ESSnet Big Data

Specific Grant Agreement No 2 (SGA-2)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/.....

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2016.010-2016.756**

Work Package 6

Early estimates of economic indicators

Deliverable 6.8

Example of calculated concrete estimates for one of the economic indicators with quality assessment of the input, throughput and output phase of the process

Version 2018-28-05

Prepared by: Henri Luomaranta (Finland), Marco Puts (Netherlands), Grzegorz Grygiel (Poland), Alessandra Righi (Italy), Pedro Campos (Portugal), Črt Grahonja, Tomaž Špeh (SURS)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

Contents

1 Summary 3
2 Idea and goal 3
3 Acquiring the data 4
3.1.1 Raw data 4
3.1.2 Semi-edited data 5
4 Reading the files, editing the data 5
4.1.1 Imputing the data 5
5 Using the data for flash (rapid) estimates 7
5.1.1 Use for Gross Domestic Product nowcasts 7
5.1.2 Use for Industry Turnover Index nowcasts 9
5.1.3 Continuation of estimating with newer data 13
6 Conclusion 18
6.1 Literature 19

1 Summary

This document describes calculated concrete estimate of Slovenian GDP 45 days after the reference period with impact and quality assessment of results. It describes implementation of pilot carried out at Statistics Slovenija (SURS) in particular:

- big data source (traffic sensors data) and statistical areas where used,
- other statistical data sources combined with investigated big data sources,
- methods used and impact on quality of results
- data treatment and related methodological and IT issues.

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 Turnover Index.

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

In this report we show how NSIs can address a major quality issue, namely the timeliness, by using a range of micro level data sources accumulated in the registers well before the official release is made, by employing large dimensional econometric models, to form an initial quick estimate of the target indicator. With the nowcasting we are able to publish preliminary results for GDP well before the official release.

We have shown that this does not necessarily lead to too large revisions, but adds significantly to the quality of official statistics through timeliness dimension.

This line of work can proceed in multiple directions:

- Other data sources can be explored with the methodologies we have presented and possibly in relation to other indicators
- Other modeling frameworks are possible
- A real-time application can be programmed, especially relying on the traffic loops data
- These methodologies can be implemented into the production systems of official statistics, and their added value is not limited to nowcasting the GDP or some aggregate final indicators, but could be explored in order to impute some missing components of the aggregated figures.

Such approaches can be easily implemented across the entire European Statistical System.

2 Idea and goal

The Statistical Office of the Republic of Slovenia calculates the statistics of gross domestic product (GDP) every quarter of a year. 70 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 (rapid) 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 regressor in a linear regression method 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.

Traffic sensors data was chosen, because we believe that these data have a good correlation, especially if only the subset, representing cargo vehicles on regional roads, is used. Our assumptions are that cargo traffic moves most of the goods across Slovenia and is the main logistic choice behind import, export and goods acquisition for industries and shops. As such it could reasonably be used as a connection between these unobserved quantities and our target variables. We aim to exploit this connection.

3 Acquiring the data

After some research we found that most[1] of the data can be acquired from the Ministry of Infrastructure. They gave us multiple choices for the format of data.

3.1.1 Raw data

The first choice was raw data from every traffic sensor placed on the Slovenian roads. Since there are different kinds of sensors that count different categories of traffic, this would mean we would need to merge the sensors on the same counting spot according to a formula that would adequately distribute these differing categories. The number of categories differs according to the version of the sensor, as is shown in the table:

Table 1: Traffic sensor versions

QLD3	Sensor QLD3 counts all vehicles
QLD5	Sensor QLD5 distinguishes 5 vehicle categories
QLD6	Sensor QLD6 distinguishes 10 vehicle categories
QLTC8	SensorQLTC8 distinguishes 10 vehicle categories
QLTC10	Sensor QLTC10 distinguishes 10 vehicle categories
QLD	Counted with different versions of sensors

Source 1: Ministry of Infrastructure of the Republic of Slovenia

Sensors have some common features. Every sensor counts traffic on 2 channels, these being the 2 opposing lanes on regional roads or the ordinary and fast lane on speedways and highways. The counting interval is also the same for every sensor at 15 minutes. The data output file is a text file with 11 categories of vehicles, regardless of the number of categories a sensor actually counts. The uncounted categories are not marked, but are filled with zeroes. The data also contain other information, such as the highest, lowest and average speed in the interval, the average of specifically personal vehicles, the average time gap between vehicles, the occupancy of the lanes and the temperature at observation time.

[1] The data we are talking about here refers to national roads of the regional and highway variety. Additionally, municipality road sensors can be acquired at the administrative offices across Slovenia. SURS also has access to sources which include traffic sensors at national borders.

3.1.2 Semi-edited data

Additionally, we could choose 'edited' data. The data would come already prepared with categories distributed according to the formula used by the Ministry of Infrastructure into **count spots** data. The final product is joint annual data for each count spot, which includes categories for Motorcycles, Personal Vehicles, Light Goods Vehicles, Medium Goods Vehicles, Heavy Goods Vehicles, Semi-Trailers, Tractor-Trailers, and Buses. The vehicle counts are given for 2 directions. However, on speedways and highways some of the count spots only count in one direction (the only available sensors that make up these counting spots are on one side). The available time intervals of vehicle counts are 15-minute, 1-hour or daily intervals.

The 15-minute interval of semi-edited data was chosen for our work. We gained it from the Ministry on an external drive in the form of 3529 text files in 24 GB of size. With the choice of edited data we skipped quite a number of steps towards final data; however, after careful examination it turned out that the text files were not in an easy-to-read format and furthermore there were mistakes in the texts. Nevertheless, we decided to use this source (for now), as the unification of raw sensor data outweighed the work needed to repair mistakes in the files. The chosen time interval of each observation was 15 minutes for the period between 2011 and 2016.

4 Reading the files, editing the data

We prepared a number of programs that read the files, checked for missing values and errors, and stored the data. Many of these programs take advantage of multi-thread processors in our computers. We created a new object class in Python, defined upon the *pandas* package class *DataFrame*, which stored all the yearly datasets for a single count spot. After checking for missing values, it was clear that not all of the count spots were useful for our work. We discarded every count spot that had 85% or more of the periods missing throughout the whole observation time or had all sensors turned off for a whole year. Afterwards we aggregated the data on a monthly basis and then imputed the missing values.

4.1.1 Imputing the data

We decided to impute data based on methods that used each count spot's "neighbours" and the yearly growth of traffic. We tested different imputation methods and in the end decided on the following one:

$$imp\left(d_k^{m,t} g_k^{m,t}\right) = \begin{cases} \left(\dfrac{\sum_{s_k \in K' \subseteq s(k)} g_{s_k}^{m,t}}{\sum_{\substack{s_k \in K' \subseteq s(k) \\ Z^d \forall s \in K'}} g_{s_k}^{m,t-l}} g_k^{m,t-l} \right)\left(1 - d_k^{m,t}\right) + d_k^{m,t} g_k^{m,t}, & \sum_{\substack{s_k \in K' \subseteq s(k) \\ Z^d \forall s \in K'}} g_{s_k}^{m,t-l} \neq 0 \\ g_k^{m,t-l}\left(1 - d_k^{m,t}\right) + d_k^{m,t} g_k^{m,t}; & d_k^{m,t-l} = 1, \\ d_k^{m,t} g_k^{m,t}; & otherwise, \end{cases}$$

where S is the set of all count spots;

$g_k^{m,t}$- is the number of ***all observable*** traffic of one vehicle category through count spot k in the month m of year t – this amount is often unknown to some degree, due to failure or inactivity of the sensors in the count spot, and imputations try to estimate it;

$d_k^{m,t}$ is the share of periods with ***existing/observed*** data in the month m of the year k through count spot k, and this number is either 0 or $\frac{days\ in\ motnh-20}{days\ in\ month} \leq d_k^{m,t} \leq 1$. It stands to reason that months that need imputation have this share <1;

$imp\left(d_k^{m,t} g_k^{m,t}\right)$ is the imputed value;

s_k is a neighbour of the count stop k;

$s(k)$ is the set of neighbors for the count stop k, defined as $s(k) = \left\{ s_i;\ i = 1 \dots 4 | s_i = \arg\min_{\substack{s_i \in S \\ s_i \neq k}} \left(\Delta_e(k, s_i)\right) \right\}$;

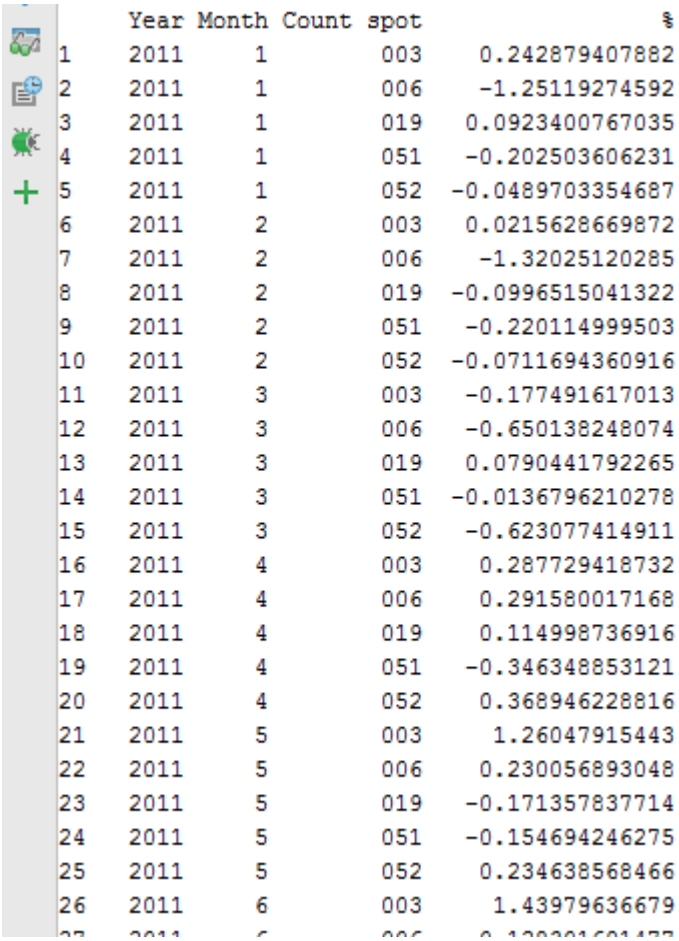
K' is a subset of the neighbors, which adheres to some conditions; and

$\Delta_e(k, s_i)$ is the Euclidian distance between count spots k and s_i.

We also had to take into account for missing data in one of the directions on some of the count spots on the highways. We modified slightly our program to choose from adequate neighbours in those instances.

After the imputation was finished we tested the accuracy of imputed data against real values and got very good results. At the individual level, the imputation of a period (month) was usually less than 5% off, while in a testing set of 5 count spots the effect of the imputation of a period resulted in an error of less than 1%.

Picture 1: An example of the effect of an imputation on a period on a set of 5 count spots



	Year	Month	Count spot	%
1	2011	1	003	0.242879407882
2	2011	1	006	-1.25119274592
3	2011	1	019	0.0923400767035
4	2011	1	051	-0.202503606231
5	2011	1	052	-0.0489703354687
6	2011	2	003	0.0215628669872
7	2011	2	006	-1.32025120285
8	2011	2	019	-0.0996515041322
9	2011	2	051	-0.220114999503
10	2011	2	052	-0.0711694360916
11	2011	3	003	-0.177491617013
12	2011	3	006	-0.650138248074
13	2011	3	019	0.0790441792265
14	2011	3	051	-0.0136796210278
15	2011	3	052	-0.623077414911
16	2011	4	003	0.287729418732
17	2011	4	006	0.291580017168
18	2011	4	019	0.114998736916
19	2011	4	051	-0.346348853121
20	2011	4	052	0.368946228816
21	2011	5	003	1.26047915443
22	2011	5	006	0.230056893048
23	2011	5	019	-0.171357837714
24	2011	5	051	-0.154694246275
25	2011	5	052	0.234638568466
26	2011	6	003	1.43979636679

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).

5 Using the data for flash (rapid) estimates

5.1.1 Use for Gross Domestic Product nowcasts

After the imputation was done, we have used the final data as a secondary regressor in a nowcasting method. The nowcasting method that consists of PCA and linear regression was used to find the best fit of quarterly enterprise turnover data onto GDP values. Afterward, we predicted the next period with the optimal model.

We have tested many combinations of regressors, by using the nowcasting method that consists of PCA and linear regression with parameters on the main regressors and adding traffic datasets as secondary regressors. We expected that the best result would be obtained when using cargo traffic on regional roads. Since Slovenia is a transit country, the so-called crossroads between East and West and North and South of Europe, our view was that the inclusion of highway traffic would be detrimental to our results. In order to justify our assumptions we used a dataset of full traffic data, a

dataset of only cargo traffic data, a dataset of traffic on regional roads and a dataset of only cargo traffic on regional roads.

We have then used these linear regression models to nowcast the estimates of GDP in three points in time (namely second, third and fourth quarter of 2016). The results are quite clear; the traffic data improves the quality of estimates in every occasion in comparison of solely use of industry data, while consistently the best PCA condition seemed to be the *80%*[2], which takes just enough first few principal components to explain 80% of the variability of the microdata.

Table 2: Absolute maximum 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)
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
po10	282.41	313.32	248.34	216.84	207.59
zadnja5	102.70	110.33	113.10	118.91	56.78

Source 2: Own

Table 3: Absolute mean 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)
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
po10	255.25	225.39	211.97	144.45	174.05
zadnja5	66.96	72.78	66.87	48.17	37.20

Source 3: Own

[2] The PCA method is used because the number of variables in microdata is too large for linear regression and therefore needs to be reduced. In our models the names *75%, 80%, 85%* and *90%* mean PCA conditions that take just enough first few principal components (variables) to explain 75%, 80%, 85% or 90% of the variability of the microdata. The chosen principal components are eigenvectors with the highest eigenvalues of the covariance matrix of the microdata.
The PCA method with PCA condition *po10* takes just enough first few principal components to have at least 10 times more time points from the beginning to the ending of the time span.
The PCA method with PCA condition *zadnja5* takes all eigenvectors (principal components) whose eigenvalues have at least a 5% share among all of the eigenvalues.

As can be seen from Table 2 and Table 3, our assumptions about traffic data were right. This can be seen by the fact that in almost every instance any use of traffic data corrects the estimates a little. However, we were pleasantly surprised when we checked for the best combination. Use of any model with the exclusion of *90%* resulted in better accuracy when using cargo data on regional roads. The errors between the official values of GDP and our estimates are on average reduced by a factor of 4, comparing it to the non-traffic example. Furthermore, the maximum absolute error was around 2.5 times smaller when using traffic data in comparison to when not using traffic data. Even the best model without usage of traffic data was still worse than some of the models with included traffic data (but it must be said that we suspect that there was a high chance of overfitting in the best non-traffic model). Lower errors when comparing all roads to regional roads or all vehicle categories to cargo vehicles also support our assumptions.

As can be seen from Table 4 below, the relative errors of estimations when using traffic data are seldom more than one percent.

Table 4: Estimates and errors of no-traffic data and traffic data sets

Period	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
		po10	9523.429	9548.396	2.08	1.83
		zadnja5	9678.580	9713.673	0.49	0.13
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
		po10	9400.231	9545.563	2.92	1.42
		zadnja5	9631.749	9640.023	0.53	0.44
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
		po10	9366.567	9439.871	2.91	2.15
		zadnja5	9544.756	9590.680	1.07	0.59

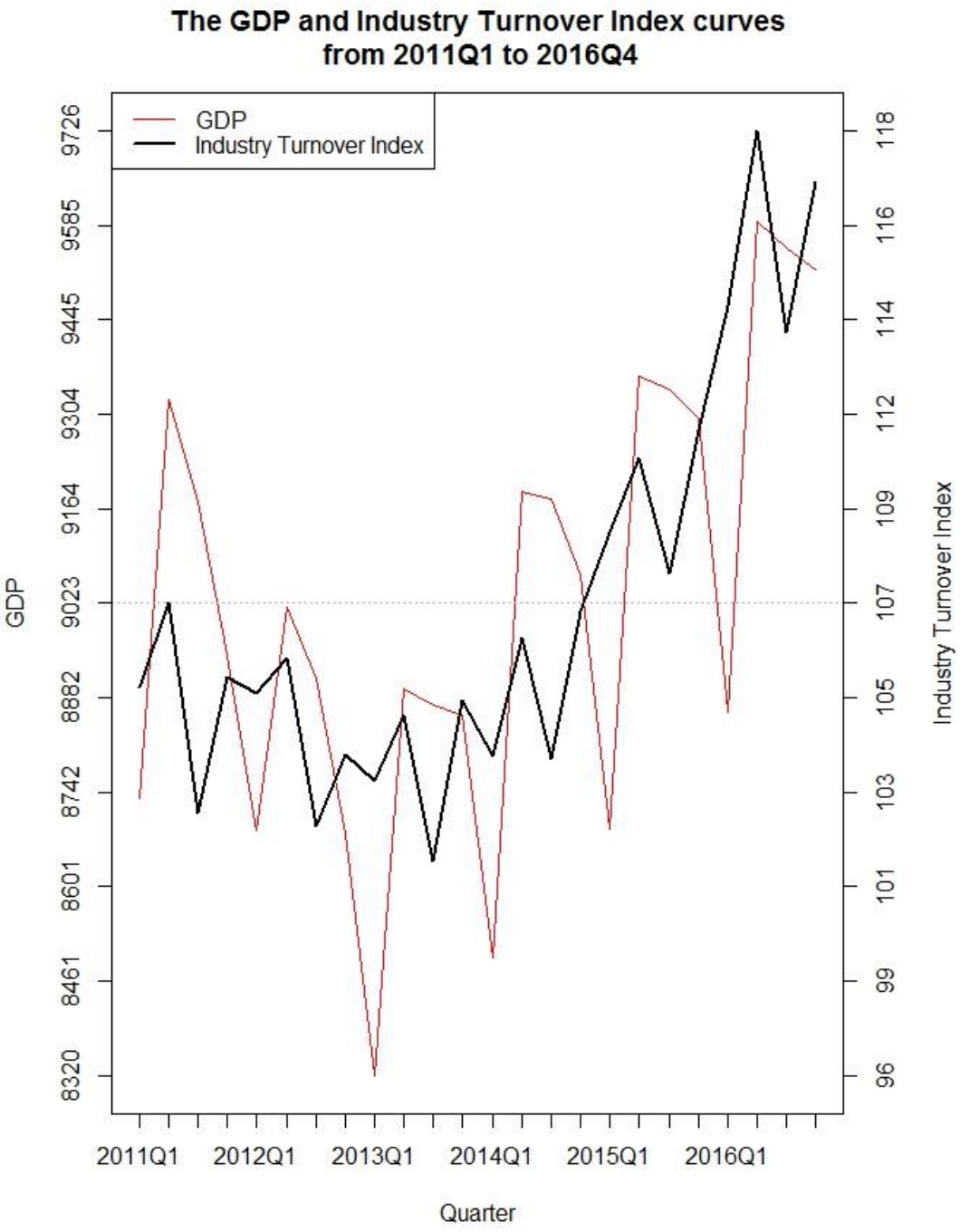
Source 4: Own

5.1.2 Use for Industry Turnover Index nowcasts

While these are really promising results, we are aware that an assessment on three temporal points is far from optimal. This is why we searched for a longer time series where traffic data could be used. With the assumption that industry turnover data represent a big part in the evaluation of GDP and therefore the two might correlate well, we tried using the traffic data as primary or secondary

regressors for Industry Turnover Index nowcasts. Encouraged by the results above we expected that traffic data could give good assessments.

Picture 2: GDP and Industry Turnover Index time series



As before, we tested different linear regression with PCA models with different parameters. Furthermore, we also used different sets of data for regression variables. We decided to test industry data as fitted with industry data alone, industry and traffic data as primary regressors, industry data as primary and traffic data as secondary regressors, and traffic data alone as primary regressors. The learning phase was conducted on the data from 2011 to 2014 (64 temporal points) and the testing phase was conducted on the remaining 24 temporal points. With a larger pool of

results we could also improve our criteria for an optimal model. We decided to choose the best model according to the *Root Mean Squared Forecast Errors* (RMSFE), which takes into account the accuracy of the forecast in every testing temporal point and also their variability. Therefore, the best model is such that has the best combination of closest average to the true values and the minimal deviation from that value in every temporal point in its monthly forecasts in a pre-agreed period.

The actual procedure was to test each parameter at the end of every year, and use the acquired optimal setup when nowcasting monthly values in the following year. An example of this will be given with the next table.

Judging from the RMSFEs of the models (in Table 5), the results are the same as in the above testing. In terms of data combinations, the best appear to use both the industry and traffic data. Meanwhile, the best PCA parameters seem to be either *70%* or *80%*. The *zadnja5* model is also close to being the optimal model.

The following table shows the RMSFE of tested models:

Table 5: Root Mean Squared Errors for some linear regression - PCA models

	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**
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: Own

Based on the results of testing, in both years we would pick the model with the PCA parameter *80%* and both sets of data as primary regressors as our optimal model. The actual assessment of the Industry Turnover Index would be produced in the following way:

1. At the end of 2015 we would test different models and according to the results choose the optimal model (PCA: *80%*, both sets of data as primary regressors).
2. We would use this model for every monthly nowcast of 2016 and declare the assessments as our flash estimates.
3. At the end of 2016 we would again test all the models and choose the new optimal model (in this case the same one) to be used throughout 2017.

Using this strategy our estimates for the indices compared to their real values in the months of 2016 are:

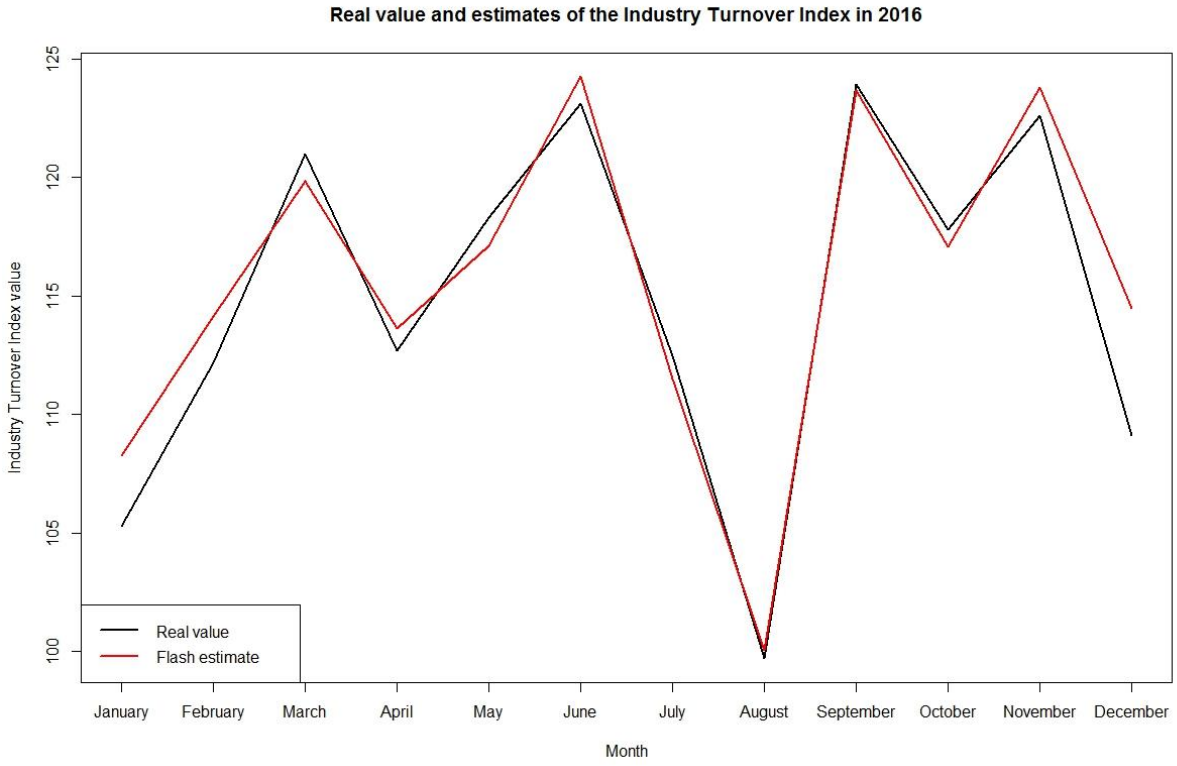
Table 6: Real values and estimates of the Industry Turnover Index in the year 2016

Month	Real value of the Industry Turnover Index	Flash estimate of the Industry Turnover Index	Absolute Errors	Relative Errors (in %)
January	105.3	108.27	2.97	-2.82
February	112.2	114.12	1.92	-1.71
March	121.0	119.82	1.18	0.98
April	112.7	113.63	0.93	-0.83
May	118.3	117.11	1.19	1.01
June	123.1	124.25	1.15	-0.93
July	112.5	111.51	0.99	0.88
August	99.7	100.05	0.35	-0.35
September	123.9	123.68	0.22	0.18
October	117.8	117.08	0.72	0.61
November	122.6	123.79	1.19	-0.97
December	109.1	114.47	5.37	-4.92

Source 6: Own

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 graph of comparisons of the two series below.

Picture 3: Comparison of the real values and estimates of the Industry Turnover Index in 2016



5.1.3 Continuation of estimating with newer data

Obtaining new data from 2017, we have tried to reproduce these results. The overall experiment was conducted on additional 4 time periods – every quarter of 2017. Once again the estimations were calculated using a PCA method with five different parameters and using the traffic data in different ways: either a total sum of traffic data in an interval as a secondary regressor or individual count spots as primary regressors to be trimmed by the PCA method.

The newer results for 2016 and 2017 are presented in the table below.

Table 7: Absolute maximum 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)
75%	155.07	266.88	158.51	154.98	135.00
80%	200.08	155.68	135.22	576.83	230.38
85%	324.01	111.88	321.65	419.63	267.41
90%	261.71	221.11	295.35	274.29	318.61
zadnja5	224.39	273.89	269.87	416.65	218.09

Table 8: Absolute mean 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)
75%	81.07592	106.34516	102.92959	55.07030	60.96715
80%	50.84946	68.59371	53.96622	26.16266	30.22127
85%	1.605270	53.435054	19.905594	79.987767	113.048431
90%	62.86737	128.62754	120.81747	170.62065	79.05392
zadnja5	62.93502	111.45412	174.36118	47.16852	57.26990

As we can see from these tables, some findings from the previous experiments seem to still hold true. Estimating with all vehicles on all roads gives worse results than using no traffic at all. Likewise, using only cargo traffic on all roads is not the best choice either. From both tables we can see that the *90%* parameter gives bad results, as the maximum error throughout the time intervals is big, and the mean error shows that estimations with non-traffic regressors perform better. However, the estimation is not the best possible. We can also see that in these new periods an alternative to the cargo traffic on regional roads has arisen for good estimation. Using all vehicles on regional roads has the same low mean error, but its maximum errors are quite big compared to the cargo traffic on regional roads.

In the end it seems like the best series is still the cargo traffic on regional roads series. Just as with the previous data, we directly compare official values to the values of estimation without and with traffic data. The official values 2016 are again included, since the GDP series is subject to revisions and the values are different.

Table 9: Estimates and errors of no-traffic data and traffic data sets

Period	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	9755.848669	*75%*	9641.769	9620.851	1.17	1.38
		80%	9655.510	9601.865	1.03	1.58
		85%	9627.988	9660.667	1.31	0.98
		90%	9689.009	9531.461	0.69	2.30
		zadnja5	9531.461	9541.131	2.30	2.20
2016Q3	9748.845959	*75%*	9642.126	9699.763	1.10	0.50
		80%	9681.626	9637.750	0.69	1.14
		85%	9656.819	9658.668	0.94	0.93
		90%	9716.181	9464.554	0.34	2.92
		zadnja5	9704.672	9530.760	0.45	2.24
2016Q4	9686.031852	*75%*	9570.723	9700.913	1.19	0.15
		80%	9629.284	9582.602	0.59	1.07
		85%	9722.534	9553.433	0.38	1.37
		90%	9593.896	9600.487	0.95	0.88
		zadnja5	9646.320	9646.767	0.41	0.41
2017Q1	9395.205718	*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
		zadnja5	9300.724	9308.459	1.01	0.92
2017Q2	10197.88759	*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
		zadnja5	10182.237	10248.388	0.15	0.50
2017Q3	10187.24812	*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
		zadnja5	10273.918	10347.609	0.85	1.57
2017Q4	10265.53999	*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
		zadnja5	10277.841	10339.279	0.12	0.72

As the GDP statistic is one that is revised with new data, the official values changed through the year. We are now trying to devise a strategy of estimation around the changing values: common sense dictates that the target variable in the training set should change according to the current

information! Since revisions of values happen in every quarter for the current year and in the last quarter for the previous year, it is important to consider which information is fed to the algorithm at which point in time.

With the new data we were able to create new estimates for the Industry Production Index. As with the GDP series, this too is subject to revisions and the old 2016 values have changed, although not by much. The table below presents the new values for 2016, using the same PCA – linear regression parameters as before: the *80%* PCA parameter and using industry data and count spots as primary regressors to be fed to the PCA method. Following our strategy of calculating the best PCA parameter at the end of a year to use throughout the next year, new estimates for 2017 are also included using the parameters we determined at the end of 2016: again the *80%* PCA method with both industry and traffic data as primary regressors.

Table 10: Estimates for the Industry Production Index

Month	Real value of the Industry Production Index	Flash estimate of the Industry Production Index	Absolute Errors	Relative Errors (in %)
January 16	105.3	102.15	-3.15	-2,99
February 16	112.1	108.34	-3.76	-3,36
March 16	120.9	120.28	-0.62	-0,51
April 16	112.9	112.80	0.20	0,18
May 16	118.1	117.63	-0.47	-0,40
June 16	123.0	119.70	-3.30	-2,69
July 16	112.4	114.07	1.67	1,48
August 16	99.6	98.61	-1.00	-1,00
September 16	123.7	123.98	0.28	0,22
October 16	117.7	121.89	4.19	3,56
November 16	122.4	122.40	-0.00	0,00
December 16	109.0	110.63	1.63	1,49
January 17	112.6	111.95	-0.65	-0,57
February 17	115.6	113.12	-2.48	-2,15
March 17	135.7	135.12	-0.58	-0,43
April 17	113.4	118.32	4.92	4,34
May 17	127.3	127.35	0.05	0,04
June 17	131.8	129.14	-2.66	-2,02
July 17	120.2	122.79	2.59	2,15
August 17	106.8	105.72	-1.08	-1,02
September 17	133.2	130.71	-2.50	-1,87
October 17	132.8	132.16	-0.64	-0,48
November 17	134.0	135.92	1.92	1,43
December 17	115.3	112.68	-2.62	-2,27

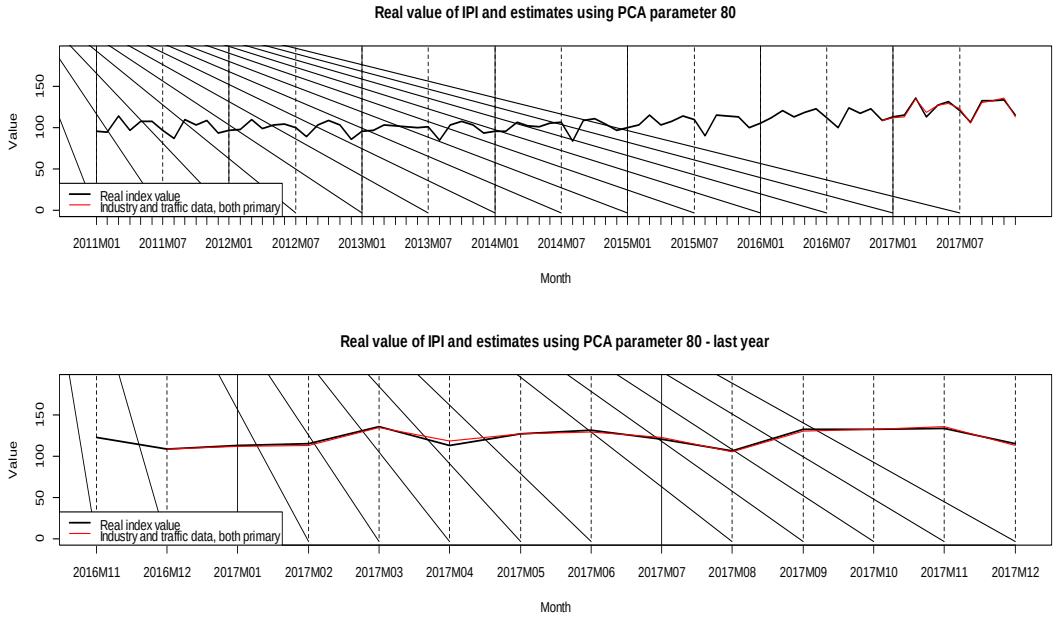
As we can see, the estimates are consistently good, with none of them being over 5% off of the real index values. According to our algorithm however, in the next period we are to use a different series for the estimation process. The best estimation according to the RMSFE is the one using cargo traffic 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.

The reasoning behind these decisions is shown in the following table:

The figure shows how our estimations for 2017 (in red) correspond to the real values of the index (in black).

Picture 5: Real and estimated IPI



6 Conclusion

We have shown that the use of linear regression and PCA methods together with traffic loops data can produce accurate estimations of some early economic indicators. Additionally, since data acquired by smart sensors such as traffic density are measured automatically, we could replace some of the sources used in the normal calculation of the statistic with these data and reduce the burden of survey respondents, or use it as an additional burden-free source for more accurate calculations.

What we gathered during our processing of data and the errors is that quite a lot of errors arise from specific characteristics of given datasets (metadata errors, weighting circumstances, editing) and cannot be generalized to the whole or at least to a big part of the Big Data field. The solutions to most of these individual problems need to be specifically made for them. Lucky for us the source of traffic sensor is fairly consistent and the solutions we devised in this pilot can be applied in the future as well without (much) changes. Furthermore, some of these problems are sensor-specific and while each problem needs to be addressed individually, they are part of a common set of problems that arise when working with other smart sensors as well. As such a lot of solutions on other smart sensor data can probably be adapted from today's (a good example is missing data due to malfunctions).

We have shown that the use of linear regression and PCA methods together with traffic loops data can produce accurate estimations of some early economic indicators. Additionally, since data acquired by smart sensors such as traffic density are measured automatically, we could replace some of the sources used in the normal calculation of the statistic with these data and reduce the burden of survey respondents, or use it as an additional burden-free source for more accurate calculations.

What still needs to be done: in some years enough quarters will pass to efficiently use RMSFE criteria for optimal model selecting on GDP. At that moment we will be able to accurately assess whether it is better to flash estimate directly GDP or if we should focus on rapid estimations of its less-timely components and then use them as regular data in the normal estimation process for GDP. In addition, it has to be investigated what precision is required in order to produce reliable earlier estimates, explore the possibilities to nowcast only some components of quarterly GDP (for which all data are not available on time), include other big data (and other) sources and also test some additional methods for nowcasting.

Going outside of national statistic territory, traffic data could be researched for use in travel/tourism statistics, population mobility statistics, regular traffic statistics, etc.

6.1 Literature

1. New steps in big data for traffic and transport statistics, CBS; https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/recente-berichten/new-steps-in-big-data-for-traffic-and-transport-statistics
2. Traffic Loop Data, M. Puts, B. Meindl, L. Ma, P. Del Rey; https://statswiki.unece.org/download/attachments/109252755/Marco%20Puts.pdf?version=1&modificationDate=1425992553561&api=v2
3. Traffic loop data for transport statistics, P. Struijs; https://unstats.un.org/unsd/trade/events/2014/Beijing/presentations/day1/afternoon/4.%20Statistics%20Netherlands%20Traffic%20Loop%20Data%20for%20Transport%20St.pdf
4. Fornaro, P., & Luomaranta, H. (2017). Aggregate fluctuations and the effect of large corporations: Evidence from Finnish monthly data. Economic Modelling. doi:10.1016/j.econmod.2017.11.012
5. Julie Josse, Francois Husson (2016). missMDA: A Package for Handling Missing Values in Multi- variate Data Analysis. Journal of Statistical Software, 70(1), 1-31. doi:10.18637/jss.v070.i01
6. The Elements of Statistical Learning. (2009). Springer Series in Statistics. doi:10.1007/b94608
7. James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, 23(6):405–430, 2004. ISSN 1099-131X. doi: 10.1002/for.928. URL http://dx.doi.org/10.1002/for.928.