ESSnet Big Data

Specific Grant Agreement No 1 (SGA-1)

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 **11104.2015.007-2016.085**

Work Package 6

Early Estimates

Deliverable 6.1

Potential Big Data and other sources with business cases for the aim of early estimates

Version 2016-16-06

Prepared by:

Boro Nikic (SURS, Slovenia)

ESSnet co-ordinator:

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

Table of contents

1. INTRODUCTION

The aim of WP6 - Early estimates was to investigate potential of big data and others sources in order to combine theme for purposes of early estimates. The WP6 team was consisted from four NSIs:

Finland (Henri Luomaranta)

Netherlands (Piet Daas)

Poland (Anna Nowitzka)

Slovenia (Boro Nikić - WP6 coordinator)

The main goal of the WP6 tem was to explore how a combination of (early available) multiple big data sources, administrative and existing official statistical data could be used in creating an existing or new early estimates for official statistics. The study included exploration of;

- big data sources and statistical areas where those sources could be used
- other administrative and statistical sources which could be combined with investigated big data sources
- possible business cases which could be tested in SGA2 period
- data collection, data linking, data processing, methodological and IT issues
- results of one or two pilots which may help us to determine the most prosperous business case for SGA2. Proposed pilot were Nowcasts of Turnover Indices or (and) Consumer Confidence Index

This deliverable focuses on results of study related to list of potential big data sources and proposed business case for SGA2.

2. INVESTIGATION OF BIG DATA AND OTHER SOURCES

2.1 Investigation of possible data sources

One of the results of brainstorming sessions at some of the NSIs involved in WP6 (and WP7), questionnaire set to participated NSIs in ESSnet Big Data project and discussion of members of WP6 team the initial list of possible big data and other sources and possible statistics which could be calculated out of detected data sources was prepared.

Table1: List of possible data sources with statistical domain where they could be employed

STATISTICAL DOMAIN	DATA SOURCES	STATISTICS
Tourism (1)	Mobile phone data, traffic counters at border crossing (including recognizing the number of plate of the vehicle), flight and train tickets, surveys…	Number of foreign tourists, number of (tourists) vehicles passing the country,
Tourism (2)	Mobile phone data, surveys…	Number of tourists, lengths of trips…
Population mobility	Mobile phone data, surveys…	Number of (short) travels per day, average travelled distance per day…
Health statistics	E-health recipes, personal health cards, pharmacies, surveys…	Use of medicines (by age groups, gender, territory…)
Agriculture	Airplane or satellite images, surveys…	Utilized agricultural area, arable land, share of permanent crops in unutilized areas…
Quick and dirty statistics (in all statistical domains)	NSI data & e.g. google trends tool	Flash estimates of all kind of early statistics
Statistics for the internal NSI purposes	Newsfeeds, social media data	Monitor and detect the statistical products and areas which occur in statistical and other web news

		• Detect new statistical products which are very frequent and not covered by NSIs production yet • Detect the statistical products produced by NSIs for which there is almost no demand • This information which help the management of NSIs (together with stakeholders) to decide for which statistical product there is high public demand
Economic indicators: • **Gross domestic product (GDP)** • **Consumer price index (CPI)** • **Retail sale** • **Balance of payments** • **Economic sentiment indictors**	Big Data: Job vacancies ads from job portals, traffic loops, Social data (Twitter, Facebook, etc.), supermarket scanner data, bank transaction data, news feeds/messages, Registers and existing sources: Statistical Register of Employment, data from the Employment Agency, tax data, wages and salaries Surveys: Turnover data from various short-term surveys, Business confidence index, Consumer confidence index	Flash and (or) intermediate estimates of economic indicators

2.2 Combining of data sources

When we think of combining of data sources in the traditional statistical production we mostly think of combining them on micro level. If common identificator exists the linking of data is quite straight forward otherwise the various record linkage methods are applied in order to derive Id in data set where Id is missing. In the area of big data the issue of combining of different data set is more complicated. Often the (big) data sources are completely different, so we are not able to employ record linkage techniques or one of data sets contain unstructured data where we need to employ big data techniques like machine learning in order to link data.

The other possibility for linking of data sources is linking on macro level. Here we whether try to aggregate all data sets on level which has common identificator whether we include data in nowcasting models.

Nowcasting1 is a very early estimate produced for an economic variable of interest over the most recent reference period calculated on the basis of incomplete data using a statistical or econometric model different from the one used for regular estimates. Soft data should not play a predominant role in nowcasting models. Nowcasts may be produced during the very same reference period for which the data is produced.

Conducting one of the pilots during SGA1 ESSnet project several nowcasting methods were under investigation. Among them the most promising in sense of practical implementation was Principal components analysis method. The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

When nowcast early indicators with PCA model big data and other sources could be combined in two ways:

- as a regressors in nowcast equation $y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + \beta_1 y_1$. Variables $x_1, x_2, \cdots, x_k$ are principal components of given set of (big) data and variable y_1 is aggregated set of combined (big data) data source.

In the pilot conducted at SURS where we tested the PCA model in order to estimate Real turnover index in industry (time series of interest) the Real turnover of industrial enterprises (time series of enterprise data used for determination of principal components) was combined with Economic sentiment indicator used as an additional predictor in linear regression.

- as a micro data in nowcast equation $y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + \beta_1 y_1$. Variables $x_1, x_2, \cdots, x_k, y_1$ are principal components of given two sets of (big) data

[1] Overview of GDP flash estimation methods; Eurostat 2016 Edition

Table 2: example of combined data from industry survey and traffic loops data

	P1003162	P1003164	P1003166	P1003168	P1003170	P002	P003	P006	P010
1	216218	44818,67	71895,67	56055,67	260288,3	1824,333	7714,333	15132	4092,667
2	245734,7	80276,67	60211	39325,67	217410,7	1419,667	10247,67	19597,67	5346,667
3	296705	47200,67	64541	75624,33	269679	1392	9770,667	19843,33	6352
4	231986,3	55985	94126	65412	276388,7	1088	8770	23579	5138,667
5	264973	38550,67	66172,67	59620,67	169702,7	1098,333	7693,667	19336,67	4771,333
6	276598,3	29820,67	40195,67	76714	182335	1050,667	9514	19132,67	6456,333
7	247292	19871,67	64405,33	88768,67	198847,3	1064,667	9937	17248,33	6919
8	272853	16915	93066,67	78192	229577	883,6667	7929,667	15822	4068,667
9	300986,3	18163,67	52228	51534,33	185970,3	775	6209	12693,33	3409,333
10	272550,7	13266,33	61242,67	47449,67	246491,3	922	8151	16869	4837
11	288410,3	NA	74888	73008,67	284151,3	1226,333	9448,667	19202,67	5885,667
12	290073	NA	69280,33	63277	202978,3	1091,333	7900,667	16063,67	3566,551
13	310173	NA	83583	57082	162391,7	788,3333	4983,333	13022	2827,333
14	316162,7	NA	75582,67	78835,33	147334	1039,667	6120,333	15912,33	5133,889
15	334938,7	NA	61854,67	57812	262239,3	1191,434	5967,333	16233	5925,667
16	375113,7	NA	84696	56782,67	190403,3	922,3333	5605	15378,67	3462,677
17	342045	NA	72032,33	39791,33	115945	736,6667	5063	13228,67	3653,667
18	334273,7	NA	69265	42082,67	136967,7	957,6667	5777,333	16466,67	5296,667
19	386572,7	NA	65479,67	273164,3	215871,3	1241,667	5768,333	16714	6395,333
20	449406	NA	76241	59811,67	129889	986,3333	5590,333	16164,33	3956
21	404387	NA	37974,33	30490,67	NA	753,6667	4553	13784	3841,333
22	444126,7	NA	78102,67	43745,33	NA	1030,667	5870,667	18748,67	5892,667
23	438757,7	NA	86084	39027,67	NA	1148	5895,667	17713,33	7566
24	492560,3	NA	36321	65785	NA	1196,333	5971,667	17673,33	4992,667

In the table 2 it is shown how data from different sources are combined. Variables denoted by P1003162, P1003162… represent turnovers form monthly survey on industry. Variables denoted by P002, P003… represent density of traffic from traffic sensors data. Such data is than used (after normalisation) in procedures of determination of principal components.

3. BUSINESS CASE FOR SGA-2

3.1 Early estimates of economic indicators:

During the SGA1 period WP6 team explored big data and other sources which could be combined for purposes of early estimates and conduct two pilots at Statistics Finland and Statistical Office of the Republic of Slovenia based on early estimates of some economic indicators. Statistics Finland testes a series of shrinkage and factor analytic methodologies to compute nowcasts of the main Finnish turnover indexes, using continuously accumulating firm-level data. They showed that the estimates based on large dimensional models provide an accurate and timelier alternative to the ones produced currently by Statistics Finland, even after taking into account data revisions. In particular, it was found that the turnovers for some economic sectors could be estimated with high accuracy five days after the reference month has ended, giving more accurate and faster predictions compared to the first official internal release. Statistical Office of the Republic of Slovenia worked on PCA model where the sharable application was created and tested on real industry indices where promising

nowcasting results were obtained. This method was also tested on some big data sources such are online job vacancy data.

Moreover all of the counties involved in WP6 (Finland, Poland, Netherlands, Slovenia) and some of the other NSIs involved in ESSnet Big Data projects expressed quite high interests for earlier estimates of main economic indicators produced at NSIs. Due to the expressed interest for investigation in area of early estimates and the fact that many big data sources could be associated to early economic indicators it was decided to propose the pilot on early estimates of economic indicators.

The aim of the pilot is to investigate big data and other existing sources for calculating flash and (or) intermediate estimates of economic indicators. Early estimators of economic indicators which will be considered are:

- Gross domestic product (GDP)
- Consumer price index (CPI)
- Retail sale
- Balance of payments
- Economic sentiment indicators

Some work will also be dedicated to exploring possible new leading economic indicators.

During the conducting of the pilot the correlation of the data sources and early economic indicators is planned to be explored and according to the results (detected combining sources and testing early economic indicator), various models for flash and (or) intermediate estimates will be tested. The most promising estimator is GDP, but the pilot will not limit itself to GDP due to the fact that results of analysing data sources could propose calculation of (better) estimates of other economic indicators.

3.2 Data sources

Many big data, statistical and other administrative sources could be linked to early economic indicators. As one of the results of SGA1 the list of possible sources which could be combined for purposes of estimates of early economic indicators was prepared (Table). Some of the sources have been already investigated for some of them there is issue with their accessibility. It should be also taken into account availability of time series of certain data source due to our goal to nowcast of economic indicators.

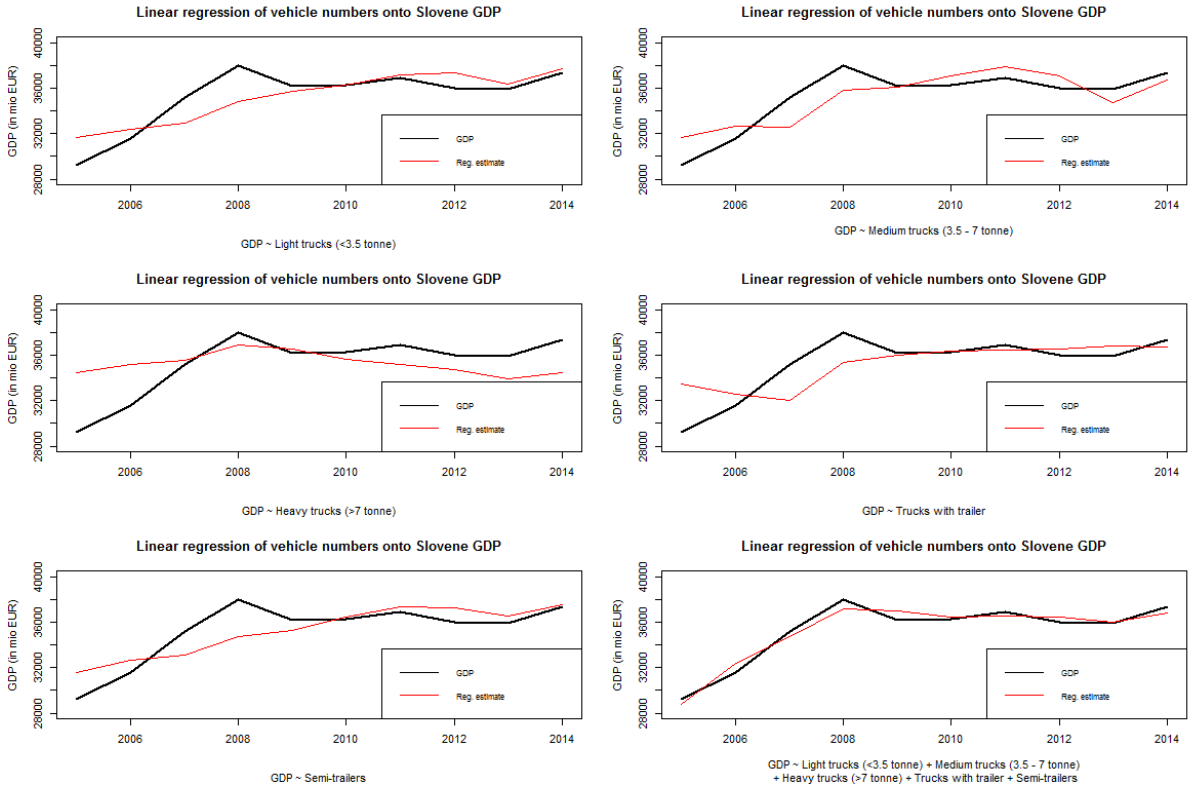
Table 3: overview of possible sources to be investigated

Big Data	**Registers and existing sources**	**Surveys**
Job vacancies ads from job portals	Statistical Register of Employment	Turnover data from various short-term surveys
Traffic loops	Data from the Employment	Business confidence index

	Agency	
Social data (Twitter, Facebook, etc.)	Tax data	Consumer confidence index
Supermarket scanner data	Wages and salaries	…..
News feeds/messages	…	
Bank transaction data		

One of the data sources which could be easiest acquired is traffic sensor data. Additional advantage is availability of times series of traffic data which is not the case for many other big data sources. First results which have been obtained at Statistical Office of the Republic of Slovenia (Image 1) shows quite a fit between curves which show movement of annual GDP and movement of estimates of annual GDP based of various annual aggregates of traffic density in Slovenia using data in period 2005-2014. Model used for estimation was simple linear regression.

Image 1: Estimates of annual GDPs using traffic sensors data



At the image 1 it could be seen 5 examples of estimate of annual GDP due to the aggregated categories of vehicles on Slovenian roads. Categorises which have been used are

- Light trucks (up to 3,5 T)
- Medium trucks (3,5 -7 T)

- Heavy trucks (more than 7 T)
- Trucks with the trailer
- Semi-trailers

At the example at the bottom right all categories of vehicles were used as a regressor. Surprisingly the best results were obtained where all vehicles are taken into account. However based on initial encouraging results more detail analysis of traffic loops data (most of work planned for SGA2) has been started.

After some research it has been found that the data can be acquired from the Slovenian Ministry of infrastructure (and municipalities for local traffic. They gave us multiple choices for the format of data and they also provided us with sample of micro data. Samples of raw data were dived on row and "so called" edited data. Row data presents counted categories of counted vehicles per traffic loop while edited data represents time series of data of one or more traffic loops which were places at the same location point.

Raw data

The data is raw data from every traffic sensor placed on the Slovenian roads. As there exist 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 4:

Table 4: Categories of detected vehicles by sensors

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

Sensors have some common features. Every sensor counts traffic on 2 channels, this 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 and it is 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.

Traffic sensors

In 2015 there were 659 sensors in Slovenia which were not manual.

On the website promet.si there is information about traffic sensors (https://www.promet.si/portal/sl/stevci-prometa.aspx, 26.1.2017) which give you on the fly information about current traffic situation. Those traffic sensors covered all highways and other roads in Slovenia as well. There is also available the map of all traffic loops in Slovenia which allows us to geo locate their exact location

(http://www.di.gov.si/fileadmin/di.gov.si/pageuploads/Prometni_podatki/2015_karta_stm.pdf)

Table 5: Legend of the description of the record for one of the traffic sensors

Variable	Description
STM	**ID of location of traffic loop**
ID	**ID of traffic loop**
CAS	**Date (yyyy/dd/mm)**
A01	**Channel 1** Motorcycles
A11	**Channel 1** Cars (also with trailer)
A21	**Channel 1** Vans (also with trailer)
B11	**Channel 1** Light trucks
B21	**Channel 1** Medium trucks
B31	**Channel 1** Heavy trucks
B41	**Channel 1** Heavy trucks with trailer
B51	**Channel 1** Vehicles with semi-trailer
C11	**Channel 1** Buses
C21	**Channel 1** Busses with trailer
XX1	**Channel 1** Unknown vehicles
VA1	**Channel 1** Average speed
VAVG1	**Channel 1** Average speed in 10 min interval
VMIN1	**Channel 1** Lowest speed in 10 min interval
VMAX1	**Channel 1** Highest speed in 10 min interval
S01- S151	**Channel 1** Peed classes
GAP1	**Channel 1** Average distance between two vehicles
OCC1	**Channel 1** Traffic density
A02	**Channel 2** Motorcycles
A12	**Channel 2** Cars (also with trailer)
A22	**Channel 2** Vans (also with trailer)
B12	**Channel 2** Light trucks
B22	**Channel 2** Medium trucks
B32	**Channel 2** Heavy trucks
B42	**Channel 2** Heavy trucks with trailer
B52	**Channel 2** Vehicles with semi-trailer
C12	**Channel 2** Buses
C22	**Channel 2** Busses with trailer
XX2	**Channel 2** Unknown vehicles
VA2	**Channel 2** Average speed
VAVG2	**Channel 2** Average speed in 10 min interval
VMIN2	**Channel 2** Lowest speed in 10 min interval
VMAX2	**Channel 2** Highest speed in 10 min interval
S02- S152	**Channel 2** Peed classes
GAP2	**Channel 2** Average distance between two vehicles
OCC2	**Channel 2** Traffic density
TEMP	**Temperature**

Roads

There are 12 categories of roads in Slovenia which is very important to know due to the possible influence of traffic density of some roads on economic indicators (e.g. excluding the foreign vehicles which cross the country). One of the steps in SGA2 is investigation of what type of roads and what categories of vehicles are most correlated with economic indicators.

Table 6: categories of roads

Id of the road	Category	Evidence number
AC	Motorways	A1-A9
HC	Expressways	H1-H9
G1	Main roads I	1-99
G2	Main roads II	101-199
R1	Regional roads I	201-399
R2	Regional roads II	401-599
R3	Regional roads III	601-999
LC	Local roads	001xxx-499xxx
JP	Public paths	501xxx-999xxx
LG	Main city roads	001xxx-499xxx
LZ	Collection city roads	01xxx-499xxx
LK	City/town roads	001xxx-499xxx

More detail description of the categories of roads could be found at

http://www.stat.si/StatWeb/File/DocSysFile/8025

4. CONCLUSIONS

The first aim of ESSnet Big data WP6 (SGA-1) was to find out which pilot would combine multiple (big) data sources and have a real potential to be implemented (by at least two countries) in SGA-2 period. The proposed pilot is *Early estimators of economic indicators*. The proposal of the pilot has been made and positive response from other countries shows that we are on right track. According to the plan the WP6 team worked on following deliverables:

- Detailed business plan for the pilot was prepared
- List of tasks per each country was prepared (involved in SGA- 2)
- Initial investigation of available data sources in participated counties was done

The seconed aim of the report was to display which of the two proposed pilots have greater potential in order to be implemented during the first wave of pilots. After the first few month of investigation it had been found out that pilot *NowCasting turnover indices* is much more feasible in terms of available data. Another advantage of this pilot is models for nowcasting which will be tested during the next period. The project team has seen the clear connection between those models and proposed pilots on early estimates where we could use the experiences from nowcasting the *NowCasting turnover indices.*