the values of TOTALNSA for period from 1. 1. 2015 to 1. 1. 2022 using the four estimated models.

Firstly, let's have a look on various prediction metrics and plot the predicted and true values. Clearly, the last, full model outperforms the simpler models. More interestingly, the model containing trend and season components that performed better in-sample than the first two models, is now the worst in all prediction metrics (table 3).

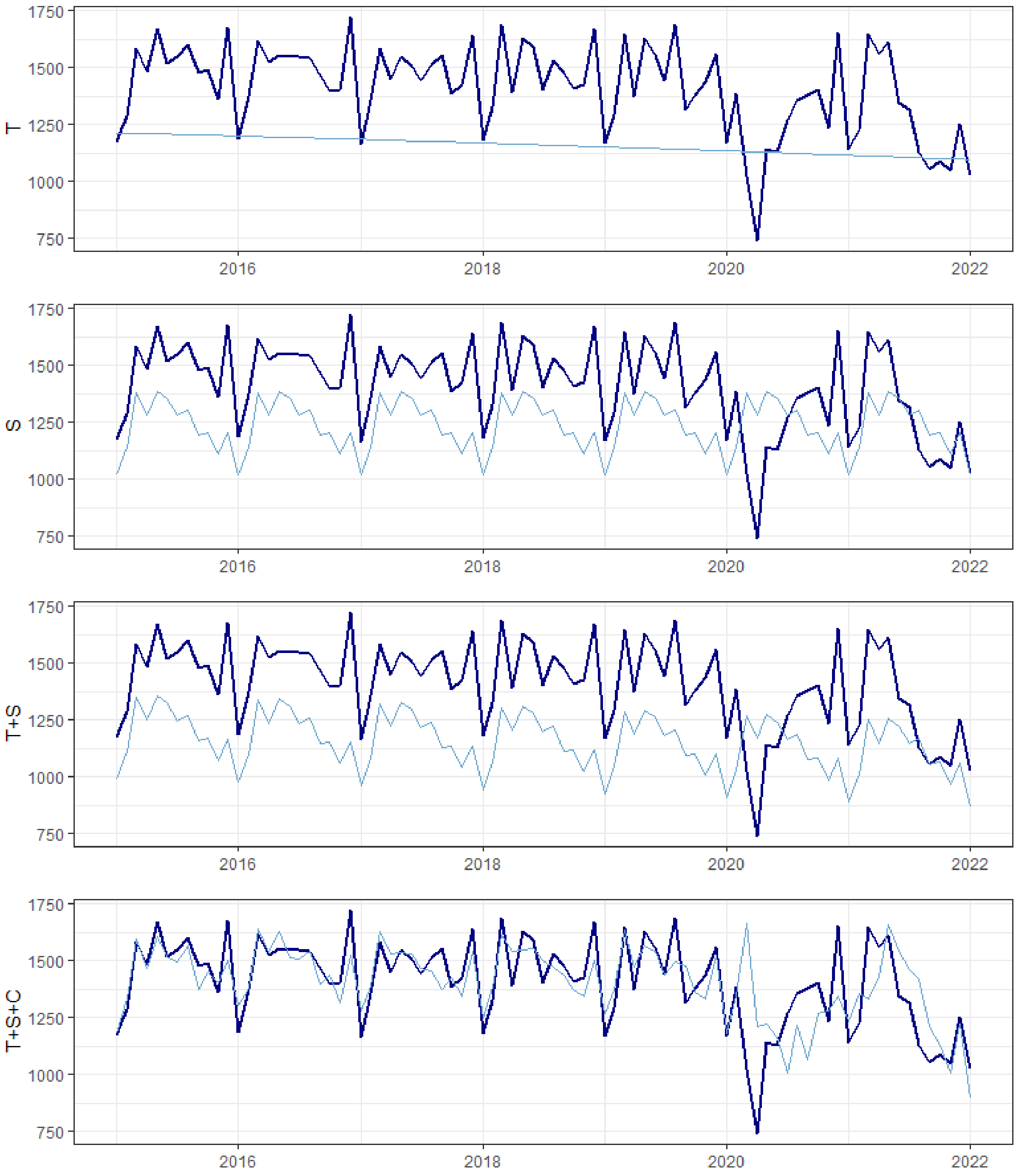

Figure 5: Predictions

	T	S	T+S	T+S+C
MAE	268.423	217.993	278.568	95.021
RMSE	309.759	243.041	300.732	139.841
MAPE	18.1%	15.5%	19.5%	7.4%

Table 3: Prediction Metrics

4 Specification tests

Ultimately, we will have a closer look on the prediction errors and test if the specifications of the models are proper. We will inspect if there is any residual serial correlation in the prediction errors and if the prediction errors are normally distributed. The later is clearly rejected as table 4 and figure 6 shows.

	T	S	T+S	T+S+C
X-squared	6.111	76.663	89.999	117.576
p-value	0.047	< 0.0001	< 0.0001	< 0.0001

Table 4: Jarque-Bera test of normality of prediction errors

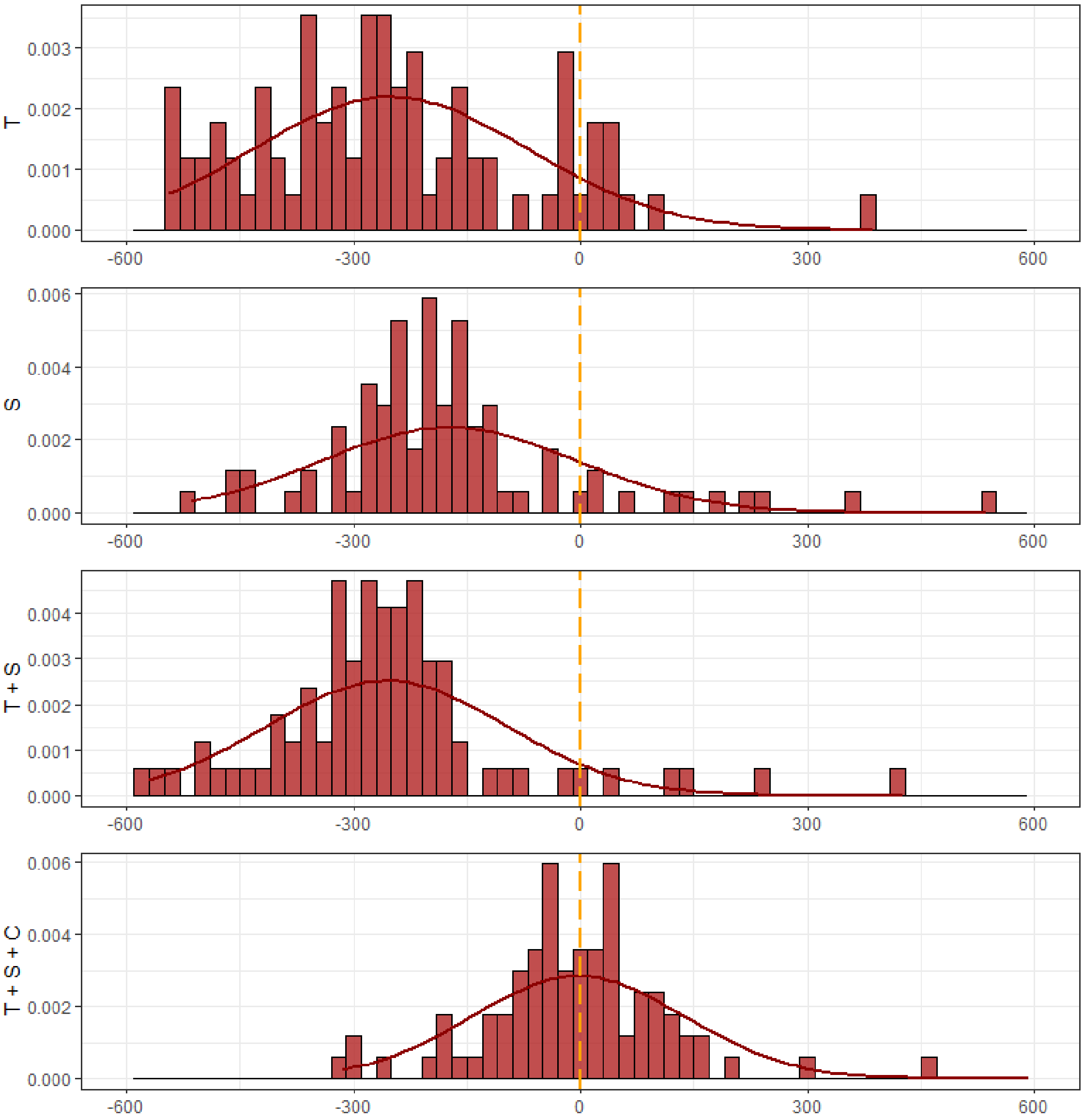

Figure 6: Prediction errors distributions

The correctness of the specification will be tested by creating regression models of prediction error on its lagged values. If the F test of insignificance of parameters is rejected, the model is not correctly specified.

	Dependent variable:			
	Prediction Error			
	T	S	T+S	T+S+C
Err_{t-1}	0.168	0.505***	0.477***	0.268**
	(0.110)	(0.113)	(0.113)	(0.114)
Err_{t-2}	0.139	0.090	0.065	−0.018
	(0.111)	(0.126)	(0.125)	(0.117)
Err_{t-3}	0.250**	0.101	0.068	0.062
	(0.110)	(0.114)	(0.114)	(0.114)
Constant	−110.376**	−50.652**	−100.020***	−3.308
	(44.301)	(24.812)	(34.453)	(15.494)
Observations	82	82	82	82
R^2	0.147	0.377	0.295	0.075
Adjusted R^2	0.115	0.353	0.268	0.039
Residual Std. Error (df = 78)	170.861	139.961	137.763	140.283
F Statistic (df = 3; 78)	4.493***	15.760***	10.901***	2.094
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5: Specification tests

The only model where we do not reject the null hypothesis that all of the variables are not statistically significant is the last full model. The three former models still contain residual serial correlation.

Based on the prediction metrics and properties of the prediction errors we would prefer using model that captures all the time series components.

Conclusion

In this exercise we constructed four various regression models to predict the sales of cars in the US (TOTALNSA). The only data used for the prediction was the time series itself. We created dummy variables to capture trend and seasonality in the data and used lagged values of the variable to model cyclic behaviour of the time series. The data was available from 1. 1. 1976 to 1. 1. 2022 in monthly frequency.

Firstly, we visualized the time series and split the data to save a portion for the out-of-sample predictions and evaluation of the models. Secondly, we specified and estimated the models and compared them based on how they can fit the data, based on properties of their prediction errors and of course based on their ability to predict future values - forecast metrics.

The only model that was not systematically underpredicting was the model containing all the components - trend, season and cycle, this model also perform the best both in-sample and out-of-sample.