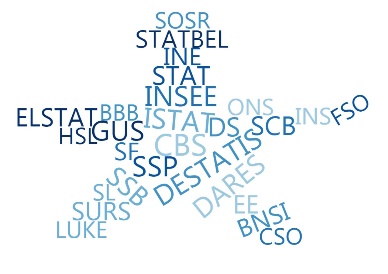
 **DRAFT, Version 24/04/24** 

**ESSnet Big Data II**

**Grant Agreement Number:**

**NUMBER — 847375 — 2018-NL-BIGDATA**

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata>

[<https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en>](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)

**Workpackage WPB**

**Implementation – Online Job Vacancies**

**An example in estimation of GDP components using OJA**

**(April 2020)**

|  |
| --- |
| **Prepared by: Črt Grahonja (SURS, Slovenia)**  **Date of present version: April 2020** |

Workpackage Leader:

Tomaž Špeh (SURS, SI)

e-mail address: tomaz.speh@gov.si

mobile phone: +38651672116

An example in estimation of GDP components using OJA

**Contents**

[1. Technical considerations 3](#_Toc39751803)

[2. Data description 3](#_Toc39751804)

[2.1 Job portals data 3](#_Toc39751805)

[2.2 Turnover from the sale of services: VAT and Indices 4](#_Toc39751806)

[3. Researching models 5](#_Toc39751807)

[3.1 Linear regression 7](#_Toc39751808)

[3.2 ARIMAX 7](#_Toc39751809)

[3.3 Lagged MD data fitted to the Index 9](#_Toc39751810)

[4. Conclusion 10](#_Toc39751811)

**List of figures**

[*Figure 1: The OJAs series* 3](#_Toc39751812)

[*Figure 2: The Index and VAT series; the scale's color on each side corresponds to the line color.* 4](#_Toc39751813)

[*Figure 3: VAT and OJA series* 5](#_Toc39751814)

[*Figure 4: Index and OJA series* 6](#_Toc39751815)

[*Figure 5: Index estimations and realizations using a linear regression* 7](#_Toc39751816)

[*Figure 6: the ACF and Partial ACF graphs for the Index series* 8](#_Toc39751817)

[*Figure 7: Index estimations and realizations using a ARMAX model* 9](#_Toc39751818)

[*Figure 8: Linear regression and ARIMAX estimations with lagged MD data* 10](#_Toc39751819)

**List of tables**

[*Table 1: Correlations between economic indicators and OJAs* 5](#_Toc39751820)

[*Table 2: Correlation of the Index data and lagged OJAs* 6](#_Toc39751821)

An example in estimation of GDP components using OJA

# Technical considerations

This reports aims to explore a viable use of job vacancy advertisements (OJA, *also JVA – job vacancy advertisements*) data to estimate two statistics related to the national accounts field: the VAT from the turnover from the sale of services (hereof called VAT) and the index of the turnover from the sale of services (hereof called Index).

The analyses have been conducted during the penultimate week of April 2020 (20th to 24th). They refer to 2 sets of scraped data and a component in the calculation of Slovenian’s GDP, both prepared by the Slovenian Office. The data cover the period of January 2017- February 2020, i.e. a period of 38 months. Additionally, two more months of data are available for OJA data. Statistical analyses and modelling have been conducted using the open source software R.

# Data description

## Job portals data

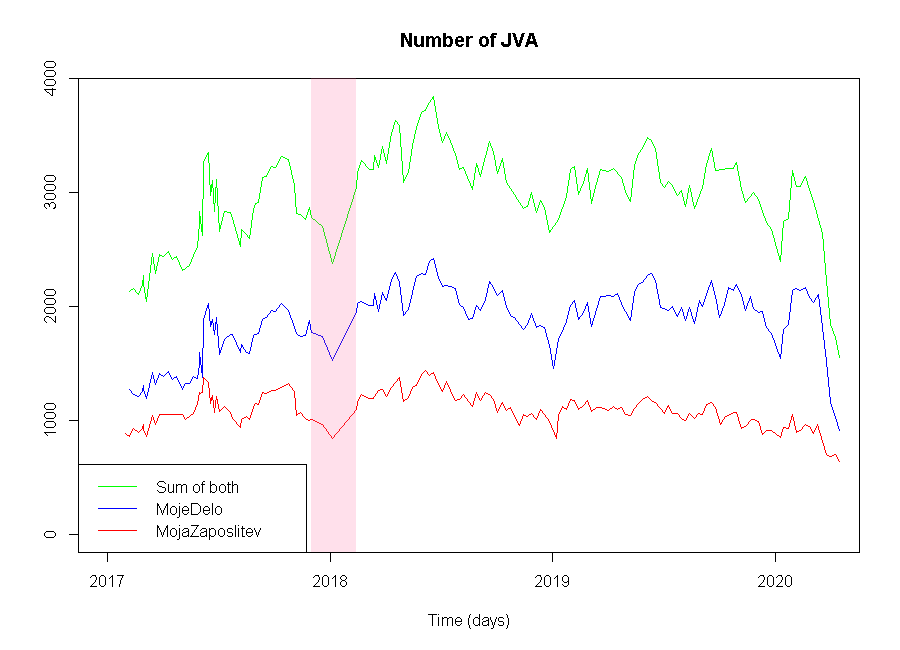


Figure 1: The OJAs series

The two main Slovenian job portals in Slovenia are MojeDelo.si (hereof MD) and MojaZaposlitev.si (hereof MZ). They are usually scraped every Tuesday since February 2017. However, due to the nature of internet sometimes scraping fails, which means unscraped periods have to be rescraped on a different day (usually the next one). This explains the mismatch in time points in some weeks. Furthermore, a period of erratic scraping occurred around late 2017 and early 2018, which was the result of personnel change in the office and other aggravating factors related to third party scrapers’ shortcomings. Between December and February only 4 observations were recorded, with at least one scrape a month. Due to the change in staff the method of scraping is different between the two periods (pre- and post-2018, respectively), but the scraped content is the same in both. The time series for both and their sum can be seen in the above figure (Figure 1), as well as the period with loss of data in pink.

Targets of scraping are whole lists of job advertisements present on each job portal, which include a lot of extra information that might be useful in the future. These data are semi-structured, with a lot of fields easily classified and an unstructured job description. However, in this analysis we focused only on the total number of OJAs. In general, MZ data represent a third, while MD data represent the other two thirds of all scraped data. There is a clear yearly seasonality with peaks in the summer and has the lowest point round around New Year. Up until the last few months, the number of OJAs was quite stable; however from mid-February 2020 there is a clear and uncharacteristic trend downwards, which can probably be attributed to the effect of the COVID-19 virus. It must be explicitly stated, that this is not a scraping error, the number of OJAs on both pages have declined steeply.

## Turnover from the sale of services: VAT and Indices

The target variables are the VAT value from the turnover from the sale of services and its index.

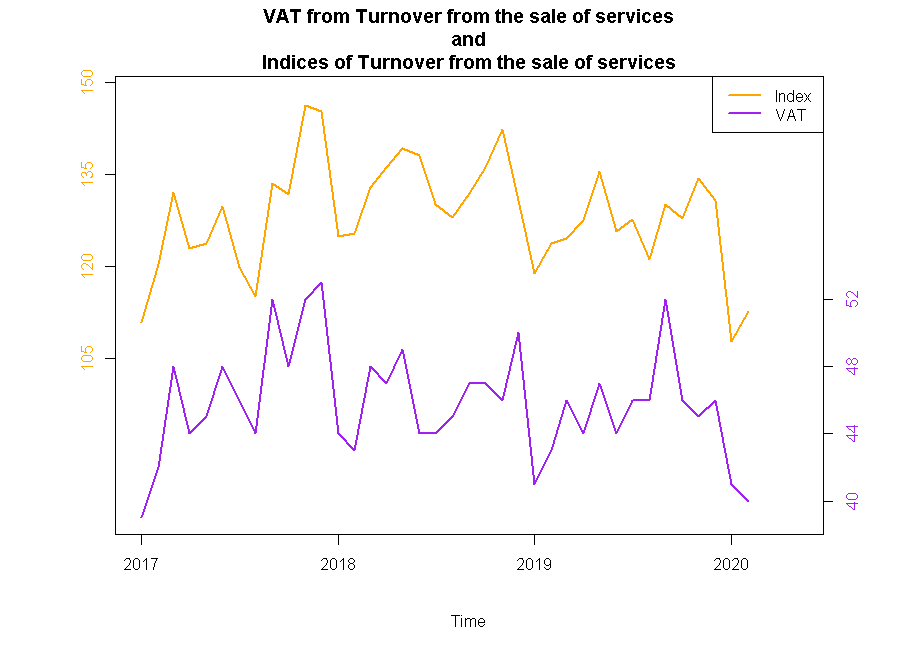


Figure 2: The Index and VAT series; the scale's color on each side corresponds to the line color.

These statistics are used in the calculation of the country’s GDP by the National Accounts department. They are prepared every month with a rather long delay, which initially prompted our analysis.

# Researching models

We have tried to transform OJA data in a few different ways. First we analyzed visually the daily periods and monthly data (Figure 3 and Figure 4). The next step was to sum weekly data for every month and use the sums as a monthly series. This was discarded in favor of an average of weekly data on a monthly basis, to preserve the scale of observations between shorter and longer months and to account for the erratic scraping period.

All series were split into a training part (January 2017-December2019) and testing part (January 2020 and onward). This means that the training part consists of 36 data points and the test part 2 data points for economic indicators and 4 points for OJA data. Here already an issue with the shortness of our data is apparent. Due to the low number of observations, we had to include the uncharacteristic fall into the test data, otherwise we would not have enough test periods.

Consider the correlation values between both training series:

|  |  |  |  |
| --- | --- | --- | --- |
|  | MojeDelo OJA | MojaZaposlitev OJA | Sum of both |
| VAT | 0.094 | 0.113 | 0.108 |
| Index | 0.346 | 0.261 | 0.347 |

Table 1: Correlations between economic indicators and OJAs

There seems to be correlation present between our series, however they are much more pronounced with the Index. This is why from this point on we only worked on estimating the Index series.

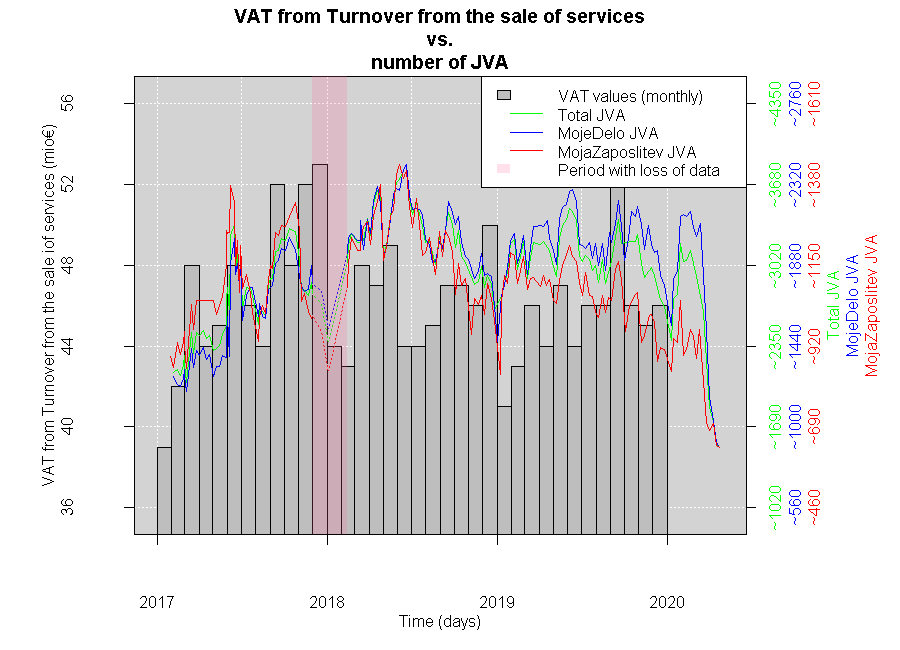


Figure 3: VAT and OJA series

In the next chapters, we will show our results for a linear regression model and an ARIMAX model that estimates an economic indicator with help from OJA data. Above (Figure 3) the VAT series with OJAs superimposed is plotted. On the left side the scale represents the target series, while all three right scales represent the values of OJAs on a scrape day. Each scale’s color corresponds to the line color of the respective OJA series. While the values are the same used in the Figure 1, the series are normalized to better visualize fitness with the target variable. The same format is used in the below figure (Figure 4), where the Index series is plotted.

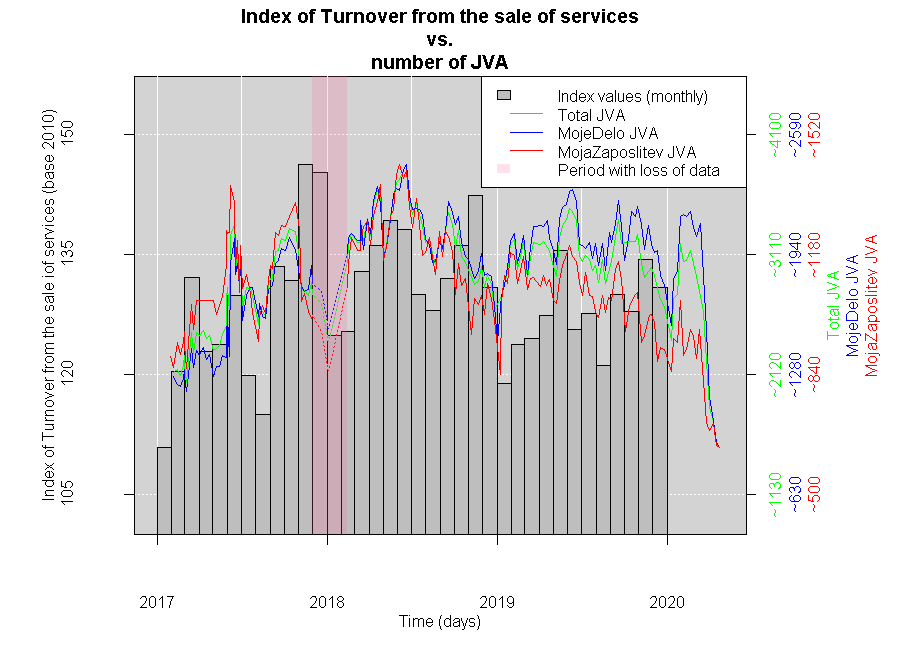


Figure 4: Index and OJA series

We also noticed a possible time lag between the data series, which seems even more probable when all data points are included due to the recent downturn. Thus, we have also checked how data behaves with up to 4 periods of OJAs lagged (we excluded the test data in this exercise). Following the above argumentation, only results for the Index data are shown.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MojeDelo OJA | MojaZaposlitev OJA | Sum of both |
| Index/OJA 1 lag | 0.244 | 0.024 | 0.189 |
| Index/OJA 2 lag | 0.105 | -0.096 | 0.042 |
| Index/OJA 3 lag | 0.308 | 0.121 | 0.274 |
| Index/OJA 4 lag | 0.280 | 0.146 | 0.252 |

Table 2: Correlation of the Index data and lagged OJAs

Against our reasoning, it seems lagged data are marginally worse, the only real candidate being OJA data lagged for 3 periods. Such lagged data will also be analyzed in this document further down. As the reader can observe, the correlation is best (if useful) for the MD series, which will be the only one presented in the rest of the document.

## Linear regression

From observing the above correlation results, we already suspected that MD data will work best for our estimations. This was proved by fitting a usual linear regression of OJAs to the Index series.

Fitting a linear regression

the results are statistically relevant, however the low adjusted statistics indicates that some other regressor would fit better or that some regressors are missing.

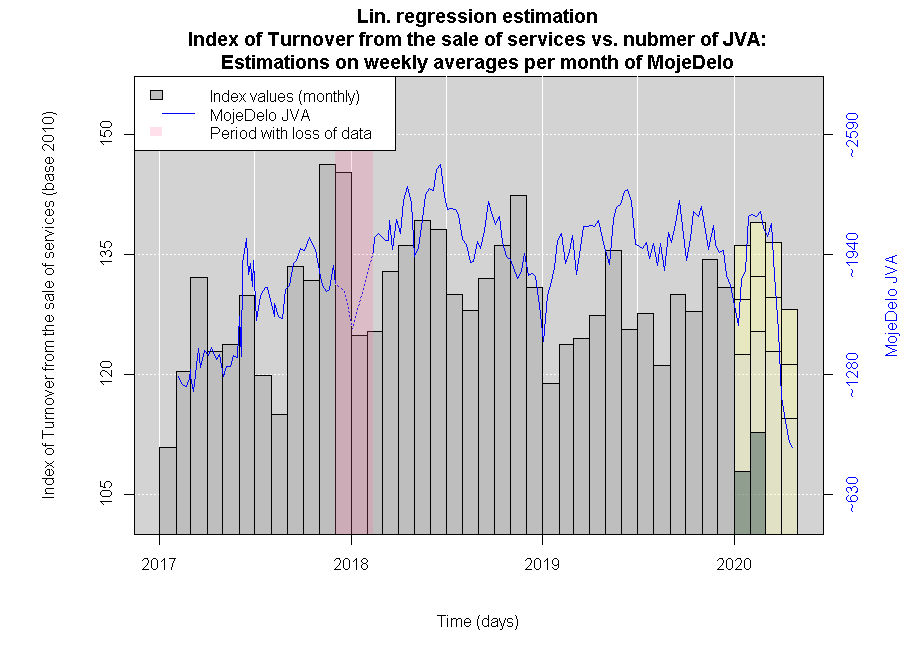
While we will not delve into results of combinations of other series we tried, it is worth noting, that none of them resulted in a better fit. 

Figure 5: Index estimations and realizations using a linear regression

The above graph (Figure 5) depicts predictions of 4 further periods with 2 realizations of the target series. We notice that the target values in early 2020 fall to the lowest level observed. Meanwhile the MD series begins to turn down only after some months. This leads us to believe that the target variable and the regressor are connected through some time-dependent component.

## ARIMAX

The general form of the ARMAX model is a normal ARMA model with an added exogenous variable.

where is the constant term, represents the exogenous variable with parameter , the linear combination of parameters and past values of the target variable represent the AR polynomial and the linear combination of parameters and past values of the white noise variable and includes a current term.

We followed the usual procedure to determine whether a time series possesses the ARMA characteristics and of which degree:

* the Augmented Dickey-Fuller test,
* the Ljung-Box test,
* ACF and Partial ACF curves analysis,
* log-likelihood and AIC comparisons.

The Augmented Dickey-Fuller statistics (0.017 for lag 0 and 0.019 for lag 1 and higher than 0.05 for larger lags) and the Ljung-box statistic (0.015) indicate stationarity and influence of past values. The ACF and Partial ACF curves, seen below (Figure 6), seem to show a presence of a MA term, but not of an AR term. Lastly, comparing log-likelihood and AIC values, the optimal parameter choice presents itself as a model.

Therefore our model with parameters to estimate is an:

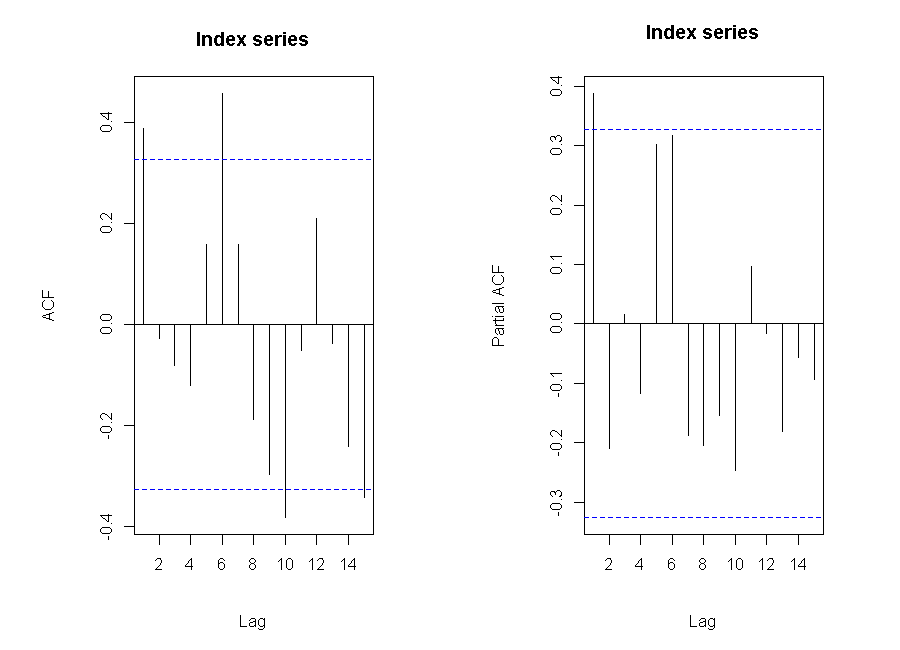


Figure 6: the ACF and Partial ACF graphs for the Index series

Running model fitting to the ARIMAX, the following results are obtained:

However, after using test data to calculate predictions, these are still very incongruent with actual realizations, leading to a fall of MD data a few months too late. This can be clearly seen in the plot below (Figure 7). As such, we must try a different approach to find the unknown time relation.

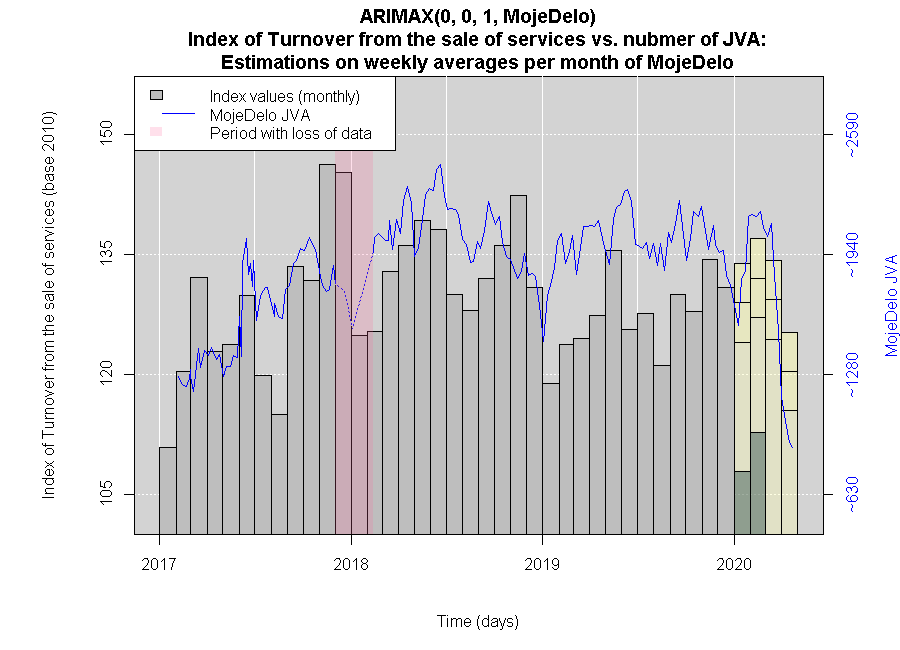


Figure 7: Index estimations and realizations using a ARMAX model

## Lagged MD data fitted to the Index

As described above in Table 2, only one real candidate crops up by the correlation criterion. With a lag of 3 months, we try both the linear regression and ARIMAX models once again. Below are the results of these parameter estimations.

* Linear regression:

* :

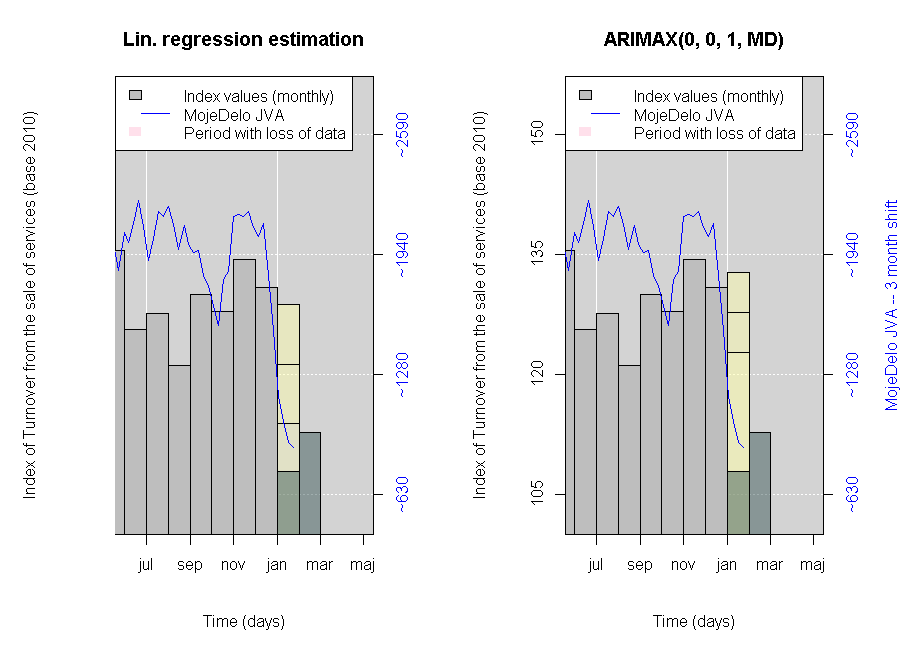


Figure 8: Linear regression and ARIMAX estimations with lagged MD data

The results of these estimations can be considered worse than the results of non-lagged data. The linear regression parameters are not statistically relevant and the adjusted is lower than before. The ARIMAX model seems to be worse as well. The log-likelihood is lower, while the AIC is higher, which indicates, that the original model better explains the link between the Index and the MD series.

# Conclusion

Unexpectedly, it seems that our assumption about time lags was wrong. However, if we extend our train data with the new periods and redo modelling, a new reality is visible.

Linear regression parameters and statistics are also statistically significant according to the t-values:

It seems clear, that the unlagged model does not actually explain the target data and the lagged model is much better at describing the Index than any of the previous. This leaves us with some problems:

* the data does not yet include enough periods, to provide a reliable model, we would need at least one more year of scraping,
* OJA data seems to lag too much to be useful for forecasting or nowcasting the chosen economic indicators, even with its timeliness.

Even with these issues, however, it’s reasonable to assume that both variables are linked. This implies a level of connectivity between OJA data and other economic indicators as well. In the future more analyses could determine a better and more advantageous use of OJA data for estimations in other fields. Furthermore, given the timeliness of OJA data, early estimations could be produced as early as the first week of every month, and subsequent estimations would improve the early values, giving us a dense time series of early estimations of diverse important indicators.