**Use of alternative data sources as flash estimates of economic indicators**

Črt Grahonja, SURS, crt.grahonja@gov.si

**Abstract**

*The Statistical Office of the Republic of Slovenia calculates the statistics of gross domestic product (GDP) every quarter of a year. 60 days after the reference period is unfortunately the quickest we can publish such statistics as the timeliness of GDP data is limited by the survey evaluations of some of the components that make up GDP. However, the use of flash estimates could speed up this process. On the basis of investigation of various big data sources, we had the idea to use the data we acquired from traffic sensors and use them as primary and secondary regressors in a linear regression model for nowcasting GDP 45 days after the reference period. Nowcasting is a method of calculating estimates on the basis of unknown present or near-future values with the use of a known correlator.*

*The article describes our work and the process of nowcasting indicators from the point of data acquisition to the end results on GDP and also on a known GDP correlator, the Industry Production Index. We also touch on what could be extended in the future such as component estimation, model accuracy improvement and data processing improvements.*

*We wish to show how useful such data can be and what was needed to be done before these data could actually be used.*

*Different types of knowledge were used while composing the selected process for economic indicator estimation. These include new skills in the fields of information technology and methodology, and knowledge in the respective subject matters.*

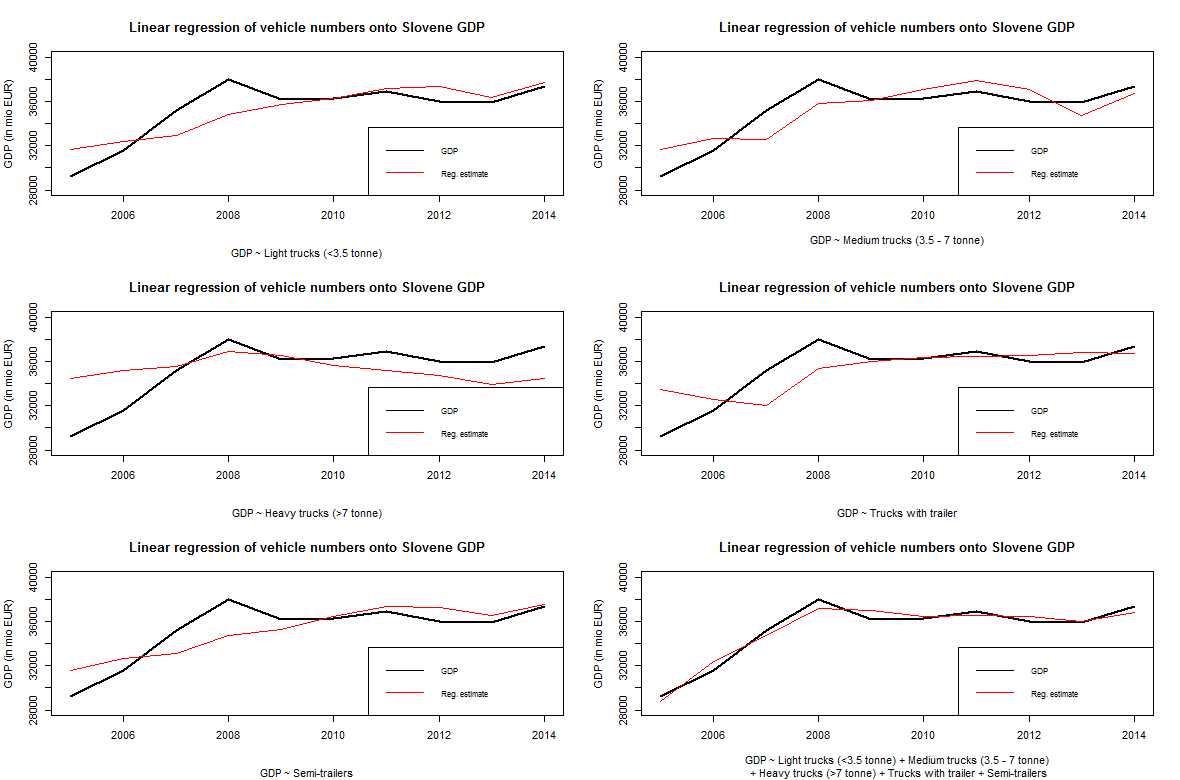
**Keywords:** nowcasting, big data, linear regression, GDP, economic indicators

# Idea and goal

Due to the needs of our users, we wanted to find a way to shorten the interval between collected data and available gross domestic product (GDP) statistics. However, the interval cannot be shortened more than a certain degree, since some of the components are not available until 45 days after the reference period, which is also our goal for the interval length. The main idea is to use flash estimates for timelier results. We turned to nowcasting to improve this interval. Nowcasting is a method of estimating untimely statistics in real or very short time using well-fitted correlators as placeholders for unknown components of the statistics.

We assumed that in Slovenia good choices for such estimation would be a linear combination of companies’ turnover data (one of the main components in the standard GDP estimation) and a big data source in the form of traffic densities, which are a good correlator to GDP, as can be seen in **Picture 1**.

**Picture 1: Linear regression of yearly cargo vehicle densities compared to GDP**



Source 1: SURS

Our target variable is quarterly constant GDP prices from 2011 to 2017, while the independent variables are a set of monthly turnovers of companies and a set of traffic data in 15-minute intervals gathered from traffic loops on public roads throughout the whole country. These variables go through a principal component analysis (PCA) factor model to decrease their numbers. Then their principal component vectors are used in a simple linear regression model.

# Acquiring the data

Since the companies’ turnovers are a component compiled at SURS, these data are available in-house and it is possible to access them for targeted analysis.

After some research we found that most traffic data can be acquired from the Ministry of Infrastructure. These data are ’semi-edited’: the data would come already prepared with distributed categories of vehicles according to a formula that is undisclosed by the Ministry of Infrastructure. The data are split into five cargo vehicle and four non-cargo vehicle categories. The vehicle counts are given for 2 directions. However, on speedways and highways some of the traffic loops only count in one direction (the only available sensors that make up these loops are on one side). The time interval of each observation is 15 minutes.

## 2.1. Coverage

We deem the sensors’ data complete. As they are obtained by the Slovenian Ministry of Infrastructure, all existing traffic sensor data are accessible. The sensors are present throughout the country on every motorway and regional road; however, it must be said that the number of sensors on the motorways is lacking.

A case of over-coverage should be mentioned, since we are not able to differentiate between foreign and domestic traffic in these data. If this difference is found to be important in the future, some kind of linkage with border traffic sensors will be needed to assess the number of foreign vehicles.

## 2.2. Comparability over time

In the Slovenian case, all existing traffic sensor data are obtained by the Ministry of Infrastructure and therefore can be expected to be consistent and available for a long time.

# Reading the data

After checking for missing values, it was clear that not all of the traffic loops were useful for our work. We discarded every traffic loop that had 85% or more of the periods missing throughout the whole observation time or had all sensors turned off for a full year or more. Afterwards we aggregated the data on a monthly basis and then imputed the rest of the missing values.

## 3.1. Imputing missing data

For the rest we decided to impute data based on methods that use each loop’s “neighbours” and the yearly growth of traffic. We tested four different imputation methods and in the end decided on a method that uses four neighbours’ traffic values as the donor. The following formula defines the imputation function:

(1)

In the equation (1) is defined as the set of all traffic loops. For each loop the number represents *all observable* traffic of one vehicle category in month of year , while is the share of periods *actually observed*. If this number is lower than , we set it to 0 (so that it is less costly in terms of computation power). It stands to reason that months that need imputation have this share under 1. The target value represents the imputed value calculated with neighbouring loops of the loop from the set of neighbours , defined as (where is the Euclidian distance between loops and . is a subset of the neighbours that adheres to conditions set in the formula. The imputation is executed individually for each direction, as not every loop counts in both directions.

The accuracy of imputed data was tested against real values with good results. On the individual level, the imputation of a period (month) was usually less than 5% off, while in multiple testing sets of five count spots the effect of the imputation of a period resulted in an error of less than 1%. This type of testing was chosen because it let us analyse neighbouring loops’ values and scales.

On the whole we imputed 2.38% of missing periods (2.30% on regional roads and 3.16% on highways) and had a 2.76% increase in traffic values (2.3% on regional roads and 4% on highways).

# Processing and measurement errors

## 4.1. Processing errors when dealing with traffic sensor data in Slovenia

Even though the data were semi-edited, we still carried out some pre-processing:

* Editing: the data were aggregated into monthly values for each category.
* Weighting: due to long periods of missing data, we excluded those loops that were not working for a full year or more, or had more than 15% of observations missing. Some hesitation concerning lost information was present, but we decided that the density of traffic loops on regional roads is enough to not affect our process. Loops on motorways were less concerning as they were deemed less suitable for our work due to the merely-crossing nature of traffic in Slovenia. Additional weighting happens in the PCA step and linear regression process. This weighting is fully automated by the process and dependent on parameters used.
* Imputations: they were used for missing observations in monthly data. Precision and similarity in terms of scale on small values was paramount. Hence, an extensive analysis of the scale and shape of data and performance tests of four different algorithms were executed.
* Metadata and plausibility errors: a lot of these were present and had to be dealt with. Unstructured (and changing) format of metadata, erroneous time interval notations, use of same characters in variable names and as delimiters, and variable names present in microdata tables were just some of the issues we had to solve.

## 4.2. Measurement errors

In the case of traffic sensors, measurement errors mainly arise from sensors’ malfunctions and/or switching off. This leads to partly or entirely non-existent data for some periods. Missing periods, empty variables, and suspicious uncharacteristic values may indicate periods of malfunction. Unfortunately, these data come from a second-hand source, so we cannot exactly know when a sensor malfunctioned, especially if the malfunction occurred during a 15-minute interval. It’s a fact that we cannot know if a sensor is incorrect without performing a simultaneous manual count on the road it covers. We performed under the assumption that the data given are without measurement errors with the exception of missing data.

# Using the data for flash (rapid) estimates

## 5.1. Use for gross domestic product nowcasts

Using PCA and linear regression we estimated constant GDP prices with industry turnover as primary regression and traffic data as secondary. Testing many combinations, we predicted the current and next period with the optimal model. We expected that the best result would be obtained when using cargo traffic on regional roads. Due to the transit nature of traffic in Slovenia, our view was that the inclusion of motorway traffic would be detrimental to our results. To test our assumptions, we used a dataset of full traffic data, as well as datasets of only cargo traffic data, all traffic on regional roads and cargo-only traffic on regional roads.

## 5.2. Result of GDP estimation

As can be seen from **Table 1** below, the relative errors of estimations when using traffic data are seldom more than 1%.

**Table 1. Estimates and errors of no-traffic data and traffic data sets**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Quarter** | **Official values of GDP (in million EUR)** | **PCA method** | **No traffic data estimates** | **Traffic data as secondary regressor estimates** | **Absolute values of relative errors of the 1st est. (in %)** | **Absolute values of relative errors of the 2nd est. (in %)** |
| **2016Q2** | 9725.868 | *75%* | 9596.824 | 9640.344 | 1.33 | 0.88 |
| *80%* | 9576.406 | 9708.404 | 1.54 | 0.18 |
| *85%* | 9525.735 | 9614.310 | 2.06 | 1.15 |
| *90%* | 9594.405 | 9630.760 | 1.35 | 0.98 |
| **2016Q3** | 9682.643 | *75%* | 9613.220 | 9617.677 | 0.72 | 0.67 |
| *80%* | 9630.709 | 9640.059 | 0.54 | 0.44 |
| *85%* | 9632.367 | 9628.078 | 0.52 | 0.56 |
| *90%* | 9693.590 | 9679.113 | 0.11 | 0.04 |
| **2016Q4** | 9647.458 | *75%* | 9520.605 | 9546.125 | 1.32 | 1.05 |
| *80%* | 9702.478 | 9635.077 | 0.57 | 0.13 |
| *85%* | 9551.981 | 9480.159 | 0.99 | 1.73 |
| *90%* | 9653.184 | 9541.028 | 0.06 | 1.10 |

Source 2: SURS

The next table presents prediction accuracy measures for each combination in the models.

**Table 2. Absolute maximum and mean absolute errors for different traffic count spots datasets**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **PCA method** | **Without traffic data (in million EUR)** | **All vehicle categories, all roads (in million EUR)** | **Cargo vehicle categories, all roads (in million EUR)** | **All vehicle categories, regional roads (in million EUR)** | **Cargo vehicle categories, regional roads (in million EUR)** |
| **Absolute maximum errors** | *75%* | 129.04 | 151..68 | 136.85 | 194.45 | 101.33 |
| *80%* | 149.46 | 154.75 | 146.71 | 105.55 | 42.58 |
| *85%* | 200.13 | 228.54 | 207.52 | 207.62 | 167.30 |
| *90%* | 131.46 | 232.92 | 136.62 | 204.39 | 106.43 |
| **Mean absolute errors** | *75%* | 108.44 | 120.71 | 111.17 | 111.67 | 83.94 |
| *80%* | 85.47 | 88.94 | 88.26 | 68.65 | 24.14 |
| *85%* | 115.30 | 134.35 | 119.59 | 150.90 | 111.14 |
| *90%* | 49.38 | 82.96 | 51.11 | 143.69 | 68.356 |

Source 3: SURS

As can be seen from **Table 2**, our assumptions about traffic data were confirmed. The best model really is computed with cargo data on regional roads. The most important discovery is that these methods return accurate results in a timely manner!

It can also be seen that traffic considerably improves the results. Compared to the non-traffic example, the errors between the official values of GDP and our estimates are reduced by a factor of 4 on average. Furthermore, the maximum absolute error was around 2.5-times smaller. The best PCA parameter seems to be *80%*.

Using data for 2017, we tried to reproduce these results. Once again the results are quite good, with less than 1% difference between the official values and our estimations. However, this time the traffic does not improve the results compared to no traffic models.

**Table 3. Estimates and errors of no-traffic data and traffic data sets**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Quarter** | **Official values of GDP (in million EUR)** | **PCA method** | **No traffic data estimates** | **Traffic data as secondary regressor estimates** | **Absolute values of relative errors of the 1st est. (in %)** | **Absolute values of relative errors of the 2nd est. (in %)** |
| **2017Q1** | 9395.21 | *75%* | 9355.285 | 9317.238 | 0.43 | 0.83 |
| *80%* | 9419.886 | 9336.230 | 0.26 | 0.63 |
| *85%* | 9285.980 | 9305.568 | 1.16 | 0.95 |
| *90%* | 9133.499 | 9275.574 | 2.79 | 1.27 |
| **2017Q2** | 10197.89 | *75%* | 10137.822 | 10103.581 | 0.59 | 0.92 |
| *80%* | 10201.657 | 10118.330 | 0.04 | 0.78 |
| *85%* | 10111.837 | 10096.056 | 0.84 | 1.00 |
| *90%* | 10130.813 | 10178.251 | 0.66 | 0.19 |
| **2017Q3** | 10187.25 | *75%* | 10151.038 | 10077.054 | 0.36 | 1.08 |
| *80%* | 10164.924 | 10045.334 | 0.22 | 1.39 |
| *85%* | 10148.311 | 10002.641 | 0.38 | 1.81 |
| *90%* | 10152.910 | 10505.859 | 0.34 | 3.13 |
| **2017Q4** | 10265.54 | *75%* | 10110.472 | 10224.349 | 1.51 | 0.40 |
| *80%* | 10065.457 | 10099.572 | 1.95 | 1.62 |
| *85%* | 10346.287 | 9998.1344 | 0.79 | 2.61 |
| *90%* | 10188.307 | 10232.035 | 0.75 | 0.33 |

Source 4: SURS

Why does this happen? We suspect that the time series is not long enough to create good enough fitting models. But the issue might also be in the models themselves. In the future we will test more complicated fitting models, such as ARIMA or ADL.

## 5.3. Use for Industry Production Index flash estimates

In regard to the time series length issue, we examined another economic indicator, one that correlates with GDP and has more temporal points: the monthly industry production index (IPI).

With a greater pool of estimations we improved our criteria for an optimal model. We decided to choose the best model according to the *Root Mean Squared Forecast Errors* (RMSFE). We also tested traffic loops as primary regressors.

## 5.4. Results of the IPI estimations

Judging from the RMSFEs of the models (in **Table 4**), the results are similar to the GDP tests. In terms of data combinations, the best model appears to use both the industry and traffic data. The best PCA parameters seem to be either *70%* or *80%*. **Table 4** shows the RMSFE of tested models:

**Table 4. Root Mean Squared Errors for some linear regression - PCA models**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **RMSFE** | PCA: *70%*;  Primary: Industry, Traffic | PCA: *70%*;  Primary: Industry;  Secondary: Traffic | PCA: *80%*;  Primary: Industry, Traffic | PCA: *80%*;  Primary: Industry;  Secondary: Traffic | PCA: *90%*;  Primary: Industry, Traffic | PCA: *90%*;  Primary: Industry;  Secondary: Traffic | PCA: *zadnja5*;  Primary: Industry, Traffic | PCA: *zadnja5*;  Primary: Industry;  Secondary: Traffic |
| **Year** |  |
| **2015** | | 2.90 | 2.50 | 2.25 | 2.84 | 4.45 | 5.97 | 2.82 | 2.47 |
| **2016** | | 2.25 | 2.74 | 2.03 | 3.28 | 3.58 | 3.11 | 4.14 | 2.94 |

Source 5: SURS

Based on these results our procedure is to find the best PCA parameter and dataset combination in a given year with RMSFE as a criterion and use it in the next year for flash estimates. Using this procedure, the *80%* parameter and both sets as primary regressors are used in 2016 and 2017 to calculate flash estimates.

Our estimates for the indices compared to their official values in the months of 2016 and 2017 are shown in **Table 6**. We can see that in most cases the relative errors are in the 1% limit, and the largest relative error is less than 5% off of the true value. This can be seen even better on the picture of comparisons (**Picture *2***).

**Table 5. Official values and estimates of the Industry Turnover Index in 2016**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Month** | **Real value of the Industry Turnover Index** | | **Flash estimate of the Industry Turnover Index** | | **Absolute Errors** | | **Relative Errors (in %)** | |
| **2016** | **2017** | **2016** | **2017** | **2016** | **2017** | **2016** | **2017** |
| **January** | 105.3 | 112.6 | 108.27 | 111.95 | 2.97 | -0.65 | -2.82 | -0,57 |
| **February** | 112.2 | 115.6 | 114.12 | 113.12 | 1.92 | -2.48 | -1.71 | -2,15 |
| **March** | 121.0 | 135.7 | 119.82 | 135.12 | 1.18 | -0.58 | 0.98 | -0,43 |
| **April** | 112.7 | 113.4 | 113.63 | 118.32 | 0.93 | 4.92 | -0.83 | 4,34 |
| **May** | 118.3 | 127.3 | 117.11 | 127.35 | 1.19 | 0.05 | 1.01 | 0,04 |
| **June** | 123.1 | 131.8 | 124.25 | 129.14 | 1.15 | -2.66 | -0.93 | -2,02 |
| **July** | 112.5 | 120.2 | 111.51 | 122.79 | 0.99 | 2.59 | 0.88 | 2,15 |
| **August** | 99.7 | 106.8 | 100.05 | 105.72 | 0.35 | -1.08 | -0.35 | -1,02 |
| **September** | 123.9 | 133.2 | 123.68 | 130.71 | 0.22 | -2.50 | 0.18 | -1,87 |
| **October** | 117.8 | 132.8 | 117.08 | 132.16 | 0.72 | -0.64 | 0.61 | -0,48 |
| **November** | 122.6 | 134.0 | 123.79 | 135.92 | 1.19 | 1.92 | -0.97 | 1,43 |
| **December** | 109.1 | 115.3 | 114.47 | 112.68 | 5.37 | -2.62 | -4.92 | -2,27 |

Source 6: SURS

According to our algorithm, in the next period we are to use cargo traffic on all roads with the *80%* PCA parameter. To check whether this is true, we will need to wait for the end of 2018 to get new data. For now we are pleased with the results, as they continue to inspire confidence in the algorithm and our reasoning.

**Picture 2: Graph of official and estimated values of IPI in 2016 and 2017**



Source 7: SURS

# Conclusion

Our paper shows that flash estimates are a viable choice of estimation. With the right choice of regression variables, the estimates are very accurate. This means that such estimation could be used in the future to publish timelier outputs. However, in our experiment we also found out that even relatively simple new sources need tailored pre-processing, which can limit the usefulness of such approaches. Because of this an in-depth analysis of both sources and data suitability is needed before they are used.

A big opportunity presents itself with the option to output accurate experimental estimations of untimely statistics between current reference dates.