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 6.2

Recommendations about IT tools for collection of data for purposes of Consumer Confidence Index and NowCasts of Turnover Indices

Version 2016-16-06

Prepared by:

Henri Luomaranta (Statistics Finland)
Piet Daas (CBS, Netherlands)
Anna Nowitzka (GUS, Poland)
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..... **Fout! Bladwijzer niet gedefinieerd.**

2 Explanation of idea and data preparation……..3

3 Explanation of data processing ……5

4 Conclusions..... 17

1. INTRODUCTION

The aim of the deliverable "Recommendations about IT tools for collection of data for purposes of Consumer Confidence Index and NowCasts of Turnover Indices" was to focus mainly of collection of big data sources which could be used as solely source or combined with other data sources for nowcasting of Consumer Confidence Index and NowCasts of Turnover Indices. Big data sources meant for purposes of Consumer Confidence Index are related to social media for which it was found out that data are mostly not available for majority of countries involved in WP6.

The second aim of deliverable was to give recommendation about IT process of preparing the data for purposes of nowcasting and IT process of nowcasting of Turnover Indices itself. Due to the fact that IT process for preparing and processing the data (independently of data source) for PCA model has been established this report focuses of description of IT process of preparing the data, processing the data and obtaining the results of nowcasting.

This report is especially important for NSIs involved in SGA-2 which could test already prepared big data methods on their data sources.

2. EXPLANATION OF IDEA AND DATA PREPARATION

2.1 Idea

The idea of using the model in SAS and in R came from colleagues from Statistics Finland. Implementation of idea was made by colleagues from Statistical Office of the Republic of Slovenia.

This nowcasting model consists of 2 stages:

1. Principal Component Analysis (PCA) is used to extract principal components from enterprise data (or big data). For each enterprise included in the model, time series of data without any missing values is needed. Then, first few principal components are chosen.
2. Linear regression is used: the time series of interest (e.g. GDP) is the dependent variable (Y) and the chosen principal components are the predictors (X_1, …, X_n). Seasonal component and other predictors can be added.

There are still some possibilities to improve this model.

2.2 Enterprise data

Enterprise data are prepared using SAS.

- SAS is used to connect to the server and to take the enterprise data.
- There are two possible forms:
    - *First form*: obdobje, P… …
      A table with variable obdobje (period of time series) and many variables for separate enterprises is made. Variable obdobje has values e.g. 2008M01 (in case of monthly time series) or 2008Q1 (in case of quarterly time series), and separate variables for enterprises are named e.g. P1000002, where 1000002 means the second company of the first source.

Table 1: Monthly turnover of enterprise data

obdobje	P1000001	P1000002	…	P1000973
2008M01	3526	.	…	66519
2008M02	4252	332	…	36012
…	…	…	…	…
2015M12	5241	412	…	71025

Enterprise data variables can have missing values.

    - *Second form*: obdobje, spr, pris_datum, id, tip_pod
      A table with variables obdobje, spr, pris_datum, id, tip_pod is made. Variable obdobje has values e.g. 2008M01 (in case of monthly time series) or 2008Q1 (in case of quarterly time series). Variable spr is the variable of interest (e.g. turnover). Variable pris_datum tells us when the information became available or was received. Variable id has different values for each enterprise (e.g. P1000002, where 1000002 means the second company of the first source). Variable tip_pod can have values s (raw data) or u (edited data).
- If we have data in the second form, they are automatically transformed to the first form later in R.

- If we have monthly enterprise data and quarterly time series of interest data, the enterprise data are automatically transformed from monthly to quarterly using mean[1] later in R.

In case of using Big Data sources (micro level) data must be structured in the same way as enterprises data. The table 2 shows example where enterprise data (variables P1...) are combined with traffic sensor data (variables P0…).

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

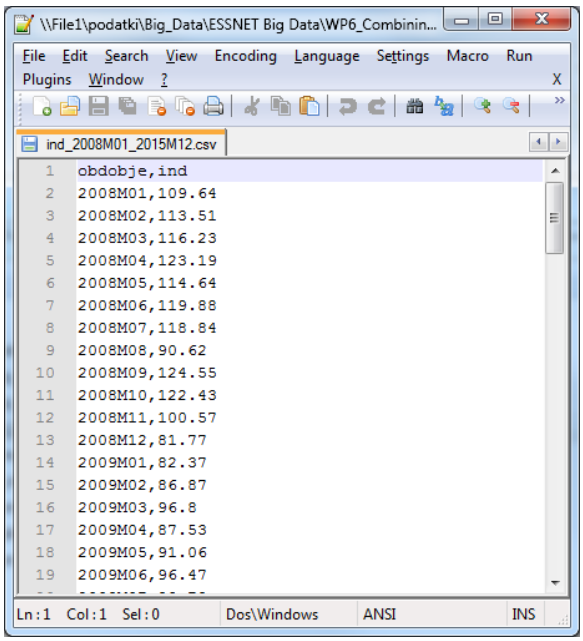
	P1003162	P1003164	P1003166	P1003168	P1003170	P002	P003	P006	P010
1	216218	44818,67	71895,67	56055,67	260288,3	1824,333	7714,333	15132	4092,667
2	245734,7	80276,67	60211	39325,67	217410,7	1419,667	10247,67	19597,67	5346,667
3	296705	47200,67	64541	75624,33	269679	1392	9770,667	19843,33	6352
4	231986,3	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

2.3 Time series of interest data

Time series of interest is the time series that we wish to estimate, nowcast. Data are prepared in csv file. It has two variables: variable obdobje has values e.g. 2008M01 (in case of monthly time series) or 2008Q1 (in case of quarterly time series), and variable e.g. ind (the name of the time series of interest) has original values of the time series. The file is stored in appropriate subfolder.

[1] If e.g. an enterprise variable has value for all the months of the quarter, their mean is the value for the quarter; if value for at least one month of the quarter is missing, then the value for the quarter is missing.

Example: ind_2008M01_2015M12.csv[2]:



2.4 Seasonal component and other predictors

Seasonal component can be used as an additional predictor in linear regression. Seasonal component is derived from time series of interest. Other predictors (eg traffic loop data on macro level) can be used as additional predictors in linear regression. They have the same form as time series of interest data.
Example: klima_2008M01_2015M12.csv

2.5 Model in R

The nowcasting model is programmed in R.

- See **obdelava_v15.R** in subfolder 3_Obdelava. Use this code to estimate (nowcast) 1 time series of interest many times (or only one time).
- See **obdelava_v15_enkrat.R** in subfolder 3_Obdelava. Use this code if you want to estimate (nowcast) 1 time series of interest only one time and you want to run separately different parts of the code to see the intermediate results (statistics, charts …).

3. EXPLANATION OF DATA PROCESSING

3.1 Introduction

[2] Only time series of interest data can have 1 period less than the other data! E.g. enterprise data and other predictors have data from January 2008 until December 2015, but time series of interest has data from January 2008 until November 2015 or December 2015.

The R code *obdelava_v("number of version").R* is used to estimate one time series of interest many times. Often, parameters can have different values and the ones that are currently not used, are commented (so when one wishes to use them it is just needed to uncomment them).
You can select different beginning (parameter zacetek), ending (parameter v_konec, v_konec_prej), microdata sources (parameter v_pod_vir), time series of interest (parameter vr_vir), condition for choosing principal components (parameter v_izberi_prve_pogoj), optimization direction in linear regression (parameter v_optim_regresija_smer), optimization condition in linear regression (parameter v_optim_regresija_pogoj), seasonal component as a predictor (parameter v_sezona) and other predictors (parameter v_dr_vir), you can take into account the date when information was available or was received (parameter t_plus) ... The names of output files can have parameter values included or the names can have other suffix (parameter izhod_pripona). All the parameters are set in point 0., then you select the whole code (Ctrl + A) and run it (Ctrl + Enter).

3.2 Time series of interest:

Could be any time series in csv format (with monthly or quarterly data)
e.g. Real turnover index in industry:
vr_vir <- "ind_2008M01_2015M12.csv"
This time series must include all the periods from parameter **zacetek** (beginning of time series) to each element of parameter v_konec (end of time series).

3.3 Data sources:

Micro data sources could be traditional (surveys…) sources or (and) big data sources. In case of big data sources, data must be structured in the same way as data from "traditional" sources.

e.g. Real turnover data for enterprises in industry (data from our sample survey; the data are already edited (imputations, etc.)
v_pod_vir <- c("podatki_vir1.sas7bdat")
Multiple data files can be chosen, e.g.
v_pod_vir <- c("podatki_vir1.sas7bdat","podatki_vir2.sas7bdat","podatki_vir3.sas7bdat")
Data files must still include NAs (because some statistics are computed to compare original data with NAs and data without enterprises with any NAs).

3.4 Principal Component Analysis:

All data that have the value for every period of a testing span (i.e. from the beginning to the end of the testing span) are included as an input in the model. Data time series/variables are standardized (mean = 0, standard deviation = 1). There are different conditions for choosing principal components parameter (parameter v_izberi_prve_pogoj) that will be used as predictors in linear regression later.

Table 3: conditions for choosing principal components

Condition	Meaning
70	take enough p. c. to explain 70% (or a bit more) of the variability of the enterprise data
75	take enough p. c. to explain 75% (or a bit more) of the variability of the enterprise data
80	take enough p. c. to explain 80% (or a bit more) of the variability of the enterprise data
85	take enough p. c. to explain 85% (or a bit more) of the variability of the enterprise data
90	take enough p. c. to explain 90% (or a bit more) of the variability of the enterprise data
po7	take only as many p. c. to have at least 7 cases (time periods) per independent variable later in the linear regression
po8	take only as many p. c. to have at least 8 cases (time periods) per independent variable later in the linear regression
po10	take only as many p. c. to have at least 10 cases (time periods) per independent variable later in the linear regression
po15	take only as many p. c. to have at least 15 cases (time periods) per independent variable later in the linear regression
po20	take only as many p. c. to have at least 20 cases (time periods) per independent variable later in the linear regression
zadnja5	take every p. c., whose eigenvalue's share among all eigenvalues is greater or equal to 5%
Kaiser	take every p. c., whose eigenvalue is greater or equal to 1

3.5 Linear regression:

Y (dependent variable) = time series of interest
X_1, …, X_n (predictors) = the chosen principal components and other predictors (parameter v_dr_vir) can also be added. For example traffic loop data could be used as additional predictor in linear model.

There are other possible optimizations:

- optimization direction (parameter v_optim_regresija_smer);
- optimization condition (parameter v_optim_regresija_pogoj).

3.6 Different testing spans:

Each testing span begins at zacetek. Each testing span has its own ending: one of the elements in v_konec.

e.g. a set of 36 testing spans in case of enterprise data:

- the first period is always 2008M01
- the last period is 2013M01 or 2013M02 or … or 2015M12

Time series of interest might not be available for the last period of the testing span; in this case we cannot estimate the difference between the estimate and the original value. But enterprise data must be available for the whole testing span.

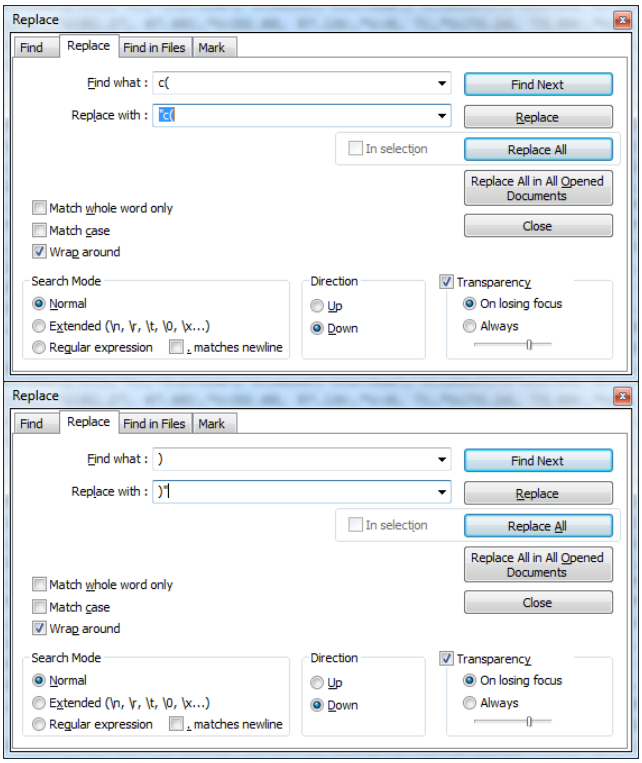
Time series of enterprises without any NAs for a given testing span (i.e. from zacetek to a given element of v_konec) are chosen (balanced method). Their data from zacetek to a given element of v_konec are used to extract principal components. The chosen principal components without the ending (i.e. given element of v_konec) are used for the model. Coefficients from this model and the ending of principal components are used to calculate the estimate of the time series of interest for the ending.

3.7 Results:

The results are presented in 3 output files.
The *basic* name: OBDELAVA_*version_time-series-of-interest_enterprise-data_conditions-for-pca_directions-for-linear-regression_conditions-for-linear-regression_predictor-for-seasonality_test-linear-regression-assumptions_other-predictors* (e.g. OBDELAVA_v15_ind_2008M01_2015M12_1_70-75_backward-forward_AIC_stl_NO_klima-ind).

- The first file: *basic*.csv.
  We can see the results for every time span, every condition for choosing principal components, every condition and possible direction for linear regression, and every possibility for predictor for seasonality.
  Test for linear regression assumptions is made or not (YES or NO). If we decided for YES, there might be some errors because of this test, so not all the results are saved. If we decided for NO, all the results are saved.
- The second file: *basic*_skupna.csv.
  We can see common results for all the results in *basic*.csv.
- The third file: *basic*_skupna2.csv.
  We can see common results for each set of testing spans with different conditions in *basic*.csv.

Before opening this file in Excel, It is recommended that one to make a copy (name it e.g. *basic*_skupna2 - C.csv) and make the following replacements (double quotes have to be around an interval c(number1,number2) so this interval will be in the same cell):



After the replacements, you can open csv in Excel and use Text to columns option with the comma delimiter.

The meaning of variables in the first output file: *basic*.csv:

Variable(s)	**Meaning**
verzija	Version of code
pot1, pot2, pot3, pot4, pot5	(Parts of) paths
zacetek	The first period in a testing span
konec	The last period in a testing span
konec_prej	The period before the last period in a testing span
v_pod_vir_ref	The names of enterprise data files; the names are separated by hyphen
v_tip_ref	The types of enterprise data; the types are separated by

	hyphen
podatki_vir_ref	The reference string for enterprise data files
t_plus	Time since the end of reference period
vr_vir	The time series of interest data file
izberi_prve_pogoj	The condition for choosing principal components
optim_regresija_smer	The direction in function step in optimization of linear regression (backward, forward or both)
optim_regresija_pogoj	The condition for optimization of linear regression (AIC, BIC or NO (i.e. no optimization))
sezona	The option to include seasonal component as predictor (»stl«) or not (»no«).
p_sezona	If seasonal component is included as predictor, this is the p-value of its regression coefficient in regression before optimization.
v_dr_vir_ref	The names of data files of other predictors included in linear regression before optimization; the names are separated by hyphen
drugi_ref	The names of other predictors included in linear regression before optimization; the names are separated by hyphen
p_drugi_s	P-values of other predictors included in linear regression before optimization; p-values are separated by hyphen
p_drugi_d	Share of p-values of other predictors included in linear regression before optimization, that are smaller than 0.05 (in %)
per	Periodicity of enterprise data (e.g. 12 for monthly time series)
l_podatki	Length of time series for each enterprise
l_vrsta	Length of time series of interest
st_podj_konec	Number of enterprises that have data for the period konec in the source file
st_podj_podatki	Number of enterprises used for the model (these enterprises have data for the period konec after removing enterprises with any NAs)
perc_st_podj	Share of number of enterprises used for the model (in %)
perc_vsota_podj	Share of value of enterprises used for the model (in %)

n_pca	Number of principal components chosen
perc_pca	How much variability of enterprise data do the chosen principal components explain (in %)
n_pred	Number of predictors after optimization of linear regression
rsqadj1	Adjusted R squared for linear regression before optimization
rsqadj	Adjusted R squared for linear regression (i.e. after optimization)
napake_maxabs	The maximum of absolute values of errors (i.e. minus residuals) in linear regression
napake_meanabs	The average of absolute values of errors in linear regression
napake_stod	The standard deviation of errors in linear regression
rel_napake_maxabs	The maximum of absolute values of relative errors (i.e. minus residuals) in linear regression (in %)
rel_napake_meanabs	The average of absolute values of relative errors in linear regression (in %)
rel_napake_stod	The standard deviation of relative errors in linear regression (in %)
vrednost	The original value of time series of interest for the last period of testing span
napoved	The estimate for the original value of time series of interest for the last period of testing span
napaka	The error of estimate for the original value of time series of interest for the last period of testing span (napaka = napoved – vrednost)
napaka_abs	The absolute value of error of estimate for the original value of time series of interest for the last period of testing span
rel_napaka	The relative error of estimate for the original value of time series of interest for the last period of testing span (in %)
rel_napaka_abs	The absolute value of relative error of estimate for the original value of time series of interest for the last period of testing span (in %)
vrednost_rast_1	Growth of original value of time series of interest for the last period of testing span to the original value of the previous period (in %)
vrednost_rast_per	Growth of original value of time series of interest for the last period of testing span to the original value of the same period

	of the previous year (in %)
napoved_rast_1	Growth of estimate of time series of interest for the last period of testing span to the original value of the previous period (in %)
napoved_rast_per	Growth of estimate of time series of interest for the last period of testing span to the original value of the same period of the previous year (in %)
rast_1_absraz	Absolute value of difference between vrednost_rast_1 and napoved_rast_1 (in percentage changes)
rast_per_absraz	Absolute value of difference between vrednost_rast_per and napoved_rast_per (in percentage changes)
test_predp2	Global Stat of global test of model assumptions (function glvma in package glvma)
p_dw	p-value of Durbin-Watson normality test (function dwtest in package lmtest) (if p-value < 0.05, we can reject independence of residuals in linear regression)
p_dw_1	p-value of Durbin-Watson normality test (function durbinWatsonTest in package car) at lag 1 (if p-value < 0.05, we can reject independence of residuals in linear regression at lag 1)
p_dw_per	p-value of Durbin-Watson normality test (function durbinWatsonTest in package car) at lag per (if p-value < 0.05, we can reject independence of residuals in linear regression at lag per)
p_shapiro	p-value of Shapiro-Wilks normality test (function shapiro.test in package stats) (if p-value < 0.05, we can conclude that the residuals in linear regression deviate from normality)
p_ks	p-value of Kolmogorov-Smirnov normality test (function ks.test in package stats) (if p-value < 0.05, we can conclude that the residuals in linear regression deviate from normality)

The meaning of variables in the second output file: *basic*_skupna.csv:

Variable	**Meaning** (all variables whose name starts with diag are rounded to 2 or 4 digits)

verzija	Version of code
pot1, pot2, pot3, pot4, pot5	(Parts of) paths
zacetek	The first period in all testing spans
zadnji_konec	The last period of the last testing span
st_koncev	Number of testing spans
v_pod_vir_ref	The names of enterprise data files; the names are separated by hyphen
v_tip_ref	The types of enterprise data; the types are separated by hyphen
podatki_vir_ref	The reference string for enterprise data files
t_plus	Time since the end of reference period
vr_vir	The time series of interest data file
izberi_prve_pogoj	The conditions for choosing principal components; the names are separated by hyphen
optim_regresija_smer	The direction in function step in optimization of linear regression (backward, forward, both, or /); the names are separated by hyphen
optim_regresija_pogoj	The conditions for optimization of linear regression (AIC, BIC or NO (i.e. no optimization)); the names are separated by hyphen
sezona	The options to include seasonal component as predictor (»stl«) or not (»no«); the names are separated by hyphen
v_dr_vir_ref	The names of data files of other predictors included in linear regression before optimization; the names are separated by hyphen
drugi_ref	The names of other predictors included in linear regression before optimization; the names are separated by hyphen
per	Periodicity of enterprise data (e.g. 12 for monthly time series)
diag_st_podj_podatki_range	Minimum and maximum (considering all simulations) number of enterprises used for the model
diag_perc_st_podj_range	Minimum and maximum (considering all simulations) share of number of enterprises used for the model (in %)
diag_perc_vsota_podj_range	Minimum and maximum (considering all simulations) share of

	value of enterprises used for the model (in %)
diag_n_pca_range	Minimum and maximum (considering all simulations) number of chosen principal components
diag_perc_pca_range	Minimum and maximum (considering all simulations) variability of enterprise data explained by the chosen principal components (in %)
diag_n_pred_range	Minimum and maximum (considering all simulations) number of predictors after optimization of linear regression
diag_rsqadj_range	Minimum and maximum (considering all simulations) adjusted R squared for linear regression (i.e. after optimization)
diag_napake_maxabs	The maximum (considering all simulations) absolute value of errors (i.e. minus residuals) in linear regression
diag_rel_napake_maxabs	The maximum (considering all simulations) absolute value of relative errors (i.e. minus residuals) in linear regression (in %)
diag_napake_meanabs	The average (considering all simulations) of average of absolute values of errors in linear regression
diag_rel_napake_meanabs	The average (considering all simulations) of average of absolute values of relative errors in linear regression (in %)
diag_napaka_maxabs	The maximum (considering all simulations) of absolute values of errors of estimate for the original value of time series of interest for the last period of testing span
diag_rel_napaka_maxabs	The maximum (considering all simulations) of absolute values of relative errors of estimate for the original value of time series of interest for the last period of testing span (in %)
diag_napaka_meanabs	The average (considering all simulations) of absolute values of the errors of estimate for the original value of time series of interest for the last period of testing span
diag_napaka_meansq	The average (considering all simulations) of squared values of the errors of estimate for the original value of time series of interest for the last period of testing span
diag_rel_napaka_meanabs	The average (considering all simulations) of absolute values of relative errors of estimate for the original value of time series of interest for the last period of testing span (in %)
diag_rel_napaka_meansq	The average (considering all simulations) of squared values of relative errors of estimate for the original value of time series of interest for the last period of testing span (in squared %)
diag_d_napake_maxabs	Share of time spans (considering all simulations) for which absolute value of error of estimate for the original value of

	time series of interest for the last period in smaller or equal to maximum absolute value of errors in linear regression; and 1 minus this share
diag_d_napake_meanabs	Share of time spans (considering all simulations) for which absolute value of error of estimate for the original value of time series of interest for the last period in smaller or equal to average of absolute values of errors in linear regression; and 1 minus this share
diag_d_napake_stod	Share of time spans (considering all simulations) for which absolute value of error of estimate for the original value of time series of interest for the last period in smaller or equal to the standard deviation of errors in linear regression; and 1 minus this share
diag_rast_1_absraz_range	Minimum and maximum (considering all simulations) absolute value of difference between vrednost_rast_1 and napoved_rast_1 (in percentage changes)
diag_rast_1_absraz_mean	The average (considering all simulations) of absolute value of difference between vrednost_rast_1 and napoved_rast_1 (in percentage changes)
diag_rast_1_absraz_meansq	The average (considering all simmulations) of squared value of difference between vrednost_rast_1 and napoved_rast_1 (in squared percentage changes)
diag_rast_per_absraz_range	Minimum and maximum (considering all simulations) absolute value of difference between vrednost_rast_per and napoved_rast_per (in percentage changes)
diag_rast_per_absraz_mean	The average (considering all simulations) of absolute value of difference between vrednost_rast_per and napoved_rast_per (in percentage changes)
diag_rast_per_absraz_meansq	The average (considering all simulations) of squared value of difference between vrednost_rast_per and napoved_rast_per (in squared percentage changes)
diag_d_test_predp2	Share of time spans (considering all simulations) for which Global Stat of global test of model assumptions (function glvma in package glvma) is equal to »Assumptions acceptable.«; and 1 minus this share
diag_d_p_dw	Share of time spans (considering all simulations) for which p-value of Durbin-Watson normality test (function dwtest in package lmtest) is greater or equal to 0.05;

	and 1 minus this share
diag_d_p_dw_1	Share of time spans (considering all simulations) for which p-value of Durbin-Watson normality test (function durbinWatsonTest in package car) at lag 1 is greater or equal to 0.05; and 1 minus this share
diag_d_p_dw_per	Share of time spans (considering all simulations) for which p-value of Durbin-Watson normality test (function durbinWatsonTest in package car) at lag per is greater or equal to 0.05; and 1 minus this share
diag_d_p_shapiro	Share of time spans (considering all simulations) for which p-value of Shapiro-Wilks normality test (function shapiro.test in package stats) is greater or equal to 0.05; and 1 minus this share
diag_d_p_ks	Share of time spans (considering all simulations) for which p-value of Kolmogorov-Smirnov normality test (function ks.test in package stats) is greater or equal to 0.05; and 1 minus this share
diag_d_p_sezona	Share of time spans (considering all simulations) for which p-value of seasonal component's predictor's coefficient is smaller than 0.05; and 1 minus this share
diag_p_drugi_d_mean	Average of shares of p-values (considering all simulations) of other predictors included in linear regression before optimization, that are smaller than 0.05 (in %)

The meaning of variables in the third output file (*basic*_skupna2.csv) is the same as in the second (*basic*_skupna.csv), but it is calculated for smaller groups (i.e. all time spans that have the same conditions and parameters …).

4. CONCLUSSIONS

This report gives a deep (IT) insight how to use workable application based on PCA method for purposes of nowcasting early economic indicators. Although methods for nowcasting early statistics have been known for a long time we can consider them as a “big data methods” due to the fact that we started to employ them when we faces with big data (time series) data.

Although there is a range of methods which could be used for purposes of nowcasting it is not easy task to prepare IT solution in such a way that it could be used by other statistical institute. One of the main results of whole work in WP6 is IT application which is ready to use for every NSI which is interested in this topic. Possible ways of incorporation of big data sources into the application is explained in the deliverable D1. However it has not to be forgotten that the time series of big data is needed in order to employ this kind of sources in nowcasting models. There are not many big data sources for which we could have an access to data also in pass. One of the big data sources for which data in the past is available is traffic sensor data. The plan for SGA-2 is to incorporate this data in the PCA model and test the nowcasting of early economic indicators.