“Forecasting of individual’s deposits volume of Sberbank with respect to weighted rates and macroeconomic indicators”

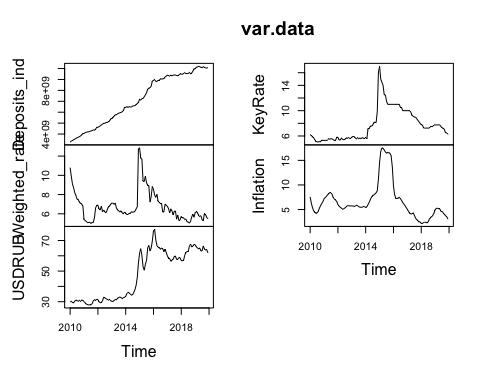
CosmoGalya

3/12/2020

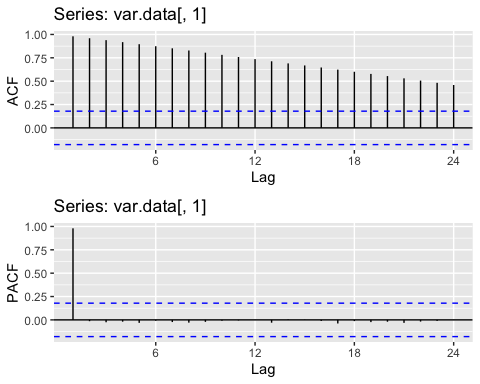
Theme of our forecasting project is “Forecasting of individual’s deposits volume of Sberbank with respect to weighted rates and macroeconomic indicators”. Firstly, we have to explain the terminology. Under weighted rates state for weighted deposit rates in Russian banking industry without Sberbank ones.

### Data description and preparation procedures

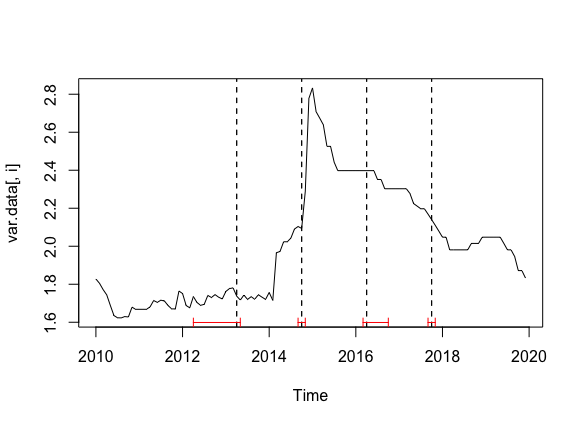
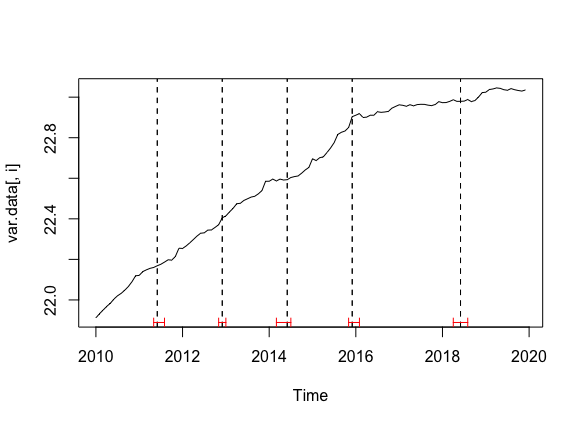
Our data is presented with volume of Sberbank deposits in rubles and macroeconomic indicators as inflation, key rate and USD/RUB exchange rate. On the graph below our data is presented ASIS.



The non-stationarity of our time series (TS) and structural break in 2015 can be easily seen from the visualization above. ACF and PACF were constructed only for Deposits data, other variables’ ACF and PACF can be found in appendix.



The main structural break we have to be careful about was in December, 2014 due to the change in central bank monetary policy: transition to float free-float exchange rate and inflation targeting policy.



In order to overcome problems in raw data, several test and procedures like Box-Cox procedure were provided to transform the data. According to the Box-Cox test results our data was transformed in following way: weighted, key rates and deposits: replaced with thier logarithms, USD/RUB with its inverse square root and inflation as it is. The estimation of Box-Cox lambda you may find in chart 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Deposits\_ind | Weighted\_rate | USDRUB | KeyRate | Inflation |
| 0.0332 | -0.1469 | -0.4828 | -0.9999 | 0.8182 |

We use auto.arima function in R to automatically select ARIMA model specification. The number of unit roots in the model is automatically checked using standard ADF test. As we cannot be sure for 100% in our data-stationarity from visual analysis (ACF, PACF analysis), ADF-tests are presented on chart 2. Even with ADF test you cannot be sure in anything, but anyway. :)

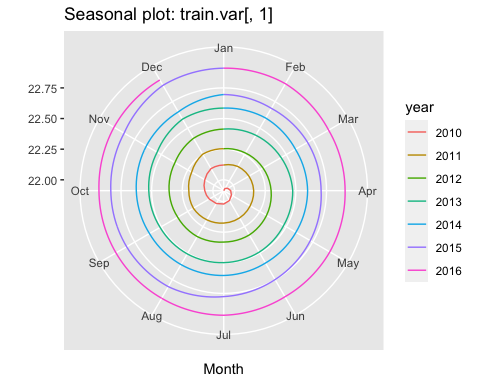
Augmented Dickey-Fuller Test

data: diff\_log\_dep Dickey-Fuller = -4.0949, Lag order = 4, p-value = 0.01 alternative hypothesis: stationary

Augmented Dickey-Fuller Test: diff\_log\_dep

|  |  |  |  |
| --- | --- | --- | --- |
| Test statistic | Lag order | P value | Alternative hypothesis |
| -4.095 | 4 | 0.01 \* \* | stationary |

Then our data was splited into train and test in the following manner: train starts from Jan, 2010 and ends in Dec, 2016; test starts Jan, 2017 and ends in Dec, 2019. We defined our forecast-horizon to be equal to 12 months, or 365 days, or 8760 hours, or 525,6 kmin, or 31,536 msec using SI-system. As a result of strucchange R package searches it was decided to add a dummy-variable to our analysis in order to solve the problem of structural break in Dec, 2014 without data elimination (unfortunately, there is lack of data from the very beginning: ~~need more data to God of Data~~). Then seasonality of our data was checked – no seasonality was found, but X-13 ARIMA has found it. We believe to X-13 ARIMA, because it was developed by US Census Bureau and Bank of Spain, they are cool guys.

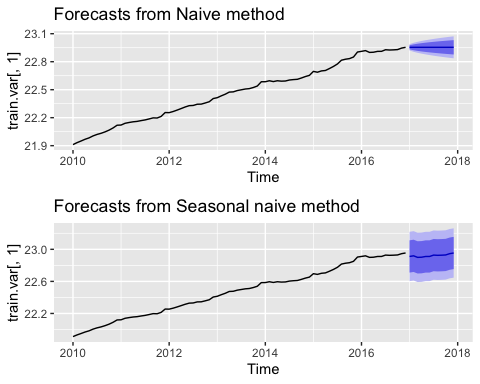


### Time series analysis

Time series analysis consists of two global parts: Naïve forecasting and Model based forecasting.

#### Naïve forecasting

Firstly, we calculated naïve and snaïve forecasts that gave us following graphs below:



The core idea of those methods is that tomorrow’s value is equal to today’s value. This tendency is being kept for a long-run as there is no adjustment to external factors or whole dynamics of TS in this methodology. Naïve method usually used for comparison with more complex forecasting methods as a benchmark. Seasonal naïve method has similar core idea but it is used for highly seasonal data.

Naïve forecasting method:

Seasonal naïve forecasting method:

s - seasonal period, k - seasonal frequency coef.

To compare out-of-sample forecasting accuracy we use RMSE measure.

Table continues below

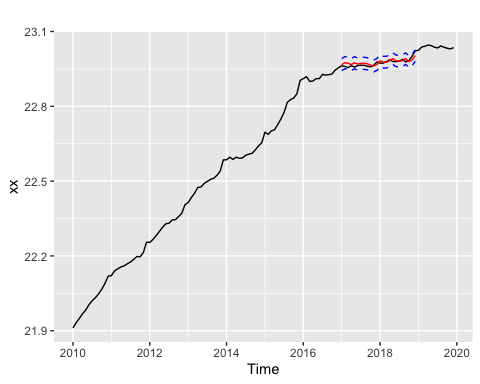
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | ME | | RMSE | MAE | MPE | MAPE | MASE |
| **Training set** | 0.01256 | | 0.01733 | 0.0137 | 0.05598 | 0.06096 | 0.09017 |
|  | ME | RMSE | MAE | MPE | MAPE | MASE |
| **Training set** | 0.1519 | 0.1569 | 0.1519 | 0.6738 | 0.6738 | 1 |

### Models

Predicting accuracy of proposed models was compared with a help of RMSE. We estimated ARIMA, ETS, ARIMAX, VAR, VAR with restriction (LASSO), VECM and RWD as benchmark. RWD is a variation of naïve forecasting that includes drift as a average historical change during the data horison. Visually it can be seen a tendency to growth in deposit value data.

#### ETS

ETS model is Holt’s exponential smoothing method, that decomposes time series into trend, seasonal and error components. There are several model specifications, depending on the form of each of these components. We use automatical specification search each time we estimate the ETS model. We do understand that for each train sample the model specification could be different. It means that each forecast made on cross-validation could be generated by different ETS specifications.

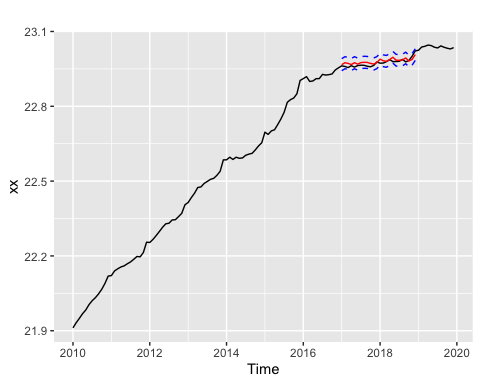


In the majority of samples (share is ) the model specification is “[MAN](https://music.yandex.ru/album/719677/track/113665)”, which means multiplicative error, additive trend and no seasonal component. The aletrnative specification (which was selected in of cases) is “MAdN”, where “Ad” stands for additive damped trend component.

#### ARIMA

ARIMA also called Box-Jenkins model is one of the classical models in applied Time Series Econometrics that explains the dynamics with lag of the variable itself and lag of error terms.

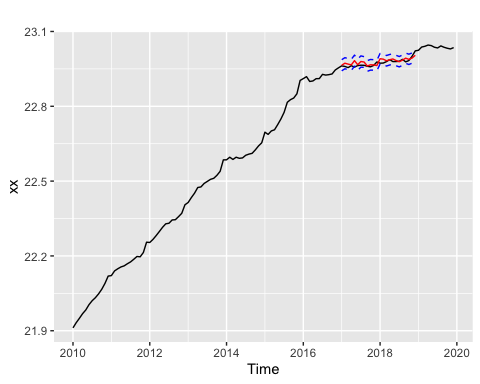
In our case only on train data model has initially (0,1,0) specification that changed to (2,1,0) and (1,1,0) though cross validation process. (I know it is rather strange spec at the first step. It’s AIC failure, BIC gave worse results.) The Visualization of the forecasting values and real values can be seen below:



ARIMA and ETS use lags only of dependent variable while the next bunch of models use not only lags of predicted variables but also lags of other external variables that can affect regressant.

#### ARIMAX

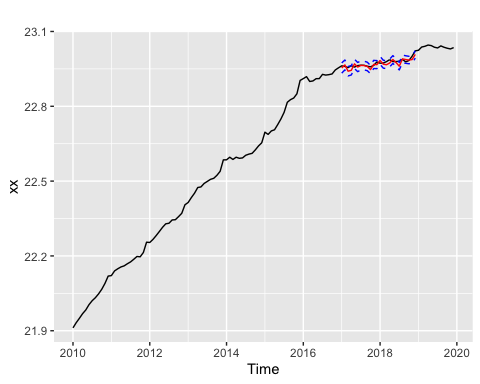
ARIMAX is a model that is characterized by forecasting throught not only with the help of dependant variable but also with lags of other variables that may affect the forecast. In our case the forecasts were calculated throught same specification as ARIMA in the first step and transform to (4,1,0) spec at the last step. In the model were used lags of weighted rate, USDRUB, Key rate and Inflation. We took 1 lag for each variable due to AIC (VARselect procedure). The graph is presented below.



#### VAR

Vector autoregerssion of order p (VAR(p) ) models are multivariate dynamic models. In this class of models the whole system of equations is estimated:

where . is the vector of endogenous variables. In our case it consists of log deposits, weighted interest rate, US Dollars to ruble exchange rate, key interest rate and inflation.



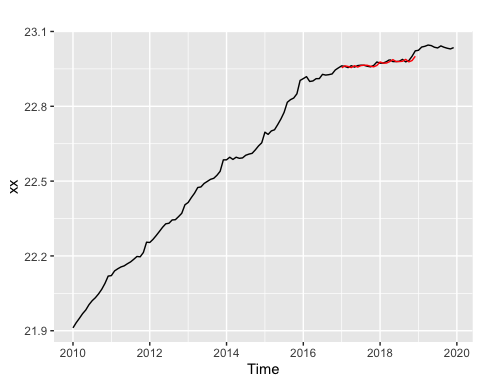
Contrary to ARIMAX models, VAR allows to predict the whole set of endogenous variables at once. It means that VAR forecasts are constructed conditional only on the previous values of endogenous variables, but not conditional on scenario values.

We use Akaike information criterion (AIC) to choose the lag length for VAR model. The maximum possible lag is , due to monthly frequency of the data. We automatically select lag length in each sample. In all samples, except the last one, selected lag length equals . On the final sample lag length is selected to be . In our forecasting exercise we calculate RMSE only for log deposit variable.

#### VAR restricted

In the standard VAR specification there are endogenous variables and lags selected. It means that there are parameters in each equation of the VAR system. Due to the lack of data this fact could lead to the model overfit. To handle this problem we regularize VAR using LASSO-type regularization rules. To do this we implement “Basic VARX-L” structure, provided by package.

LASSO regularization shifts parameters to zero values, depending on the penalty value. The higher the penalty, the lower parameters are in absolute values. To choose the penalty value, we use cross-validation on the rolling window with forecasting horizon equal to .

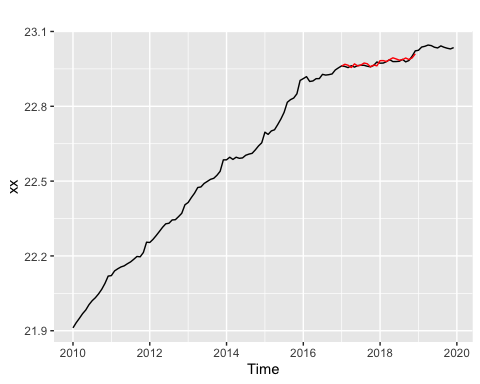


All variables, used in VAR, except key interest rate, are non-stationary. To take this fact into account, we used Minnesota prior, which shrinks coefficients towards random walk case (see Demeshev and Malakhovskaya, (2016)). Without Minnesota prior forecasts were **very** bad.

#### VECM

Vector error correction model has the following specification:

Linear combinations of endogenous variables $ Y\_{t-1} $ are cointegration relations that drive first differences of endogenous variables. To choose the number of coiintegration relations, we applied trace Johansen test without constant and trend. This is due to the fact that the representation of original series is VAR model with constant.



To make forecasts of the original series , this model should be replaced with corresponding VAR(p) representation. The inclusion of integration relations could shift coefficients. It means that forecasts could differ from the standard VAR model. We use procedures of the package to calculate forecasts.

VAR and ARIMAX use in addition other macroindicators. We calculated ARIMAX model forecasts conditional on the actual values of exogenous variables in the test sample. VAR forecasts were made by forecasting all of the endogenous variables. However, RMSE for VAR model was calculated only for log deposits. We also need to check causality before constructing VAR models.

##   
## Granger causality H0: Weighted\_rate USDRUB KeyRate Inflation do not  
## Granger-cause Deposits\_ind  
##   
## data: VAR object m.var  
## F-Test = 3.3112, df1 = 48, df2 = 170, p-value = 5.9e-09

To check forecasting power we used RMSE measure (RMSE results presented in chart below). To make forecasting exercise we split the whole sample into train and test samples in the following manner. We first estimate the proposed models on the sample till the December, 2014 and calculate forecasts on horizon 12 month, starting from the January, 2015. We then sequentially add new points with 1-month step to the train sample and shift the test sample by 1 month to the right. We then reestimate models and calculate out-of-sample forecasts for a new train sample.

Using RMSE-ratio (RMSE of the model/RMSE of the random walk benchmark) we concluded that VECM, VAR restricted and ETS as best prediction models.VAR and VECM are the best models for forecasting as it used other variables that affects directly the volume of the deposits. Restrictions that are included in those models increase the accuracy due to some problems like multicollinearity partially solved with its help.

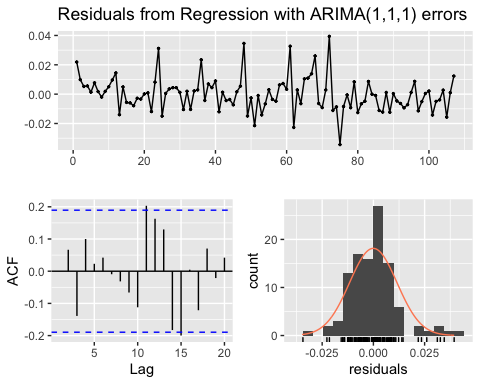
RMSE:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| arima | ets | arimax | seas | var | var\_lasso | vecm | rwd |
| 0.01108 | 0.009939 | 0.01126 | 0.01037 | 0.01029 | 0.008146 | 0.008696 | 0.01154 |
| 0.01997 | 0.01726 | 0.01828 | 0.01908 | 0.01466 | 0.01238 | 0.01357 | 0.02079 |
| 0.02822 | 0.02392 | 0.0252 | 0.02731 | 0.01982 | 0.01622 | 0.01839 | 0.02927 |
| 0.03596 | 0.02998 | 0.03256 | 0.03566 | 0.02235 | 0.01986 | 0.02284 | 0.03726 |
| 0.04475 | 0.03725 | 0.04084 | 0.04408 | 0.02684 | 0.02425 | 0.0287 | 0.04582 |
| 0.05258 | 0.04322 | 0.04804 | 0.05227 | 0.03256 | 0.02778 | 0.0337 | 0.05366 |
| 0.0601 | 0.04858 | 0.05582 | 0.06074 | 0.03894 | 0.03036 | 0.03883 | 0.06166 |
| 0.06706 | 0.05351 | 0.06329 | 0.06928 | 0.04467 | 0.0323 | 0.04385 | 0.06957 |
| 0.07416 | 0.05895 | 0.0713 | 0.07743 | 0.0534 | 0.03512 | 0.0498 | 0.0775 |
| 0.08093 | 0.06361 | 0.07957 | 0.08561 | 0.06302 | 0.03756 | 0.05576 | 0.08526 |
| 0.08741 | 0.06756 | 0.08766 | 0.09363 | 0.07065 | 0.03971 | 0.06181 | 0.09292 |
| 0.09411 | 0.07191 | 0.09588 | 0.102 | 0.08375 | 0.0417 | 0.06839 | 0.101 |

RMSE RATIO:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| arima | ets | arimax | seas | var | var\_lasso | vecm |
| 0.9605 | 0.8613 | 0.9757 | 0.8988 | 0.8913 | 0.7058 | 0.7535 |
| 0.9604 | 0.8302 | 0.879 | 0.9175 | 0.7051 | 0.5952 | 0.6525 |
| 0.9641 | 0.8172 | 0.861 | 0.9328 | 0.677 | 0.5539 | 0.6281 |
| 0.9651 | 0.8048 | 0.8739 | 0.9572 | 0.5999 | 0.5332 | 0.613 |
| 0.9767 | 0.813 | 0.8915 | 0.9622 | 0.5858 | 0.5294 | 0.6264 |
| 0.9799 | 0.8054 | 0.8953 | 0.9741 | 0.6068 | 0.5178 | 0.6281 |
| 0.9748 | 0.7879 | 0.9054 | 0.9851 | 0.6316 | 0.4924 | 0.6297 |
| 0.9639 | 0.7692 | 0.9097 | 0.9958 | 0.642 | 0.4642 | 0.6303 |
| 0.9569 | 0.7606 | 0.92 | 0.9991 | 0.6891 | 0.4532 | 0.6426 |
| 0.9491 | 0.746 | 0.9332 | 1.004 | 0.7391 | 0.4406 | 0.654 |
| 0.9408 | 0.7271 | 0.9434 | 1.008 | 0.7604 | 0.4273 | 0.6652 |
| 0.932 | 0.7122 | 0.9495 | 1.01 | 0.8293 | 0.4129 | 0.6773 |

We have to check residuals for all models if it possible :)



##   
## Ljung-Box test  
##   
## data: Residuals from Regression with ARIMA(1,1,1) errors  
## Q\* = 6.2143, df = 6, p-value = 0.3996  
##   
## Model df: 4. Total lags used: 10

To see more tests please look to the appendix.

### References

1. [The best book about forecasting ever](https://otexts.com/fpp3/x11.html)
2. [Демешев Б.Б. Малаховская О.А. МАКРОЭКОНОМИЧЕСКОЕ ПРОГНОЗИРОВАНИЕ С ПОМОЩЬЮ BVAR ЛИТТЕРМАНА](https://cyberleninka.ru/article/n/makroekonomicheskoe-prognozirovanie-s-pomoschyu-bvar-littermana/viewer)

### Appendix with sweet graphs

All graphs, code, rmd and some other things that makes [~~me cry for a week~~](https://music.yandex.ru/album/7475804) this work full you may find on GitHub public repository [following this link](https://github.com/Galunay/forecasting) :)

P.S. I miss so much the faculty of economic science, sweet Poisson and room 2112 :c