Ukrainian Catholic University
Applied Sciences Faculty

Econometrics project

russian Invasion Forecasting

Author:
Markiian Tsalyk

Ukrainian Catholic University
APPLIED
SCIENCES
FACULTY

1 June 2022

Abstract

24 February 2022 was the date when the lives of all Ukrainian people have been changed rapidly. We woke up not from an annoying alarm clock as usual, but from an air alarm. That was the stunning news, russia started the bloody war. This research was incentivized by the willingness to support Ukraine in the information war and the efforts of the digital department of Ukraine to provide as clean and unbiased data as possible.

Introduction

We have data that can be explored and analyzed. It is possible to extract some patterns from this data and do research. In this work, we will describe the analytical approach to reach such patterns and construct forecasting models with ARIMA. This modeling helps us to monitor situations, observe anomalies and predict in which direction the war will develop. It is useful as well because it can show, that data indeed is unbiased, therefore it is not Ukrainian propaganda, that's what actually is going on nowadays.

Literature

In order to perform a time series analysis, it was necessary to learn the aspects of models which can be applied and math behind them. Wooldridge's "Introduction to Econometrics" is extremely helpful here.

Also, there is some documentation about python libraries that are useful and articles about Time Series modeling. All the links are attached below.

[1] Basic Regression Analysis with Time Series
[2] Time Series with ARIMA model
[3] Time Series analysis tsa
[4] ARIMA

Data description and analysis

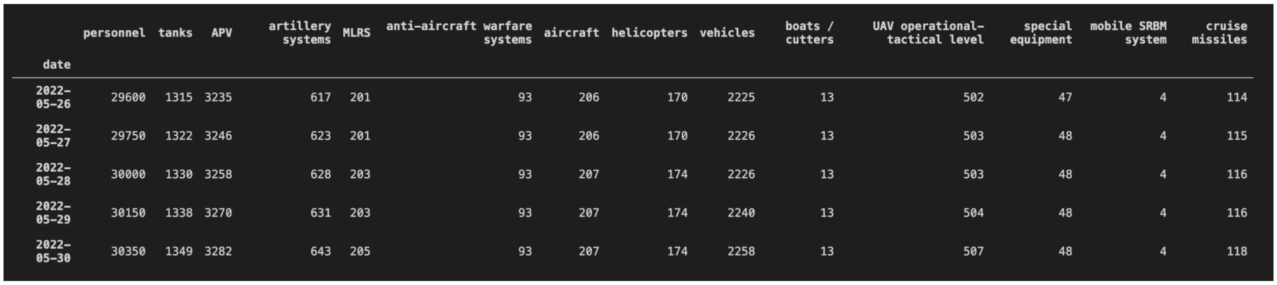
The source of the data is the Facebook page of the General Staff of the Armed Forces of Ukraine: https://www.facebook.com/GeneralStaff.ua. Losses include daily retrieving data:

- personnel
- tanks

- APV
- MLRS
- artillery systems
- anti-aircraft warfare systems
- aircraft
- helicopters
- UAV operational-tactical level
- cruise missiles
- boats/cutters
- special equipment
- mobile SRBM system
- vehicles and fuel tanks

In the dataset, we currently have *97 observations*, since the war is going on 97-th day. Also, the dataset requires data cleaning and feature engineering, because in first few weeks information was approximate, therefore appropriately marked as approximate.

Each entry is an integer type and represents the total losses of russian army till the given day (index of the data frame as it is shown in Figure 1).



	personnel	tanks	APV	artillery systems	MLRS	anti-aircraft warfare systems	aircraft	helicopters	vehicles	boats / cutters	UAV operational-tactical level	special equipment	mobile SRBM system	cruise missiles
date														
2022-05-26	29600	1315	3235	617	201	93	206	170	2225	13	502	47	4	114
2022-05-27	29750	1322	3246	623	201	93	206	170	2226	13	503	48	4	115
2022-05-28	30000	1330	3258	628	203	93	207	174	2226	13	503	48	4	116
2022-05-29	30150	1338	3270	631	203	93	207	174	2240	13	504	48	4	116
2022-05-30	30350	1349	3282	643	205	93	207	174	2258	13	507	48	4	118

Figure 1

Methodology

It is important to research what factors influence *personnel losses* the most. To do this, we construct the linear model using **OLS**. After obtaining results we can choose variables for forecasting.

Basically, personnel losses is our *dependent variable*, and technique with types of equipment are *independent* variables.

It is assumed that personnel losses are the most valuable for russian government, therefore it will be our task to choose features that strongly correlate with the death of soldiers. It is easy to do, checking whether some variables are statistically significant for our OLS modeling. Forecasting deaths of russian soldiers will be developed as well.

ARIMA model is a great choice for such type of analysis. ARIMA – autoregressive integrated moving average model. It forecasts future values, based on previous ones. Such type of forecasting is useful in our case because decisions on future war actions usually are made with the experience of the past actions. To construct the ARIMA model we need to know 3 important parameters: ρ, d, q.

d is the number of differences between observations we need to compute. It can be chosen using the **Augmented Dickey-Fuller** test. Based on the ρ-value we can take the appropriate value for d.

q is the number of lags crossing the autocorrelation threshold.

ρ is the most significant lag of the autocorrelation.

Since standard deviations σ of technique and types of equipment are much smaller than the standard deviation σ of personnel losses, we can try to make a pipeline with 2 models. At first, we forecast losses of independent variables and then, based on this forecasting, construct a linear regression with deaths of soldiers being a dependent variable.

Results

First of all, we plotted the accumulated growth of each loss and saw that in first 2 weeks incredible power was forced to capture the country. Probably, a kind of blitzkrieg was planned (a part of the plots is given in Figure 2).

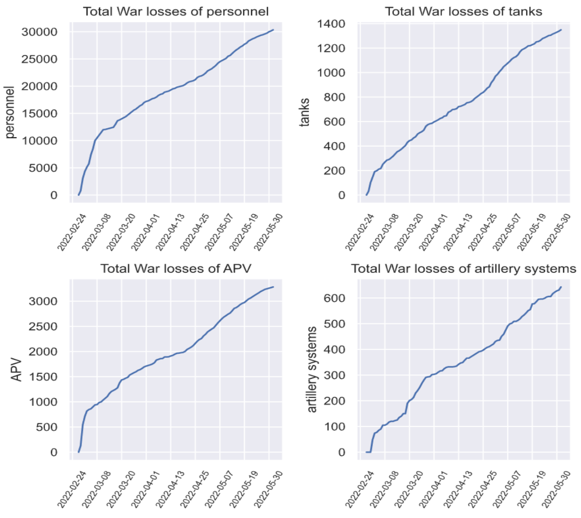


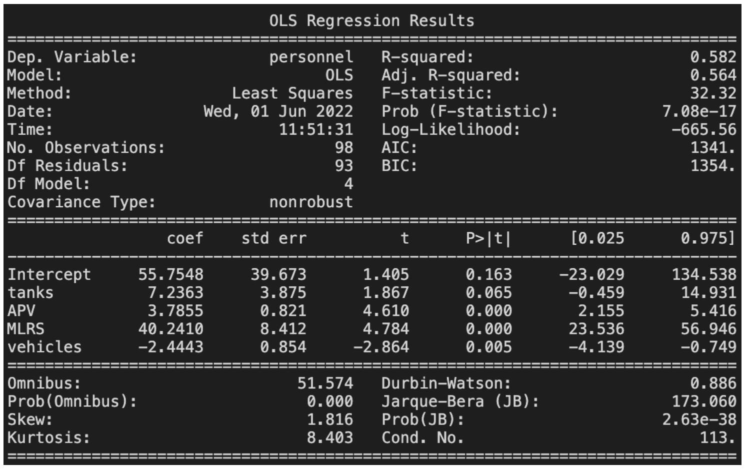
Figure 2

We have observed that tanks, APV, MLRS, and vehicles (significance **P > |t|** is given in Figure 3) make the largest impact on personnel losses, therefore there are reasons to create forecasting models of those features.

Also, here we see, that **R-squared** is *0.582*, which means that *58.2%* changes in personnel losses can be explained with our model.

The probability of the **F-statistic** is low (approximately 0), which implies that all the coefficients are not 0.

Omnibus indicates the normality of distribution of personnel losses. The larger it is, the lower probability that the data is normally distributed. This probability is given below the omnibus, it is 0 in our case, which means that the distribution isn't normal.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              personnel   R-squared:                       0.582
Model:                            OLS   Adj. R-squared:                  0.564
Method:                 Least Squares   F-statistic:                     32.32
Date:                Wed, 01 Jun 2022   Prob (F-statistic):           7.08e-17
Time:                        11:51:31   Log-Likelihood:                -665.56
No. Observations:                  98   AIC:                             1341.
Df Residuals:                      93   BIC:                             1354.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     55.7548     39.673      1.405      0.163     -23.029     134.538
tanks          7.2363      3.875      1.867      0.065      -0.459      14.931
APV            3.7855      0.821      4.610      0.000       2.155       5.416
MLRS          40.2410      8.412      4.784      0.000      23.536      56.946
vehicles      -2.4443      0.854     -2.864      0.005      -4.139      -0.749
==============================================================================
Omnibus:                       51.574   Durbin-Watson:                   0.886
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              173.060
Skew:                           1.816   Prob(JB):                     2.63e-38
Kurtosis:                       8.403   Cond. No.                         113.
==============================================================================
```

Figure 3

Next, we conducted hypothesis testing. The question that is interesting for us is whether the intensity of war actions during last 2 months (April-May) is lower compared to the beginning (February-March). Let μ_1 be the intensity of the beginning, then μ_2 is the intensity of the last months::

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

```
ttest, pvalue = stats.ttest_ind(losses_feb_mar['intensity'], losses_apr_may['intensity'])
pvalue /= 2

if pvalue > 0.05:
    print('H_0 holds')
elif ttest > 0:
    print('mu_1 > mu_2')
else:
    print('mu_1 < mu_2')

print(f't-statistics, p-value = ({ttest}, {pvalue})')
```

```
mu_1 > mu_2
t-statistics, p-value = (3.514540352723546, 0.00034125776053321657)
```

Thus, under significance level $\alpha = 0.05$ we reject the null hypothesis. Since t-statistics is larger than 0, we can conclude that $\mu_1 - \mu_2 > 0$ and then $\mu_1 > \mu_2$.

After all, we constructed the ARIMA model for personnel losses. To choose parameters, we plotted autocorrelation (Figure 4) and conducted the **Augmented Dickey-Fuller** test.

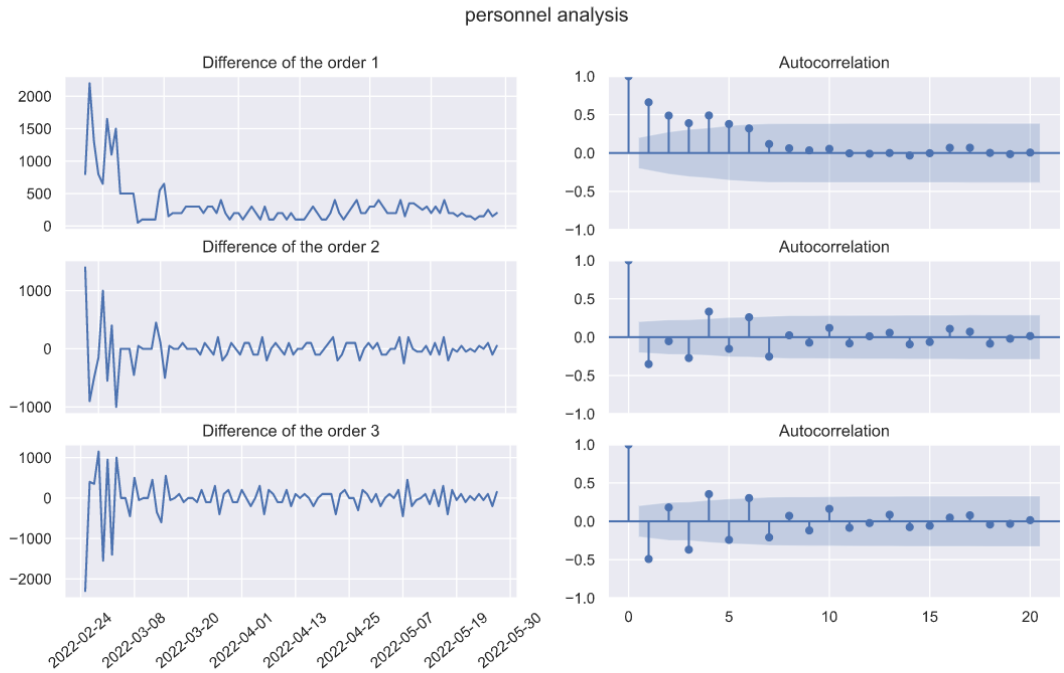


Figure 4

We ended up with **ARIMA(p=1, d=2, q=3)**. In the same way, we created models to forecast technique losses, which are significant for personnel losses. Based on this forecasting, we developed a linear regression for dead soldiers. In Figure 5 depicted forecastings for the **next month** total personnel losses with both ARIMA and Linear Regression models.

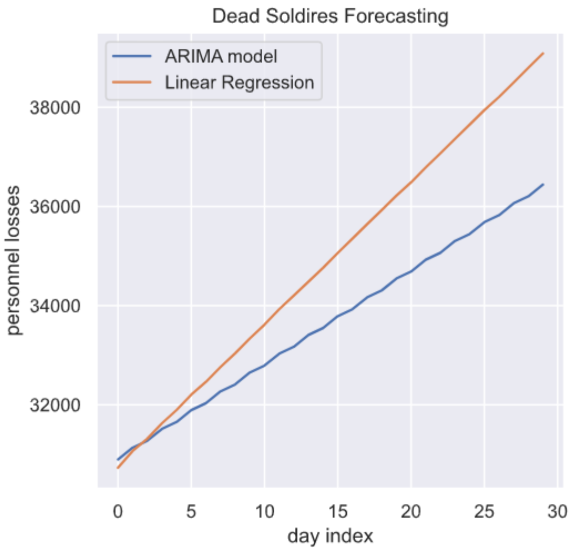


Figure 5

Limitations

One limitation is that we can't actually use a linear model straight forward to predict personnel losses because the data about the number of destroyed types of equipment won't be available, as it updates simultaneously. However, we can still use the **ARIMA** model, since it is the *Time Series* approach, and we actually know that all losses of our enemy are strongly *dependent on time (past observations).*

Next steps

The project can be developed and improved in the future. What is incredibly useful to do is to include features about a number of the weapon of each type given to the Ukrainian army in some periods of time. In such a way we can observe the dependency between this weapon and russian losses. Thus, we will have strong pieces of evidence to request the weapon we need for the world and consequently defeat our enemy.

Project on GitHub

https://github.com/Tsalyk/russianInvasionAnalysis

Conclusion

In this research, we conducted the **EDA** of the russian invasion of Ukraine.
Later we have determined the types of equipment that influence the most losses of russian personnel. We found out that these are ...

As well, we test the hypothesis about the intensity of the war actions compared to *February-March* and *April-May*. We figured out that at the beginning of the war, as was intuitively expected, the intensity was bigger.

We developed forecasting models for the deaths of soldiers and types of equipment: tanks, APV, MLRS, and vehicles.

Finally, we combined the results of ARIMA forecasting for technique losses and used obtained data to create Linear Regression with personnel losses being a dependent variable.