Customs Revenues Prediction Using Ensemble Methods (Statistical Modelling vs Machine Learning)

Jordan Simonov and Zoran Gligorov

"There are two kinds of forecasters: those who don't know, and those who don't know they don't know." John Kenneth Galbraith

Abstract

This paper considers the problems associated with prediction of customs revenues by ministries of finance and customs administrations. Accurate predictions of customs revenues result in better liquidity of the central budget, and for that reason, they are extremely important for successful management of public finances. The orthodox approach to forecasting revenues is usually based on forecasting revenues based on tax buoyancy and tax elasticity, with respect of some economic proxy. However, this approach has some shortcomings which can negatively affect accuracy, and for that reason we examine different approaches (machine learning and ensembling). Namely, nowadays in the era of Big Data and digitisation in customs, new approaches based on computer algorithms can give us better results as compared to classic modelling. The paper concludes that using ensemble methods that combine different types of heterogeneous models such as statistical modelling and machine learning can improve forecast accuracy when predicting customs revenues.

1. Introduction

Even though customs revenues are collected by customs administrations within customs procedures, forecasting the collection of such revenues is ordinarily performed by ministries of finance. Nevertheless, ministries also need predictions performed by customs administrations themselves. In this regard, we shall review the most frequently used approaches when planning the revenues.

The orthodox approach to forecasting tax revenues (including customs revenues) is usually based on forecasting revenues based on tax buoyancy and tax elasticity. Tax buoyancy measures the gross elasticity of tax revenues in relation to the respective macroeconomic variable (for instance, import or consumption). The main characteristic of this approach is that it measures the overall elasticity of taxes in relation to their base. In the tax elasticity approach, the time series needs to be first excluded from the discretionary measures of the fiscal policy to calculate the coefficient of the net tax elasticity in relation to the respective macroeconomic variable. Tax elasticity, according to Jenkins et al. (2000, p. 39), is a relevant factor for forecasting and is most often used by ministries of finance when forecasting tax revenues. Furthermore, to obtain more robust forecasts when estimating the elasticities, it is necessary to harmonise them with the business cycle in the economy, which has significant effects on revenue collection. The advantage of such forecasting is that the forecasted revenues are fully correlated with the macroeconomic indicators so that, should they increase, the revenues are expected to correlate with such an increase. However, this forecasting approach also has its shortcomings.

Macroeconomic indicators (which are usually forecasted twice a year) are used when forecasting the revenues, but from the moment of their forecasting to the moment of realisation, a certain period passes which can have negative effects on the forecast accuracy. In fact, Buettner & Kauder (2009, p.7) point out that the circumstances that the forecasters face can significantly affect the accuracy of the forecasts, and this needs to be taken into consideration when evaluating accuracy. Also, timing of the frequency of forecasts can vary (for instance, in Austria the time is 3.5 months; in Italy the time is six months; and in the Netherlands the time is 9.5 months).

To overcome the problems that occur when applying the previous approach, and with the aim of achieving more accurate forecasts, we look at some more flexible approaches that are based primarily on data-driven methods. The use of data-driven methods can be exceptionally useful, since such models provide for forecasting by using high-frequency data and are of particular benefit for cash management and early warning. The main objective of such models is making short-term inflow forecasts (daily, weekly or monthly) for a period not longer than two years (Haughton, 2008, p. 1).

2. Statistical modelling vs machine learning

Nowadays, to increase the accuracy of forecasting models, forecasters apply various approaches based on statistical modelling and machine learning. Modelling assisted by these approaches is usually done in one of the programming languages (for example, R or Python), whereby automated algorithms are used to perform the complex mathematical operations. However, before moving on to practical modelling, we shall first review the basic differences between these two approaches. The basic characteristics of and major differences between statistical modelling and machine learning are outlined in Table 1.

Table 1: Major differences between statistical modelling and machine learning

Statistical modelling	Machine learning
Formalisation of relationships between variables in the form of mathematical equations	Algorithm that can learn from the data without rule-based programming
Required to assume shape of the model curve prior to performing model fitting to the data (e.g. linear, polynomial)	Does not need to assume underlying shape, as machine learning algorithms can learn complex patterns automatically, based on the provided data
Predicts the output with 85% accuracy at a 90% confidence level	Predicts the output with 85% accuracy
Various diagnostics of parameters are performed, such as p-value	Does not perform statistical diagnostic significance tests
Data will be split into 70%/30% to create training and testing data. Model developed on training data and tested on testing data	Data will be split into 50%, 25%/25% to create training, validation and testing data. Models developed on training and hyperparameters are trained on validation data and are evaluated against test data
Models can be developed on a single dataset (training data), as diagnostics are performed at both overall accuracy and individual variable level	Need to be trained on two datasets (training and validation data), to ensure two-point validation
Mostly used for research purposes	Apt for implementation in a production environment
From the school of statistics and mathematics	From the school of computer science

Source: Adopted according to Pratap (2017, p. 43).

Statistical modelling. This type of modelling comprises a wide range of models that could be used for modelling time series. Exponential smoothing (ETS) and autoregressive integrated moving average (ARIMA) (Hyndman & Athanasopoulos, 2016, p. 290) could be considered as two of the most frequently used models in time series forecasting that allow for a complementary approach to the problem. The ETS forecasting model starts from the assumption that a certain legality of the change in observations and their random fluctuations is present in the series, whereby the alignment method gives rise to the so-called 'smoothed series', showing the basic tendency of the time series that is further used for modelling. The predictions of ARIMA forecasting models assume that future circumstances in the time series will be similar to past circumstances. Due to this feature, these models are widely used when modelling a great number of economic series that entail periodic variations.

Machine learning. The term 'machine learning' was first used in 1959 by Arthur Samuel, who was at the time working at IBM, and described it as the field of study that gives computers the ability to learn without being explicitly programmed (Gutierrez, 2015, p. 17). In the era of Big Data and digitisation in customs administrations, machine learning algorithms can be advantageous, especially in predictive analytics. In general, these algorithms can be divided into two major types: supervised and unsupervised. Supervised learning algorithms are those in which a machine learning model is scored and tuned against some smart of known quantity, while unsupervised learning algorithms are those in which machine learning derives patterns and information from data while determining the known quantity tuning parameter itself (Burger, 2018, p. 40). In this regard, we shall look at the application of the Artificial Neural Network (ANN), frequently used for modelling time series. In general terms, ANN algorithms are based on simple mathematical models of the brain and they allow complex nonlinear relationships between the response variable and its predictors (Hyndman & Athanasopoulos, 2016, p. 443). To illustrate, the human brain consists of approximately 85 billion neurons, which creates a network capable of absorbing huge quantities of knowledge, whereas the number of neurons in animals is far lower – for instance, cats have 1 billion neurons and mice have 75 million neurons (Lantz, 2015, p. 220).

Ensembles. Bates and Granger (1969, p. 451–468), in their famous paper 'The Combination of Forecasts', point out that combining forecasts often leads to better forecast accuracy. Even though this approach is more than half a century old, the point of ensembling is not very far from this idea. Namely, this is the reason for using ensembles, whereby it is always considered better if they consist of heterogeneous types of models to better cover different aspects of the time series. The output of these models is most commonly based on the average projections given by models upon the voting, weighting or other type of selection. This method is more often applied in machine learning, whereby special algorithms automatically create ensembles. A specialised type of supervised learning algorithm is ensemble learning, which is a set of algorithms that is built by combining results from multiple machine learning algorithms. These methods have become popular due to their ability to provide superior results and the possibility of breaking them into independent models to train on distributed networks. Some of the most popular ensemble machine learning methods are boosting, bagging, gradient boosting machines and random forest. It is worth mentioning Kaggle, a subsidiary of Google, and the largest online community of data scientists, frequently organises machine learning competitions involving the use of forecasting methods. Often the winning solutions are based on variations of the ensemble methods strategy (Gutierrez, 2015, p. 239). As the discussion above shows, these data-driven approaches provide a solid basis for revenue modelling. However, the question that inevitably arises is, 'which of these approaches can provide us with better projections?'.

3. Data

For the purposes of this research, a dataset with customs duties from the Republic of North Macedonia was used. This dataset consists of a univariate time series with monthly frequency from January 2014 to January 2020. Taking into consideration that the Republic of North Macedonia still applies the 1986 government finance statistic (GFS), this is on a cash basis according to their payment. To better elaborate revenue collection related to customs duties, we shall look at several basic facts related to the customs protection of the Republic of North Macedonia. Customs duties account for three per cent of the total budget revenue or 0.9 per cent of GDP. According to the report *The World Tariff Profiles* (World Trade Organization [WTO], International Trade Centre [ITC] and the United Nations Conference on Trade and Development [UNCTAD], 2018), by using the same methodology a comparison of trade weighted average was made, which in the European Union (EU) is three per cent, while in Republic of North Macedonia it is calculated at 6.3 per cent. The Republic of North Macedonia has concluded Free Trade Agreements (FTA) with the following parties: the EU, the European Free Trade Association (EFTA), Turkey (TR), Ukraine (UA) and the Central European Free Trade Agreement (CEFTA). Considering that trade with CEFTA is fully liberalised, that is zero tariff rates, it is not going to be subject to analysis in the study below. One of the most important FTAs for the Republic of North Macedonia is the Stabilization and Association Agreement (SAA) regulating foreign trade with the EU, in line with which around 70 per cent of the total foreign trade is realised. Figure 1 shows the distribution of most favoured nation (MFN) tariff rates, as well as the range of customs protection of agricultural products and non-agricultural products. The dotted lines present the simple mean for each of these product groups, respectively.

Figure 1: Distribution of MFN and FTA tariff rates by type of products

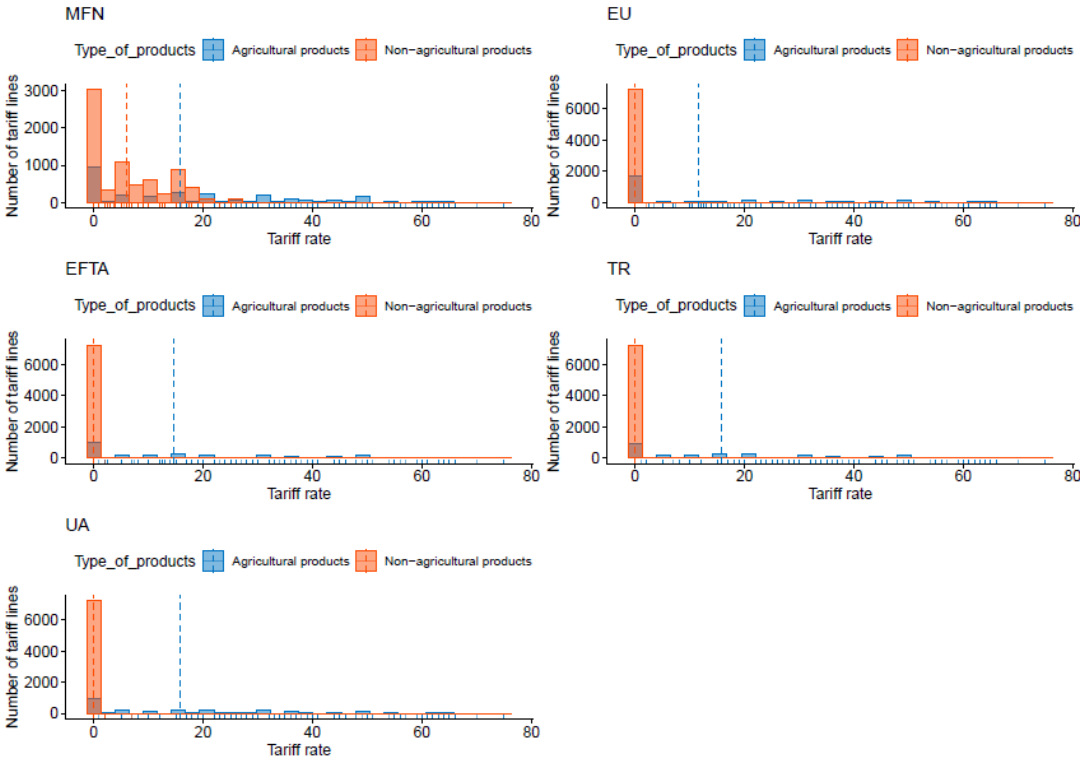
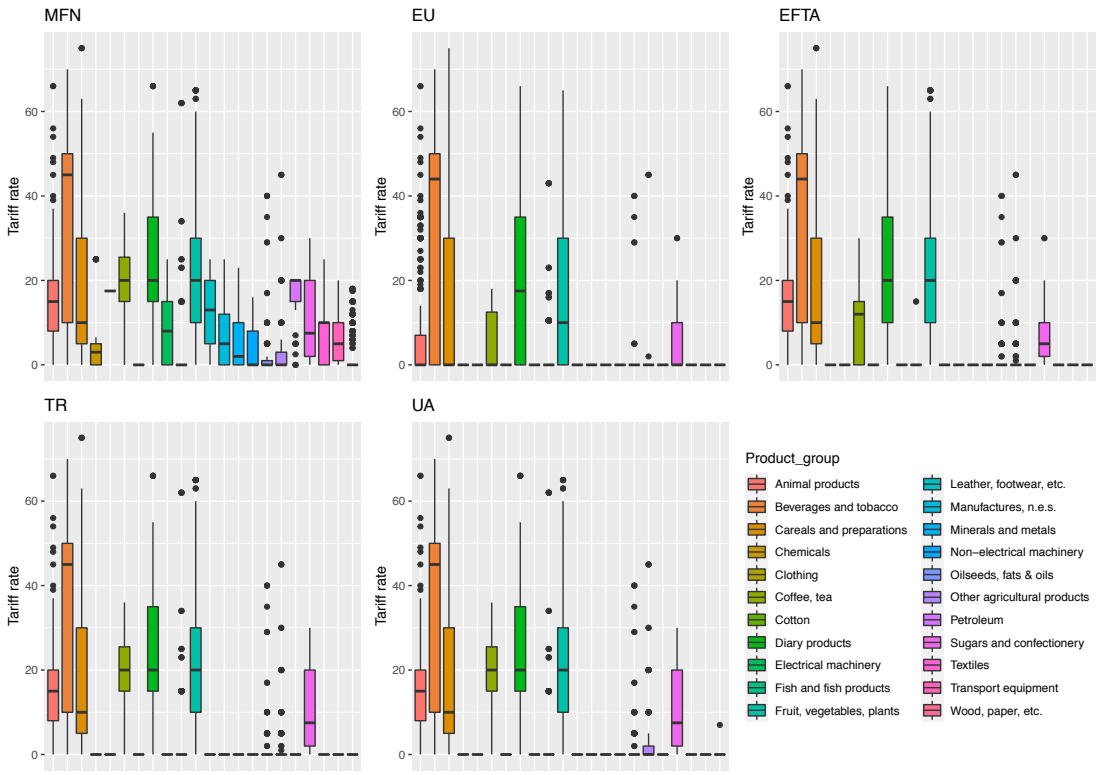


Figure 2 shows a boxplot of the descriptive statistics of tariff rates based on a five number summary (minimum, first quartile, median, third quartile, and maximum plus outliers) for the indicated 22 groups of products shown in a rectangular form. The horizontal line in the middle of the rectangles displays the median of the data, or in this case, the amount of the tariff rates. The horizontal line under the median displays the first quartile of the customs duties, while the line above the median displays the third quartile of the tariff rates for the respective group of products. The box itself shows where 50 per cent of the central data in the variation series (interquartile range) is located, and the length of each vertical line (whisker) corresponds to one and a half length of the interquartile range, while all data above these lines that are outliers are marked as dots.

Figure 2: Boxplot (MFN and preferential tariff rates)



n.e.s.: not elsewhere specified

4. Exploratory data analysis

Explanatory data analysis (EDA) is an approach to analysing data sets to summarise their main characteristics by using visual methods and statistical tests. Performing such an analysis is a mandatory precondition for successful modelling and is it highly recommended to conduct it before any forecasting of customs revenues.

To better understand data properties, find patterns and suggest a modelling strategy, we began with EDA. The descriptive statistics of monthly collection of customs duties in Table 2, with main central tendency measures, show that customs duties collection measured through simple averages account for 416.6. Standard deviation, which measures the distance from the median value ranges, is 68.6.

Although the average of customs duties collection (expressing the central tendency of the data) amounts to 416.6, the median value is lower and accounts for 411.9. This is also confirmed by the trimmed mean (which at the level of 10 per cent excludes the lowest and highest values at the daily collection and then calculates the average), thereby accounting for 413.1. The median absolute deviation is a robust measure of statistical dispersion and is more resilient to outliers in a dataset than a standard deviation. It indicates a deviation of 59.4. Monthly collection of customs duties is in the range of 249.9 to 591. The asymmetry coefficient is positive, amounting up to 0.04, thus indicating a positive skew (that is, the right tail is longer and the mass of the distribution is concentrated on the left side). In addition, the coefficient kurtosis records low values that are lower than three, thus confirming non-normal data distribution, which is 0.1 and value of the standard error is 8.0.

Table 2: Descriptive statistics of monthly collection of customs duties in MKD (Macedonian Denar)

Mean	**416.6**
Standard deviation	68.6
Median	411.9
Trimmed mean	413.1
Median absolute deviation	59.4
Min	249.9
Max	591.0
Range	341.1
Skew	0.4
Kurtosis	0.1
Standard error	8.0

Figure 3: Boxplot of customs duties of monthly collection in MKD

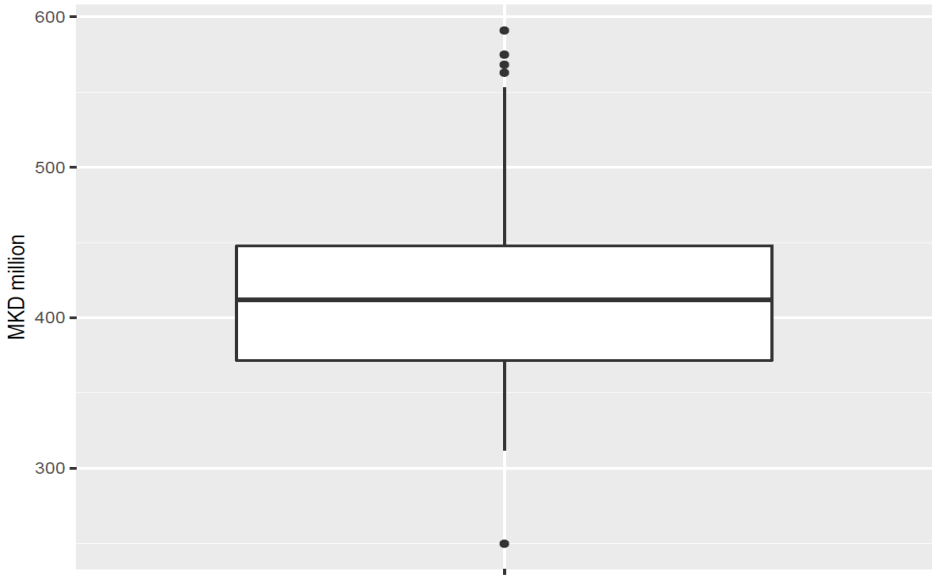


Figure 3 represents a boxplot with a graphic display of the descriptive statistics based on a five number summary (minimum, first quartile, median, third quartile, and maximum plus outliers) of the monthly collection of customs duties. The horizontal line in the middle of the rectangle displays the median of the data, or in this case, the amount of collected duties in national currency. The horizontal line under the rectangle displays the first quartile of the customs duties, while the line above displays the third quartile of the collection of customs duties. The box itself shows where 50 per cent of the central data in the variation series (interquartile range) is located, and the length of each vertical line (whisker) corresponds to one and a half length of the interquartile range, while all data above these lines that are non-standard observations (outliers) are marked as dots.

Upon visual inspection of the boxplot in Figure 3, a crucial characteristic of the time series is evident: the existence of outliers which should be adequately treated in the data preprocessing procedure, since they usually lead to non-stationarity, and may also affect the accuracy of the projections.

Figure 4 shows four charts, the first of which refers to the original time series, whereas for the rest of the three charts the STL method (Seasonal and Trend decomposition using Loess) for decomposition was applied, whereby the time series has been divided into three parts: trend, seasonal and residual component. Upon visual inspection of the lineplot showing the original time series, a trend in the data is evident, which usually leads to non-stationarity of the time series. This assertion was checked through the KPPS-test (Kwiatkowski at al., 1992, p. 159–178) whereby the result for the test statistic for critical value for a significance level of 5 per cent is 0.463. As the test-statistic is 1.79, we can see that it fails the KPSS unit root test for stationarity. The detailed results from this KPSS test are provided in Table 1A in Annex A. A characteristic of the trend component is that it is constantly increasing, and this is due to several factors such as an increase in economic activity and thus an increase in imports, which on the other hand affects the increase of revenue collection at a nominal value. The seasonal component describes the seasonal character of the data in the form of fluctuations in the time series, related to calendar cycles. The seasonal type of data may have a significant effect on the projections, and thus they should be adequately treated – that is, the data needs to be deseasonalised. The remaining component refers to the residuals from the seasonal plus trend fit.

Figure 4: STL decomposition

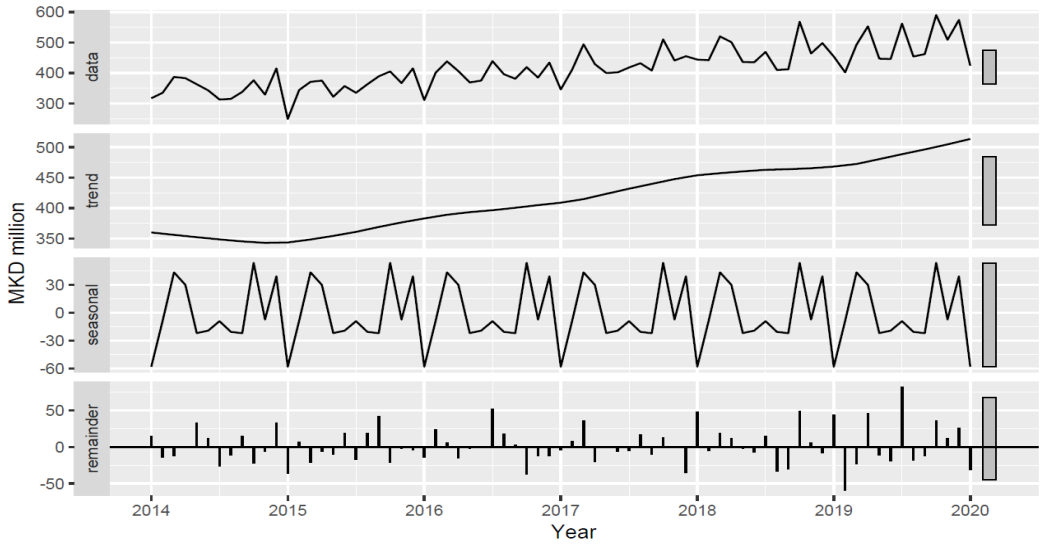
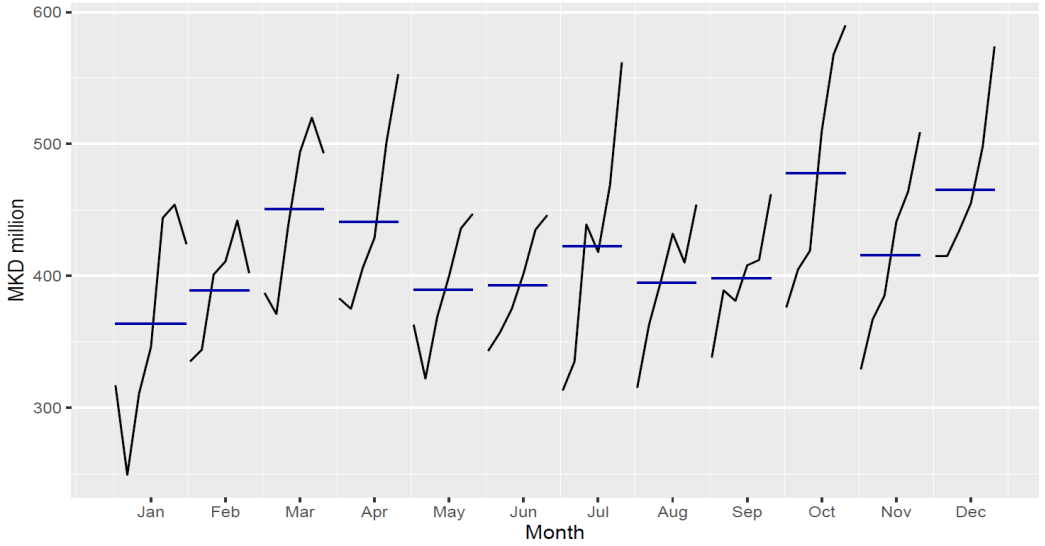


Figure 5: Seasonal plot of monthly collection of customs duties in the period of 2014−2020

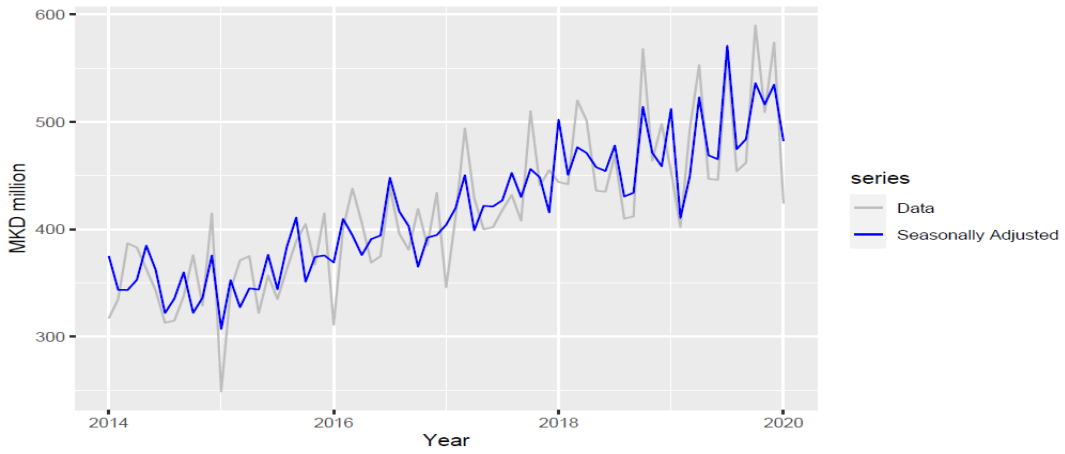


To better familiarise ourselves with the seasonal patterns of time series, a seasonal plot, shown in Figure 5, was used. The horizontal blue lines indicate the means for each month and show the changes in seasonality over time. From this plot, we can conclude that the last quarter is usually strongly affected by season, and collection of customs duties is higher as compared to other quarters. This can also be confirmed by averages shown on this plot.

Finally, it can be concluded that EDA has revealed many important features of this time series by providing mathematical and statistical proof related to seasonal patterns, non-standard observations (outliers) and non-stationarity that must not be neglected during modelling. Such volatility of the time series may be explained by the influence of seasonality as well as with the measures of

the discretionary fiscal policy, which, in the analysed period appears in the form of adoption of autonomous measures for reducing certain tariff rates. Considering that the series has a seasonal component, its seasonal adjustment was performed with the assistance of an automated procedure that uses the decomposition method developed by the US Census Bureau and Statistics Canada, also known as X11 (Hyndman & Athanasopoulos, 2016, p. 216). Additionally, we used the automated function that performs outlier replacements on the linear interpolation principle. The seasonally adjusted series is shown in Figure 6.

Figure 6: Original series and seasonally adjusted series



5. Predictive modelling

In this part we focus on the most likely scenario of income planning, or, in other words, it is a scenario in which the data source is limited and data of only one time series are at our disposal. We chose this scenario because forecasters in the real word of data often face situations when at the time of income projection they do not have the remainder of the data at their disposal because the data publication frequency may be different. For modelling purposes, we divided the data into a training set (80 per cent of the observations) with 58 observations relating to the period from January 2014 until October 2018, and a test set (20 per cent of the observations) with 14 observations relating to the period from November 2018 to January 2020. This approach is also known as hold-out and is used for training the models, based on the training set, as well as for testing the predictive performance of the test set. We applied this approach to avoid overfitting, which often occurs during such modelling when the models have good results when it comes to the training set, but when it comes to the test set they are far more dissatisfactory.

For the evaluation of the forecast accuracy, we used different types of metrics, where we calculated the errors separately for each model with: Mean Error (ME), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and first-order autocorrelation (ACF1). When making the main comparison between the errors we used MAPE and RMSE. We chose these accuracy metrics as MAPE and RMSE are commonly used when mutually comparing the errors made by different models because these metrics are more pessimistic measures since they give more weight to larger errors.

For the purposes of this predictive modelling, we chose the R ecosystem with different packages as it is one of most user-friendly languages, which operates with automated algorithm functions.

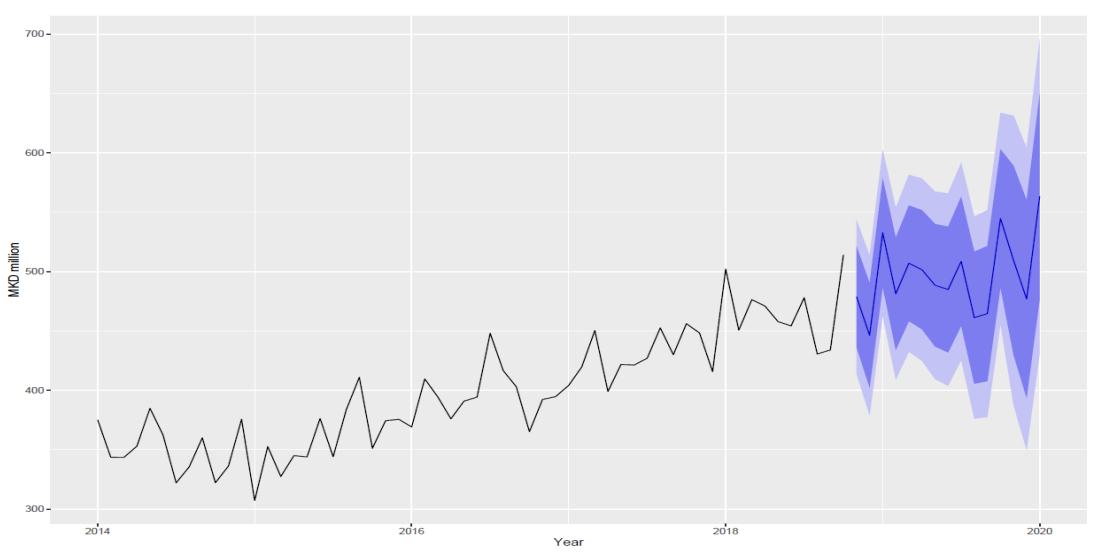
5.1 Statistical modelling

For the purpose of this kind of modelling we used the auto.arima function from the forecast package in the R ecosystem. This function enables modelling with ARIMA. Determining the order of the ARIMA model is usually a complex task and this becomes even more complicated if it concerns a model with seasonal data. Therefore, for the purposes of this modelling, the automatic ARIMA algorithm (Hyndman & Khandakar, 2008, p. 1–22) was used. This algorithm automatically selects the best model for the specified time series, combining single-root tests to convert the series from non-stationary to stationary (on an off-seasonal and seasonal basis) and to determine the order of the ARIMA model on a non-seasonal and seasonal level through the Akaike information criterion (AIC) .

By using this automated approach with the auto.arima function, the model for the time series customs duties is specified as ARIMA$(0,1,1)(0,1,0)_{[12]}$. The detailed results from this modelling are provided in Table 2A and Table 3A in Annex A.

Before deciding whether the selected model is adequate for further use and forecasting, we needed to perform some additional tests related to the residuals. It is always a good idea to check if there is any autocorrelation between the residuals or whether they are normally distributed. For that reason, we performed additional residual diagnostics. Results from this diagnostic are shown in Annex A, Figure 11. The forecast accuracy can only be determined by considering how well a model performs on the test set that was not used when fitting the model within the training set. In the training set, this model has a RMSE of 28.8, a MAPE of 5.05 and a MASE of 0.568, while on the test set the RMSE is 40.5, the MAPE is 6.61 and the MASE is 0.870. MASE metrics compares the model predictive performance (MAE) to the naive forecast on the training and the test set and has a value below one, which indicates that the model has a lower average error than the naive forecasts. Considering that the residual diagnostics showed good results, we considered this model to be adequate for forecasting and we used it to make a forecast with a horizon of 14 + 1 observations. The results are shown in Figure 7.

Figure 7: Forecast from ARIMA $(0,1,1)(0,1,0)_{[12]}$

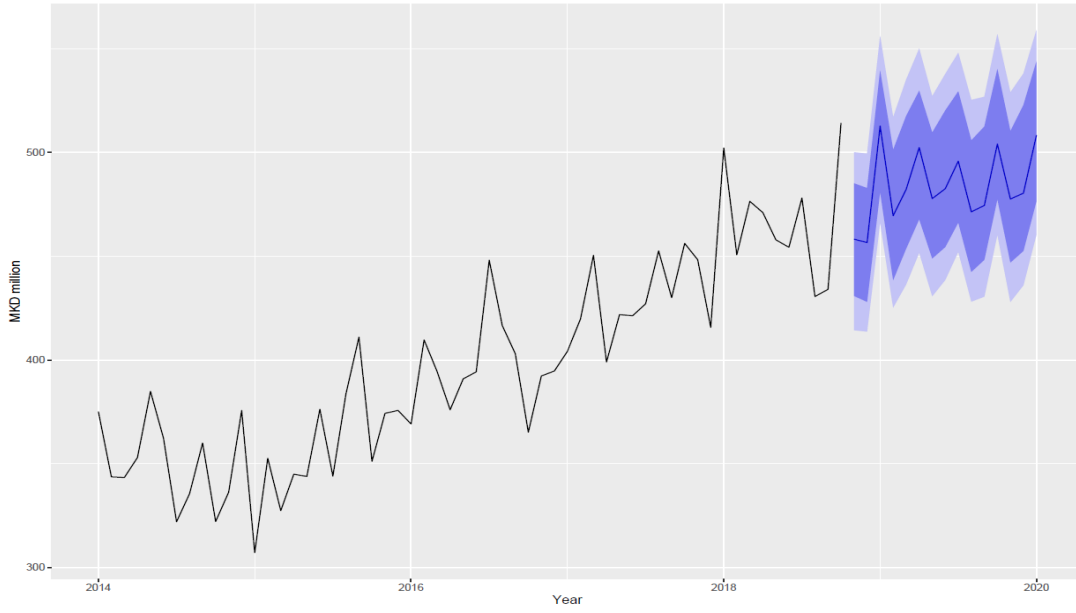


5.2 Machine learning

Given that the time series differs from other types of data series, is subject to modelling, we modelled it with a neural network that supports solving regression problems such us customs revenue forecasting.

In this case we used an automatic procedure which we performed with the assistance of the nnetar function from the forecast package in the R ecosystem, which can operate with neural network autoregression (NNAR). Here, the feature selection is automatic and lag-based features are selected. This function has four main arguments: repetitions, p, P and size. The first argument in this function repeats the number of neural networks fitted. By default, this argument is 20 unless otherwise specified. For the seasonal series argument, p is chosen from the optimal linear model fitted to the seasonally adjusted data while P = 1 by default. These functions only fit neural networks with a single hidden layer, where the last argument – size – refers to the number of nodes in the hidden layer. By default, this argument is estimated in this way: size = (p + P + 1)/2 (and rounded to the nearest integer). If the values for the arguments in the function are not specified, then they are automatically selected. When it comes to forecasting, the network is applied iteratively. For forecasting one step ahead, the nnetar function simply uses the available historical inputs and this process will continue consequently, two steps ahead (the one-step forecast as an input, along with the historical data) and so on (Hyndman & Athanasopoulos, 2016, p. 446). By using this automated approach with the nnetar function, the model for the time series customs duties is specified as NNAR $(3,1,2)_{[12]}$. The detailed results from this modelling are provided in Table 4A and Table 5A in Annex A. Results from this diagnostic are shown in Annex A, Figure 12. Considering that we already performed residual diagnostics, our next step was to evaluate the forecast accuracy. The forecast accuracy can only be determined by considering how well a model performs on the test set that was not used when fitting the model within the training set. On the training set this model has a RMSE of 21.8 and a MAPE of 4.19 and on the test set the RMSE is 49.7 and the MAPE is 7.7. Considering that the residual tests showed good results, we considered this model adequate for forecasting and we used it to make a forecast with a horizon of 14 + 1 observations. The results are shown in Figure 8.

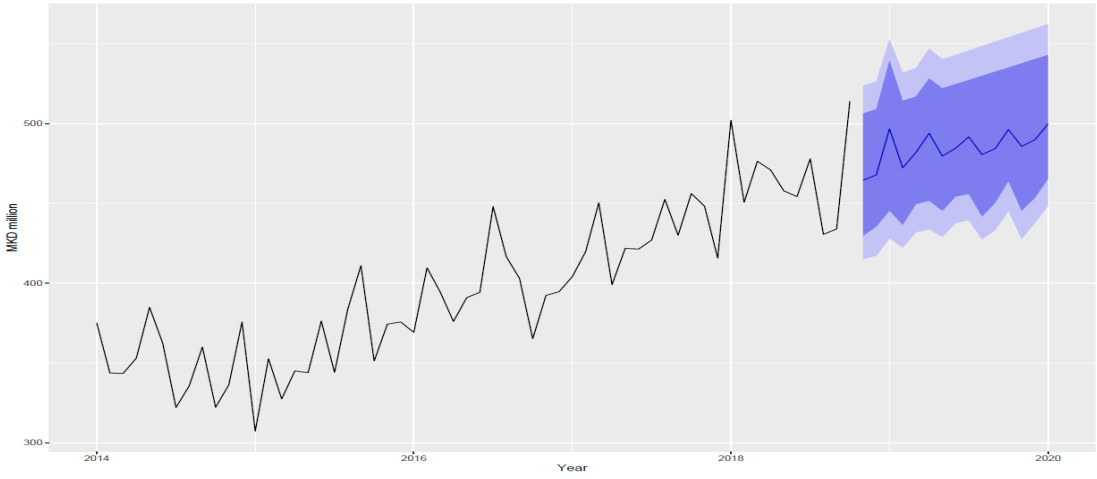
Figure 8: Forecast from NNAR (3,1,2) $_{[12]}$



5.3 Ensemble modelling

Our idea here was to combine the results of the previous models and for that purpose we used the hybridModel function from the forecastHybrid package in the R ecosystem. This function gives modelling a platform to ensemble heterogeneous time series models. In this modelling framework we used ARIMA and NNAR. We already saw the predictive performance in the separate models and now we ensembled this model. By using this automated approach, the model for the time series customs duties was comprised of the ARIMA and NNETAR models, with weights 0.452 for ARIMA and 0.548 for NNETAR. The detailed results from this modelling are provided in Table 6A and Table 7A in Annex A. Results from this diagnostic are shown in Annex A, Figure 13. In the training set this model has a RMSE of 21.8 and a MAPE of 4.00 and on the test set the RMSE is 31.2 and MAPE is 5.0.
As shown in the results, the ensemble method gave the best results compared to all previously tested models. In fact, compared to the RMSE of the test set, this model has a value of 31.2. This error was larger in all the other models, ARIMA was 40.55 and NNETAR was 49.7. The MAPE of this model has a value of 5.0 while in all other models, this error was bigger, ARIMA 6.61 and NNETAR 7.7. This example shows that we can consider the models separately as weak learners because, individually, they can have poor predictive performance, while the ensemble technique of combining different models provided better performance of the models, so that the individual error in the models for RMSE has been reduced from 49.7 to 31.2 and for MAPE it was reduced from 7.7 to 5.0. These results are shown in Figure 9.

Figure 9: Forecast from ensemble



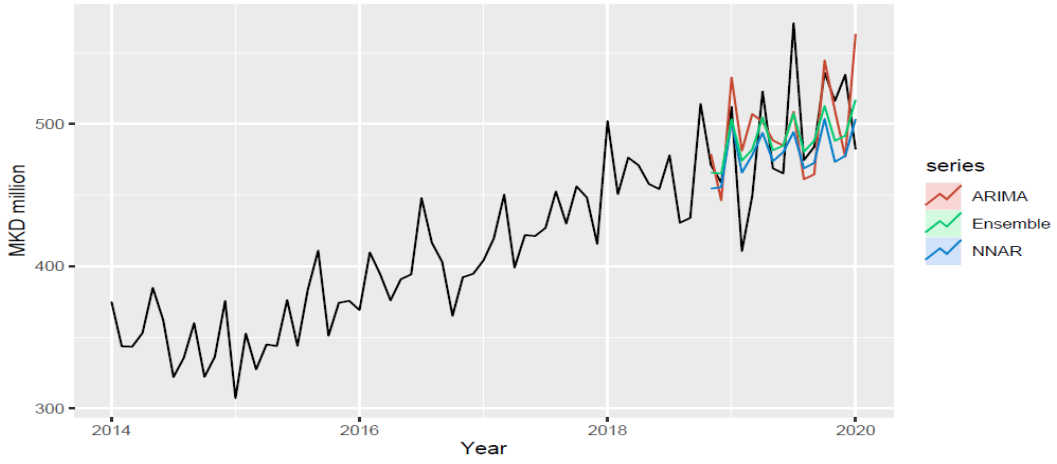
6. Conclusion

For the objectives of this research, we have modelled a univariate time series by using three different approaches within the R ecosystem.

By using statistical modelling we specified the ARIMA $(0,1,1)(0,1,0)_{[12]}$ model as the most adequate one to forecast time series. This model has a RMSE of 28.8 and a MAPE of 5.05 while on the test set the RMSE is 40.5 and the MAPE is 6.61. Statistical modelling combined with some classical models can give good results but, generally, all these models have problems when it comes to handling Big Data (for example, a sample of around one thousand rows). For this reason, it may be better to use all these models with data that have monthly, quarterly and yearly frequency or smaller samples.

Although machine learning, as part of artificial intelligence, is widely used in many other areas of classification problems (for example, customs fraud detection and detection of underpricing evasion), this research has shown how we can use this kind of modelling for forecasting time series and projecting customs duties. By using machine learning we specified the NNAR $(3, 1, 2)_{[12]}$ model as the most adequate one to forecast time series. This model has a RMSE of 21.8 and a MAPE of 4.19 and on the test set the RMSE is 49.7 and the MAPE is 7.7. As a relatively new area of time series analysis, compared to the classical statistical methods, machine learning has shown solid results on a relatively small sample and its use deserves more attention especially in more complex data. Point forecasts from ARIMA, NNAR and Ensemble are show in Figure 10.

Figure 10: Point forecasts from ARIMA, NNAR and Ensemble applied to customs duties collections



As for the two approaches,[1] the ensemble technique has shown that it can improve prediction accuracy compared to the individual models. This example shows that we can consider the models separately as weak learners because individually they can have poor predictive performance, while the ensemble technique of combining different models provided better performances of the models, so that the individual error in the models for RMSE was reduced from 49.7 to 31.2 and for MAPE it has been reduced from 7.7 to 5.0. They can reduce the forecasting error and, finally, we can conclude that the ensemble technique is certainly a game changer and must be an important addition to every forecaster's toolbox. For this reason, we strongly recommend using this technique for forecasting purposes with statistical modelling or machine learning.

References

Buettner, T., & Kauder, B. (2009). *Revenue forecasting practices: Differences across countries and consequences for forecasting performance*. CESifo Working Paper, No. 2628, Center for Economic Studies and ifo Institute (CESifo).

Burger, S. V. (2018). *Introduction to Machine learning with R*. O'Reilly Media.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*(4), 451–468.

Gutierrez, D. D. (2015). *Machine learning and data science: An introduction to statistical learning methods with R.* Technics Publications.

Haughton, J. (2008). *Manual on tax analysis and revenue forecasting*. https://cpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/8/1443/files/2019/06/TaxManual.pdf

Hyndman. R. J., & Athanasopoulos, G. (2016). *Forecasting: Principles and practice*. Monash University.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1–3), 159–178.

Lantz, B. (2015). *Machine Learning with R* (2nd ed.). Packt Publishing Ltd.

Jenkins, P. G., Kuo, C., & Shukla, G. P. (2000). *Tax analysis and revenue forecasting: Issues and techniques.* Harvard Institute for International Development, Harvard University.

Pratap, D. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

World Trade Organization (WTO), International Trade Centre (ITC) and United Nations Conference on Trade and Development (UNCTAD) (2018). *World Tariff Profiles 2018*.

Notes

1 Code is available on GitHub https://github.com/jordans78/Ensemble-Methods

Annex A

Table 1A: KPSS Unit Root Test

Value of test-statistic is: 1.7909

Critical value for a significance level of:

	10pct	5pct	2.5pct	1pct
Critical values	0.347	0.463	0.574	0.739

Table 2A: ARIMA(0,1,1)(0,1,0)[12]

Call:

ARIMA(0,1,1)(0,1,0)[12]

Coefficients:

ma1
-0.7164
s.e. 0.1148

sigma^2 estimated as 1094: log likelihood=-221.16

AIC=446.32 AICc=446.61 BIC=449.94

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	4.416098	28.81491	20.82767	1.062319	5.051560	0.5679152	-0.02734069	NA
Test set	6.142640	40.55658	31.93333	-1.724624	6.611157	0.8707370	-0.03343311	0.7099668

Table 3A: Ljung-Box test

data: Residuals from ARIMA(0,1,1)(0,1,0)[12]

Q* = 5.654, df = 11, p-value = 0.8954

Model df: 1. Total lags used: 12

Figure 11: Residual diagnostics

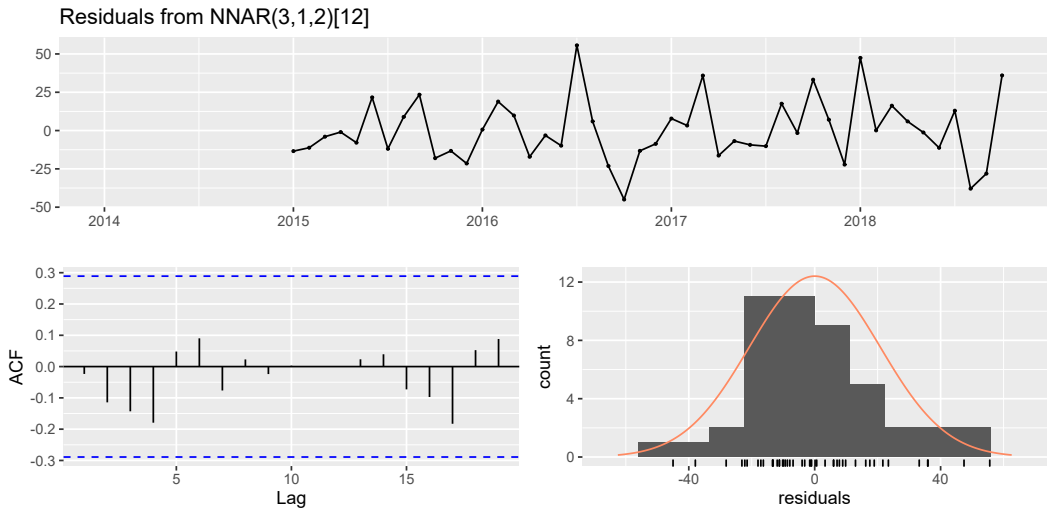


Table 4A: NNAR(3,1,2)[12]

Call:

Series: MONTLY_TS_TRAINING_SET

Model: NNAR(3,1,2)[12]

Call: nnetar(y = MONTLY_TS_TRAINING_SET, lambda = "auto")

Average of 20 networks, each of which is

a 4-2-1 network with 13 weights

options were - linear output units

sigma^2 estimated as 263.5

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.4175588	21.67883	16.63400	-0.2456791	4.043607	0.4535650	-0.05721551	NA
Test set	5.8703121	34.37212	26.86929	0.6581365	5.419174	0.7326541	-0.06010674	0.6110487

Table 5A: Box-Ljung test

data: test_res
X-squared = 7.6537, df = 11, p-value = 0.7439

Figure 12: Residual diagnostics

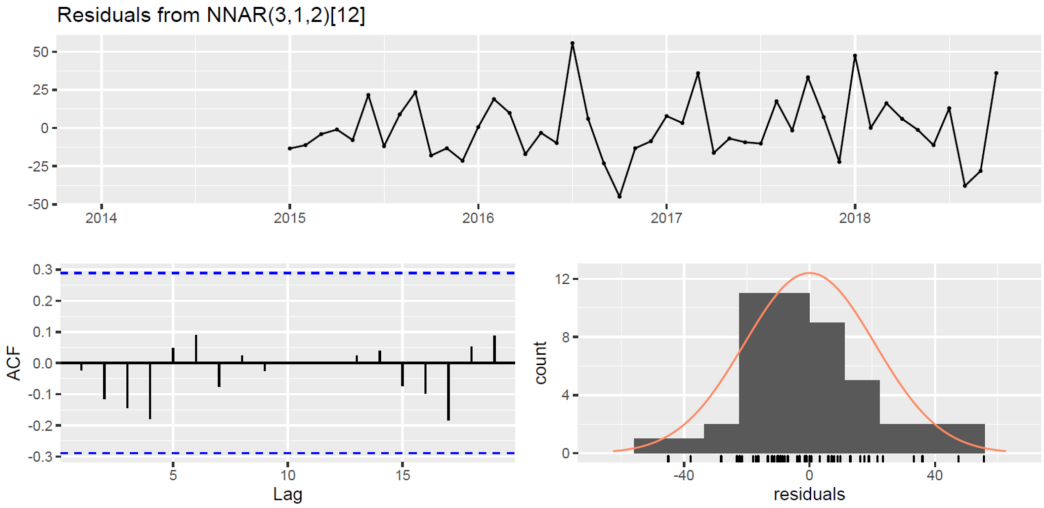


Table 6A: Ensemble model

Forecast method: auto.arima with weight 0.444

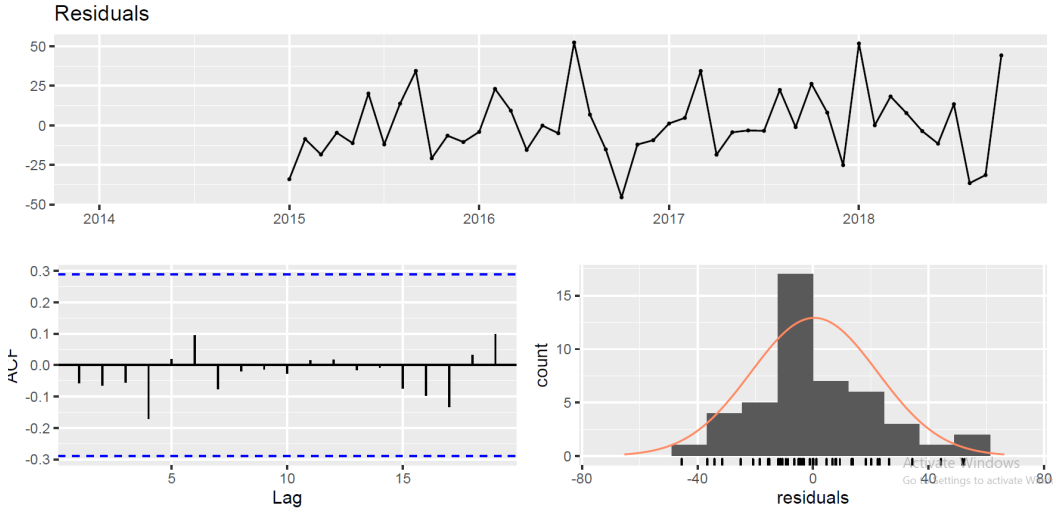
Forecast method: nnetar with weight 0.556

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.3780783	21.29728	16.40109	-0.2507778	4.002343	0.4472140	-0.03853283	NA
Test set	0.6781094	31.22528	24.72169	-0.3442233	5.055429	0.6740947	-0.03640456	0.5478217

Table 7A: Ensemble model

data: test_res2
X-squared = 2.9966, df = 11, p-value = 0.9908

Figure 13: Residual diagnostics



Jordan Simonov

Jordan Simonov has worked at the North Macedonian Ministry of Finance since 2002. Mr Simonov completed a bachelor degree in customs and freight forwarding and a master degree in European Studies during which he defended his master thesis, *Harmonising Customs legislation of the Republic of Macedonia with acquis Communautaire and future challenges of full membership in the European Union*. Currently Mr Simonov is Head of the Forecasting and Analysis Unit, where he analyses and forecasts tax and Customs revenue.

Zoran Gligorov

Zoran Gligorov has worked at the North Macedonian Ministry of Finance since 2000. Mr Gligorov completed a bachelor degree in financial management and has a master degree in economics entitled *Customs strategy of EU in accordance with customs system of the Republic of Macedonia and further excise system policy*. Currently Mr Gligorov is the Assistant Head of Department where he works on tax and customs legislation.