INTRODUCING BIG DATA IN OFFICIAL STATISTICS: RESULTS, RECOMMENDATIONS AND STRATEGIC COOPERATION

By: **Gian Luigi Mazzi**

Technical Director of GOPA Luxembourg

Gianluigi.mazzi@gopa.lu

CIRET pre-Conference Workshop
Big Data for Economic Statistics:
Challenges and Opportunities
Tuesday, 11 September 2018, Rio de Janeiro, Brazil

Introduction I

- High potential of big data to enhance statistical production
  - Not fully explored and assessed
  - Still open debate on the effective usefulness of big data
- Recent years characterized by a quickly growing up literature on big data continuously and exponentially growing up
  - in the context of the SDGs
  - for the construction of new indicators, smart statistics;
  - to enhance nowcasting and short-term forecasting of key macroeconomic infr-annual indicators
- Promising results obtained but still not conclusive evidence on the benefits from the use of big data in official statistics: some open conceptual issues remain
  - big data as a complement of an alternative to traditional sources
  - what kind of big data should be used and when
- Contribution of this paper in the field of nowcasting and short-term forecasting

Introduction II

Defining a clear typology of big data with clearly defined characterisitics.

	Type	Main Utilisation
1	Financial market data	Macroeconomics, financial sector monitoring
2	Electronic payments data	Macroeconomics, inflation, consumers behaviour
3	Mobile phone data	Labour market, sustainable development
4	Sensor data and the Internet Of Things	Sustainable development, urban monitoring
5	Satellite image data	Sustainable development, economic growth and land utilisation
6	Scanner prices data	Macroeconomics, inflation, consumers behaviour
7	Online prices data	Macroeconomics, inflation, consumers behaviour
8	Online search data	Macroeconomics, sustainable development, human behaviour
9	Textual data	Human sentiments, confidence, uncertainty
10	Social media data	Macroeconomics, sustainable development, human behaviour

Introduction III

- big data limits and accessibility problems
- most of big data sources are available on request and not for free
  - difficult to obtain consistent time-series over the time
  - lack of stability of big data production making problematic a regular update of some kind of big data
- limited time coverage of several big data times
  - regular collection process started only recently
  - structural changes characterizing the collection process and creating significant breaks in the data
- Difficult utilization of big data available only on a relatively short time-span
  - mainly in a nowcasting and short- term forecasting exercise;
  - non-significant simple size and consequently non-conclusive inference
- restricting the big data set used in this paper only to two times:
  - Google trend data;
  - Reuters data

Description of the exercise 1

- Large scale pseudo-real-time nowcasting/forecasting exercise
- main objective: estimating present and future evolution of target variables
  - key macroeconomic indicators
  - using real-time vintages for the target variables from the Eurostat historical database
- using a large number of predictors from different official and non official data sources;
- identification and estimation of several competing models and model specifications
  - Bayesian and non-Bayesian ones
- Comparative evaluation of the performance based on statistical criteria
  - reflecting forecasting accuracy

Data issues I

- Choice of the target variables based on their relevance for short-term economic analysis : monthly and quarterly variables
- Selected target variables
  - Industrial production Index IPI, monthly seasonal and calendar adjusted
  - Unemployment rate UR monthly seasonally adjusted
  - Harmonized index of consumer prices HICP monthly non seasonally adjusted
  - Gross domestic product GDP quarterly seasonally and calendar adjusted
- Further considerations on the indicators chosen
  - different production and associated revision process
  - different degree of volatility
    - IPI also chosen due to its high volatility which complicates often the nowcasting exercise
- Country coverage: for EU largest economies
  - France, Germany, Italy and UK
  - good proxy of the EU and EA: more than 80% coverage

Data issues II

- Large number of predictor selected from various data sources
  - Official statistics
  - Financial indicators;
  - big data
- data available and different frequency and with different timing:
  - quarterly, monthly and weekly data
  - ragged aged exercise taking into account the release calendar of each predictor
- both seasonally and non-seasonally adjusted data following data characteristics and availability
- big data considered expressed in a structured time-series form

Data issues III

Monthly predictors

Monthly Predictor	Seasonal Treatment	Monthly Predictor	Seasonal Treatment
Bank Lending Rate	Non-seasonally adjusted	House Price Index	Non-seasonally adjusted
Bankruptcies	Non-seasonally adjusted	Import Prices	Non-seasonally adjusted
Building Permits	Seasonally adjusted	Imports	Seasonally adjusted
Capital Flows	Non-seasonally adjusted	Job Vacancies	Seasonally adjusted
Car Registrations	Seasonally adjusted	Manufacturing Production	Seasonally and calendar adjusted
Construction Output	Seasonally and calendar adjusted	Mining Production	Seasonally and calendar adjusted
Consumer Credit	Non-seasonally adjusted	Money Supply M1, M2 and M3	Seasonally adjusted
Core Consumer Prices	Non-seasonally adjusted	New Orders	Seasonally adjusted
various CPI components	Non-seasonally adjusted	Private Sector Credit	Non-seasonally adjusted
Crude Oil Production	Seasonally adjusted	Producer Prices	Non-seasonally adjusted
Export Prices	Non-seasonally adjusted	Steel Production	Seasonally and calendar adjusted
Exports	Seasonally adjusted	Youth Unemployment Rate	Seasonally adjusted
Factory Orders	Seasonally adjusted	Gasoline Prices	Non-seasonally adjusted
Consumer Confidence Indicators and various surveys	Seasonally adjusted		

Data issues IV

- Weekly financial indicators
  - Interest Rates
    - Various maturities and spread
    - Equity indexes
    - Volatility indexes
- Big data based uncertainty indicators
  - Reuters news based uncertainty indicator
  - Google trend searches based uncertainty indicator

Data issues V

- Target variables transformed in order to achieve the second-order stationarity meaning no trend in mean and in variance
  - IPI period on period growth rate
  - UR period on period difference
  - HICP period on period growth rate
  - GDP period on period growth rate
- Logarithmic transformation not needed for UR since already stationary in variance
- Predictors subject to the best appropriate transformation consistent with the one applied to the target variable to which they are referring to
  - usually period on period growth rates for official statistics predictors
  - usually no transformation for financial data and big data

Uncertainty indices I

- Constructing the Reuters Uncertainty index using the Reuters News database:
- Using the following keyword by country

Country	Keywords 1	Keywords 2
Germany	uncertainty uncertain, uncertainty, uncertainties	Germany, German, Germans
France	uncertainty uncertain, uncertainty, uncertainties	France, French
Italy	uncertainty uncertain, uncertainty, uncertainties	Italy, Italian, Italians
UK	uncertainty uncertain, uncertainty, uncertainties	UK, Britain, British, United Kingdom, Briton

Uncertainty indices II

- Using Google trend data to construct uncertainty and risk index
- Using the following key words by country.

Country	Keywords 1
Germany	"unsicherheit" and "risiko"
France	"incertitude" and "risque"
Italy	"incertezza" and "rischio"
UK	"uncertainty" and "risk"

Modelling strategy I

- Enhancing the relevance of this nowcasting exercise by using a large number of models and model specifications.
- Univariate vs multivariate models
  - Univariate models mainly used as a benchmark
- Focusing on models specifically adapted to deal with large number of data and variables
  - Shrinkage regression
  - Sparse factor models
- Classical vs Bayesian models

Modelling strategy II

Alternative models used in the nowcasting exercise.

Models	Number of specifications	Description
Naive and AR models	7	Random walk and various AR specifications
Simple Linear Regression	6	models with up to 3lags of the target variables and only Reuter and Google uncertainty indexes as regressors
Various univariate models	8	autoarima, exponential smoothing, neural networks, etc.
DFA	16	4 different settings are used and 4 alternative data configurations
PLS	20	up to 5 factors using 4 different data configurations
SPC	20	up to 5 factors using 4 different data configurations
LASSO	8	class of machine learning methods to estimate penalized regressions using 4 data configurations
Spike and Slab	4	another class of machine learning methods to estimate penalized regressions
Data-Driven Automated Forecasting Strategies	4	choosing the model with the smallest cumulative nowcasting error and then use an averaging procedure with equal weights for the top 3, 5 and 10 models

Modelling strategy III

- Beside the pure big data configuration used in the Simple Regression Model, 4 data configuration are used in the nowcasting exercise
  - Macroeconomic and Financial indicators only (MacroFin)
  - Macroeconomic and Financial indicators + Reuters uncertainty Index (MacroFin-Reuters)
  - Macroeconomic and Financial indicators + Google uncertainty Index (MacroFin-Google)
  - Macroeconomic and Financial indicators + Reuters and Google uncertainty Indices (MacroFin-GoogleReuters)
- A total of 93 models and model combinations have been estimated
  - for each variable
  - For each country
  - Total 372 models
- Extended comparative analysis carried out based on measures of forecasting accuracy
  - Mean Absolute Error (MAE)
  - Root Mean Square Forecasting Error (RMSFE)
  - Predictive Accuracy Test of Diebold and Mariano (DM)

Nowcasting exercise

- Our timespan from January 2007 to October 2016 (118 months)
- Nowcasting exercise from January 2014
  - 34 evaluation periods from January 2014 to October 2016
  - 12 evaluation periods for quarterly data
- Computation of the nowcasting
  - Step 1: truncating the sample in January 2014 to leave 34 evaluation period for monthly data and 12 for quarterly data
  - Step 2: computing nowcasting for h = (-5, -4, -3, -2, -1) weeks before the official release updating the information every week.
  - Step 3: repeating Step 2 recursively for each official release date.

Empirical evidence: Uncertainty Indices

- Preliminary testing on the usefulness of big data uncertainty indices in nowcasting exercise
- Performing a simple regression analysis with Eurostat vintages as dependent variables and the uncertainty indices as predictor
  - IPI 2005=100
  - HICP 2005=100
  - Unemployment 1000 person
  - Using the last 4 vintages closer to the official release
- For all considered countries, both indicators perform well in the contemporaneous and in one-step ahead now-cast when the target variables are HICP and Unemployment
- Uncertainty indicators perform less well in the case of IPI

Empirical evidence: Uncertainty Indices II

DE, Industrial Production										
		GOOGLE		REUTERS			GOOGLE		REUTERS	
		Estimate	P-value	Estimate	P-value		Estimate	P-value	Estimate	P-value
		Contemporaneous Regression					Predictive Regression			
$r = 1$	α	89.147	0.000	92.776	0.000	α	99.204	0.000	94.882	0.000
	β	0.331	0.004	131.911	0.000	β	0.120	0.309	114.959	0.001
	R^2_{Adj}	0.145		0.309		R^2_{Adj}	0.001		0.208	
$r = 2$	α	89.996	0.000	93.379	0.000	α	99.523	0.000	95.477	0.000
	β	0.316	0.007	127.009	0.000	β	0.116	0.327	109.677	0.001
	R^2_{Adj}	0.131		0.286		R^2_{Adj}	0.000		0.186	
$r = 3$	α	90.057	0.000	91.585	0.000	α	99.358	0.000	94.006	0.000
	β	0.316	0.007	151.525	0.000	β	0.121	0.314	130.271	0.000
	R^2_{Adj}	0.130		0.354		R^2_{Adj}	0.001		0.237	
$r = 4$	α	90.495	0.000	90.951	0.000	α	99.224	0.000	94.191	0.000
	β	0.302	0.011	159.873	0.000	β	0.119	0.322	126.171	0.001
	R^2_{Adj}	0.138		0.355		R^2_{Adj}	0.222		0.224	
DE, Harmonised CPI										
		GOOGLE		REUTERS			GOOGLE		REUTERS	
		Estimate	P-value	Estimate	P-value		Estimate	P-value	Estimate	P-value
		Contemporaneous Regression					Predictive Regression			
$r = 1$	α	102.005	0.000	108.248	0.000	α	101.948	0.000	108.216	0.000
	β	0.204	0.000	40.618	0.000	β	0.207	0.000	41.992	0.000
	R^2_{Adj}	0.293		0.129		R^2_{Adj}	0.313		0.144	
$r = 2$	α	102.022	0.000	108.126	0.000	α	102.025	0.000	108.137	0.000
	β	0.203	0.000	41.514	0.000	β	0.205	0.000	42.351	0.000
	R^2_{Adj}	0.293		0.137		R^2_{Adj}	0.303		0.148	
$r = 3$	α	102.088	0.000	108.169	0.000	α	102.013	0.000	108.140	0.000
	β	0.202	0.000	40.999	0.000	β	0.205	0.000	42.333	0.000
	R^2_{Adj}	0.283		0.134		R^2_{Adj}	0.309		0.148	
$r = 4$	α	102.181	0.000	107.959	0.000	α	102.042	0.000	107.995	0.000
	β	0.198	0.000	42.281	0.000	β	0.203	0.000	42.836	0.000
	R^2_{Adj}	0.286		0.147		R^2_{Adj}	0.000		0.157	
DE, Unemployment in Thousands of Persons										
		GOOGLE		REUTERS			GOOGLE		REUTERS	
		Estimate	P-value	Estimate	P-value		Estimate	P-value	Estimate	P-value
		Contemporaneous Regression					Predictive Regression			
$r = 1$	α	3524.278	0.000	2884.080	0.000	α	3440.697	0.000	2876.751	0.000
	β	-20.790	0.000	-4232.205	0.000	β	-19.065	0.000	-4176.868	0.001
	R^2_{Adj}	0.220		0.093		R^2_{Adj}	0.184		0.090	
$r = 2$	α	3564.526	0.000	2929.120	0.000	α	3501.269	0.000	2907.059	0.000
	β	-21.391	0.000	-4649.320	0.000	β	-20.082	0.000	-4405.782	0.000
	R^2_{Adj}	0.231		0.113		R^2_{Adj}	0.204		0.101	
$r = 3$	α	3632.838	0.000	2942.509	0.000	α	3523.388	0.000	2939.277	0.000
	β	-22.742	0.000	-4778.569	0.000	β	-20.505	0.000	-4762.117	0.000
	R^2_{Adj}	0.254		0.116		R^2_{Adj}	0.203		0.115	
$r = 4$	α	3657.279	0.000	2994.324	0.000	α	3596.655	0.000	2960.431	0.000
	β	-22.919	0.000	-5192.654	0.000	β	-21.801	0.000	-5038.766	0.000
	R^2_{Adj}	0.264		0.140		R^2_{Adj}	52.756		0.131	

Empirical assessment : IPI

- Discordant results across countries observed when using IPI as a target variable
- Univariate models outperforming the others for Germany and France.
  - Confirming difficulties in nowcasting IPI
- Dealing with Italy, the LASSO model with Uncertainty Indices are among the top performing model along with univariate methods
- In the case of UK, the SPC5 with MicroFin-GoogleReuters returns improved now-casts across all *h.*

Empirical assessment : IPI

Model	-5w	-4w	-3w	-2w	-1w	Model	-5w	-4w	-3w	-2w	-1w
Average(4)	1.225	1.225	1.223	1.214	1.214	PLS(3)-MacroFin-GoogleReuters	1.271	1.238	1.258	1.251	1.254
Average(12)	1.154	1.154	1.18	1.166	1.166	PLS(3)-MacroFin-Reuters	1.268	1.23	1.241	1.233	1.235
Average(24)	1.118	1.118	1.125	1.113	1.113	PLS(4)-MacroFin	1.233	1.216	1.244	1.232	1.235
Naive	1.928	1.928	1.891	1.897	1.897	PLS(4)-MacroFin-Google	1.254	1.243	1.258	1.246	1.249
AR(1)	1.119	1.119	1.13	1.104	1.104	PLS(4)-MacroFin-GoogleReuters	1.251	1.237	1.252	1.24	1.244
AR(4)	1.027	1.027	1.05	1.028	1.028	PLS(4)-MacroFin-Reuters	1.235	1.218	1.235	1.222	1.225
AR(AIC)	1.045	1.045	1.097	1.045	1.045	PLS(5)-MacroFin	1.215	1.144	1.233	1.226	1.232
AutoArima	1.199	1.199	1.185	1.164	1.164	PLS(5)-MacroFin-Google	1.23	1.161	1.217	1.213	1.22
ETS	0.952	0.952	1.013	0.993	0.993	PLS(5)-MacroFin-GoogleReuters	1.222	1.168	1.228	1.226	1.233
BaggedETS	0.909	0.916	0.977	0.975	0.971	PLS(5)-MacroFin-Reuters	1.219	1.152	1.219	1.214	1.221
BATS	0.956	0.956	1.017	0.997	0.997	SPC(1)-MacroFin	1.176	1.18	1.162	1.14	1.14
TBATS	0.956	0.956	1.017	0.997	0.997	SPC(1)-MacroFin-Google	1.176	1.179	1.162	1.14	1.14
NN	1.013	1.012	1.067	1.035	1.033	SPC(1)-MacroFin-GoogleReuters	1.176	1.179	1.163	1.14	1.14
Spline	1.002	1.002	1.061	1.046	1.046	SPC(1)-MacroFin-Reuters	1.177	1.181	1.162	1.14	1.14
THETA	1.108	1.108	1.114	1.094	1.094	SPC(2)-MacroFin	1.217	1.185	1.157	1.143	1.151
Google	1.337	1.352	1.185	1.167	1.167	SPC(2)-MacroFin-Google	1.225	1.189	1.158	1.144	1.147
Google-L1	1.215	1.251	1.309	1.288	1.288	SPC(2)-MacroFin-GoogleReuters	1.225	1.187	1.15	1.144	1.144
Google-L3	1.217	1.285	1.384	1.372	1.373	SPC(2)-MacroFin-Reuters	1.221	1.184	1.149	1.143	1.135
Reuters	1.182	1.173	1.162	1.141	1.141	SPC(3)-MacroFin	1.225	1.166	1.148	1.126	1.128
Reuters-L1	1.158	1.147	1.165	1.146	1.145	SPC(3)-MacroFin-Google	1.234	1.178	1.154	1.128	1.128
Reuters-L3	1.115	1.103	1.116	1.101	1.101	SPC(3)-MacroFin-GoogleReuters	1.239	1.176	1.159	1.126	1.126
DFA(2)-MacroFin	1.231	1.185	1.157	1.143	1.144	SPC(3)-MacroFin-Reuters	1.226	1.178	1.142	1.124	1.123
DFA(2)-MacroFin-Google	1.233	1.186	1.158	1.144	1.144	SPC(4)-MacroFin	1.22	1.17	1.157	1.124	1.145
DFA(2)-MacroFin-GoogleReuters	1.234	1.186	1.158	1.144	1.144	SPC(4)-MacroFin-Google	1.216	1.21	1.134	1.111	1.127
DFA(2)-MacroFin-Reuters	1.232	1.185	1.157	1.143	1.143	SPC(4)-MacroFin-GoogleReuters	1.229	1.192	1.14	1.133	1.118
DFA(3)-MacroFin	1.246	1.181	1.158	1.14	1.14	SPC(4)-MacroFin-Reuters	1.207	1.175	1.144	1.127	1.135
DFA(3)-MacroFin-Google	1.251	1.183	1.16	1.142	1.141	SPC(5)-MacroFin	1.193	1.155	1.149	1.123	1.128
DFA(3)-MacroFin-GoogleReuters	1.252	1.184	1.16	1.142	1.141	SPC(5)-MacroFin-Google	1.191	1.163	1.159	1.118	1.139
DFA(3)-MacroFin-Reuters	1.246	1.182	1.158	1.14	1.139	SPC(5)-MacroFin-GoogleReuters	1.197	1.17	1.157	1.124	1.137
DFA(4)-MacroFin	1.239	1.229	1.163	1.149	1.146	SPC(5)-MacroFin-Reuters	1.209	1.15	1.153	1.124	1.134
DFA(4)-MacroFin-Google	1.245	1.225	1.136	1.119	1.12	LASSO-MacroFin	1.089	1.081	1.005	1.036	1.011
DFA(4)-MacroFin-GoogleReuters	1.244	1.225	1.142	1.126	1.126	LASSO-MacroFin-Google	1.085	1.069	1.024	1.023	1.03
DFA(4)-MacroFin-Reuters	1.24	1.223	1.161	1.147	1.145	LASSO-MacroFin-GoogleReuters	1.085	1.09	1.014	1.036	1.012
DFA(5)-MacroFin	1.218	1.19	1.167	1.14	1.142	LASSO-MacroFin-Reuters	1.083	1.08	1.015	1.027	1.036
DFA(5)-MacroFin-Google	1.218	1.189	1.167	1.139	1.141	EN-MacroFin	1.096	1.082	1.027	1.027	1.036
DFA(5)-MacroFin-GoogleReuters	1.216	1.187	1.166	1.139	1.141	EN-MacroFin-Google	1.098	1.082	1.029	1.041	1.03
DFA(5)-MacroFin-Reuters	1.217	1.188	1.166	1.14	1.142	EN-MacroFin-GoogleReuters	1.088	1.104	1.033	1.048	1.05
PLS(1)-MacroFin	1.215	1.191	1.187	1.168	1.168	EN-MacroFin-Reuters	1.109	1.073	1.042	1.054	1.032
PLS(1)-MacroFin-Google	1.219	1.192	1.187	1.168	1.168	SSlab-MacroFin	1.115	1.114	1.105	1.116	1.114
PLS(1)-MacroFin-GoogleReuters	1.219	1.193	1.187	1.168	1.168	SSlab-MacroFin-Google	1.116	1.114	1.105	1.113	1.117
PLS(1)-MacroFin-Reuters	1.216	1.191	1.187	1.168	1.168	SSlab-MacroFin-GoogleReuters	1.116	1.117	1.1	1.113	1.114
PLS(2)-MacroFin	1.205	1.196	1.177	1.163	1.163	SSlab-MacroFin-Reuters	1.115	1.115	1.105	1.115	1.112
PLS(2)-MacroFin-Google	1.209	1.199	1.177	1.162	1.162	Best1	0.946	0.954	1.093	1.109	1.081
PLS(2)-MacroFin-GoogleReuters	1.21	1.2	1.177	1.162	1.162	Best3	0.997	1.003	1.052	1.044	1.068
PLS(2)-MacroFin-Reuters	1.205	1.197	1.177	1.163	1.163	Best5	1.021	1.029	1.068	1.038	1.057
PLS(3)-MacroFin	1.266	1.229	1.245	1.237	1.24	Best10	1.051	1.047	1.037	1.043	1.041
PLS(3)-MacroFin-Google	1.268	1.236	1.263	1.256	1.259						

Empirical assessment : GDP I

- Important preliminary remarks to be made due to the difference frequency at which the GDP is recorded in comparison to the other variables considered
  - Larger frequency mismatch with the weekly uncertainty indicators.
  - Very limited number of now-casts (12)
  - Strong uptrend observed during the nowcasting exercise
    - Privileging univariate methods
- Considering Germany, SPC4, PLS5 and PLS3 with MicroFin and Reuters Uncertainty Index are top performing models in term of MAE. Also, SPC5 with MicroFin and Google Uncertainty Index return good results for h=5 and h=4
  - AR (AIC) returns the smallest MAE
  - PLS3 with MicroFin and Google Uncertainty Index is the top model for h=1

Empirical assessment : GDP II

- Considering France, both in terms of MAE and RMSFE, the top model is Univariate BaggedETS across all h
  - PLS5 with MicroFin and Reuters as well as DFA5 with MicroFin and Google Uncertainty Index
- Similar results for Italy where univariate models such as ETS and BaggedETS are the top performance
  - A simple linear regression using 3 lags of GDP and Reuters uncertainty Index returns small MAE across all h
- Concerning the UK, Spline and ETS are top performance across all h.
  - Similarly to Italy , Simple Linear regressions with 3lags of GDP and Reuters uncertainty Index returns small MAE across all

Empirical assessment : GDP III

Model	-5w	-4w	-3w	-2w	-1w	Model	-5w	-4w	-3w	-2w	-1w
Average(4)	0.334	0.334	0.334	0.334	0.334	PLS(3)-MacroFin-GoogleReuters	0.317	0.313	0.316	0.319	0.311
Average(12)	0.301	0.301	0.301	0.301	0.301	PLS(3)-MacroFin-Reuters	0.317	0.314	0.312	0.314	0.306
Average(24)	0.301	0.301	0.301	0.301	0.301	PLS(4)-MacroFin	0.307	0.304	0.311	0.315	0.307
Naive	0.35	0.35	0.35	0.35	0.35	PLS(4)-MacroFin-Google	0.31	0.301	0.316	0.32	0.312
AR(1)	0.32	0.32	0.323	0.323	0.323	PLS(4)-MacroFin-GoogleReuters	0.31	0.301	0.317	0.321	0.313
AR(4)	0.338	0.338	0.354	0.343	0.343	PLS(4)-MacroFin-Reuters	0.306	0.304	0.313	0.316	0.308
AR(AIC)	0.315	0.315	0.303	0.304	0.304	PLS(5)-MacroFin	0.32	0.288	0.321	0.316	0.31
AutoArima	0.348	0.348	0.348	0.348	0.348	PLS(5)-MacroFin-Google	0.33	0.291	0.327	0.322	0.316
ETS	0.329	0.329	0.329	0.329	0.329	PLS(5)-MacroFin-GoogleReuters	0.337	0.29	0.328	0.322	0.316
BaggedETS	0.267	0.266	0.269	0.253	0.26	PLS(5)-MacroFin-Reuters	0.329	0.287	0.322	0.317	0.311
BATS	0.361	0.361	0.361	0.361	0.361	SPC(1)-MacroFin	0.323	0.312	0.318	0.323	0.316
TBATS	0.361	0.361	0.361	0.361	0.361	SPC(1)-MacroFin-Google	0.323	0.312	0.317	0.322	0.315
NN	0.52	0.541	0.55	0.5	0.486	SPC(1)-MacroFin-GoogleReuters	0.323	0.312	0.318	0.322	0.315
Spline	0.341	0.341	0.341	0.341	0.341	SPC(1)-MacroFin-Reuters	0.322	0.312	0.319	0.322	0.315
THETA	0.301	0.301	0.301	0.301	0.301	SPC(2)-MacroFin	0.313	0.311	0.304	0.309	0.304
Google	0.314	0.298	0.324	0.319	0.319	SPC(2)-MacroFin-Google	0.315	0.31	0.302	0.307	0.302
Google-L1	0.33	0.313	0.328	0.315	0.315	SPC(2)-MacroFin-GoogleReuters	0.314	0.31	0.302	0.306	0.302
Google-L3	0.378	0.367	0.358	0.336	0.336	SPC(2)-MacroFin-Reuters	0.313	0.311	0.305	0.309	0.304
Reuters	0.328	0.327	0.326	0.325	0.324	SPC(3)-MacroFin	0.345	0.329	0.309	0.312	0.307
Reuters-L1	0.34	0.328	0.331	0.329	0.328	SPC(3)-MacroFin-Google	0.346	0.328	0.308	0.311	0.305
Reuters-L3	0.332	0.325	0.367	0.41	0.409	SPC(3)-MacroFin-GoogleReuters	0.345	0.326	0.307	0.31	0.305
DFA(2)-MacroFin	0.314	0.308	0.307	0.309	0.305	SPC(3)-MacroFin-Reuters	0.345	0.328	0.308	0.312	0.306
DFA(2)-MacroFin-Google	0.314	0.308	0.306	0.307	0.303	SPC(4)-MacroFin	0.352	0.331	0.311	0.314	0.309
DFA(2)-MacroFin-GoogleReuters	0.314	0.308	0.306	0.307	0.303	SPC(4)-MacroFin-Google	0.354	0.337	0.31	0.312	0.306
DFA(2)-MacroFin-Reuters	0.314	0.308	0.307	0.309	0.304	SPC(4)-MacroFin-GoogleReuters	0.354	0.334	0.309	0.312	0.303
DFA(3)-MacroFin	0.338	0.318	0.306	0.309	0.305	SPC(4)-MacroFin-Reuters	0.352	0.333	0.312	0.314	0.308
DFA(3)-MacroFin-Google	0.338	0.318	0.305	0.308	0.304	SPC(5)-MacroFin	0.365	0.337	0.311	0.313	0.305
DFA(3)-MacroFin-GoogleReuters	0.337	0.318	0.305	0.308	0.304	SPC(5)-MacroFin-Google	0.372	0.336	0.31	0.31	0.302
DFA(3)-MacroFin-Reuters	0.337	0.318	0.306	0.309	0.305	SPC(5)-MacroFin-GoogleReuters	0.404	0.355	0.309	0.314	0.304
DFA(4)-MacroFin	0.355	0.332	0.303	0.307	0.302	SPC(5)-MacroFin-Reuters	0.398	0.343	0.313	0.315	0.307
DFA(4)-MacroFin-Google	0.359	0.336	0.302	0.305	0.3	LASSO-MacroFin	0.363	0.377	0.362	0.327	0.338
DFA(4)-MacroFin-GoogleReuters	0.357	0.335	0.302	0.305	0.3	LASSO-MacroFin-Google	0.376	0.372	0.38	0.325	0.345
DFA(4)-MacroFin-Reuters	0.352	0.331	0.303	0.307	0.302	LASSO-MacroFin-GoogleReuters	0.308	0.37	0.366	0.336	0.328
DFA(5)-MacroFin	0.366	0.341	0.301	0.303	0.299	LASSO-MacroFin-Reuters	0.328	0.382	0.359	0.33	0.332
DFA(5)-MacroFin-Google	0.367	0.34	0.301	0.302	0.297	EN-MacroFin	0.326	0.35	0.348	0.326	0.346
DFA(5)-MacroFin-GoogleReuters	0.385	0.352	0.303	0.304	0.3	EN-MacroFin-Google	0.332	0.386	0.352	0.348	0.328
DFA(5)-MacroFin-Reuters	0.387	0.348	0.305	0.307	0.302	EN-MacroFin-GoogleReuters	0.374	0.374	0.365	0.321	0.327
PLS(1)-MacroFin	0.329	0.32	0.332	0.337	0.332	EN-MacroFin-Reuters	0.347	0.37	0.345	0.319	0.321
PLS(1)-MacroFin-Google	0.328	0.318	0.33	0.335	0.33	SSlab-MacroFin	0.304	0.312	0.34	0.33	0.322
PLS(1)-MacroFin-GoogleReuters	0.328	0.318	0.331	0.335	0.33	SSlab-MacroFin-Google	0.309	0.311	0.346	0.334	0.322
PLS(1)-MacroFin-Reuters	0.329	0.32	0.332	0.337	0.332	SSlab-MacroFin-GoogleReuters	0.308	0.309	0.351	0.343	0.332
PLS(2)-MacroFin	0.325	0.32	0.315	0.317	0.309	SSlab-MacroFin-Reuters	0.309	0.311	0.344	0.334	0.324
PLS(2)-MacroFin-Google	0.327	0.32	0.315	0.317	0.309	Best1	0.271	0.285	0.289	0.294	0.3
PLS(2)-MacroFin-GoogleReuters	0.327	0.319	0.315	0.317	0.309	Best3	0.311	0.292	0.321	0.317	0.317
PLS(2)-MacroFin-Reuters	0.325	0.32	0.315	0.317	0.309	Best5	0.33	0.303	0.323	0.32	0.326
PLS(3)-MacroFin	0.317	0.313	0.312	0.314	0.306	Best10	0.326	0.316	0.333	0.332	0.33
PLS(3)-MacroFin-Google	0.317	0.312	0.316	0.319	0.311						

Empirical assessment: some general considerations

- The use of big data based uncertainty indicators generally results in an improvement of the nowcasting ability of several multivariate models
  - IPI still remain a critical case
  - Results of GDP probably influenced by the strong trend characterizing the evaluation period
- Absence of homogeneity of the results across variables and countries
  - The only observed regularity is the well expected improvement of the forecasting accuracy when approaching the official release date
- As it is the case in many simulation exercises, univariate methods performs pretty well even if they can hardly be proposed as a tool for the regular production of nowcasting
- Considering the multivariate models, several of them perform well in specific cases but none of them is outperforming in a clear way the others

Some consideration on future activities

- Interesting results obtained confirming the usefulness of big data for macroeconomic nowcasting
  - Further investigations needed
  - different target variables,
  - alternative set of predictors
  - larger timespan and evaluation period
- The performance of other big data types in the nowcasting exercise still to be investigated
  - financial big data, credit card data
  - IOT ecosystem data
- Checking for the possibility of using big data to obtain higher frequency reliable estimates for key macroeconomic indicators
  - GDP inflation
- Constructing higher frequency indicators of climate and sentiment

Guidance for using big data I

- Using big data still remains a challenging exercise with a lot of risks
  - obtaining misleading results
  - drawing wrong conclusions
- How, when and what for using big data still needs to be clarified.
- Statistician using big data needs to be adequately guided on what to do and how to deal with this kind of data.
- Guidance and recommendations are particularly needed at this stage

Guidance for using big data II

- Proposing a step-by-step approach to the use of big data
  - generalizing the one already proposed in 2017
  - Handbook on rapid estimates Chapter 16
  - Jointly developed by Eurostat and GOPA experts
- Including 13 steps presented in a harmonized structure
- For each step we propose
  - A detailed presentation of the associated problem as a free text
  - on or more suggestions presented as sequence of bullet points

Step 1: Identification of the Statistical Problem

Description

Before starting the search phase for identifying possibly useful big data, it is important to clarify in the most accurate way the problem we want to address.

- For nowcasting exercise, the focus is on timeliness gains and the updating frequency.
- For estimating high frequency indicators, frequency of estimation and feasible timeliness targets should be defined.
- For new indicators, focus on defining what they are supposed to measure, their nature as qualitative or quantitative, as well as frequency of computation and, the timelines of production.

For all the three cases presented above, it is essential also to specify if we aim to use big data together with other source of information or alone.

Recommendations

- Prepare a clear and detailed problem statement for each of the statistical problems to be addressed and solved
- Clearly specify if big data are intended to be used as a complement of standard information or as the only source of information.

Step 2: Implement a Modern and Efficient IT System to Host and Use Big Data

Description

- Factors indicating necessity of specific hard and soft frameworks for big data include;
  - Large, typically unstructured data sets
  - Often available at very high frequency
  - Need for adaptation of traditional statistical and econometric modelling or exploratory techniques.
- Designing an efficient system able to collect, store and treat big data is necessary to use big data in any kind of statistical production.

Recommendations

- Implementing a modern, efficient and flexible IT system for collecting, storing and treating big data of various nature and characteristics,
- Implementing a set of statistical and econometric tools specifically adapted to deal with big data,
- Use as much as possible existing IT solutions and well known and tested software applications, such as those available in Matlab and R packages.

Step 3: Big Data Usefulness

Description

Several advantages of Big Data have been widely emphasized

- In a nowcasting context, Big Data provides potentially relevant complementary information and granular indicators.
- In the temporal dimension, providing nowcasts at a given frequency
- In the cross-sectional dimension, providing relevant information on units not fully covered by traditional coincident and leading indicators.

However, Big Data use must be evaluated in terms of potential benefits without significant deterioration of their reliability.

Recommendations

- Suggest the use of Big Data only when there are well founded expectations of their useful- ness either for fixing problems in existing nowcasts, or to improve the timeliness/accuracy of nowcasting, or to produce reliable and timely high frequency indicators or even new ones.
- Do not consider Big Data sources with doubtful or even spurious correlations with the target variable.

Step 4: Big Data Search

Description

A careful search and classification of existing data is a crucial step

- **Social Networks** (human-sourced information) contains typically loosely structured and often ungoverned data-
- **Traditional Business Systems** (process-mediated data) often collected by either private or public institutions are highly structured and include transactions, reference tables and relationships, as well as the metadata that sets its context.
- **Internet of Things** (IoT, machine- generated data) gives well-structured data, making it suitable for computer processing, but their size and speed is way beyond traditional approaches.

A priori, it is very difficult to give general guidelines on a preferred data source because the choice is heavily dependent on the target problem we are facing with.

Recommendations

- Search in the wider possible set of Big Data having clearly in mind the specificities and the characteristics of the target exercise as well as what we want to estimate (a quantitative growth indicator for something not yet measured, a qualitative sentiment or climate one, the growth of an existing indicator, etc.)
- Check for the adherence of available Big Data to what the target variable, if any, is really measuring or what we aim to measure.

Step 5: Assessment of Big Data Accessibility and Quality

GOPA Consulting Group

Description

- Evaluate potential Big Data source based on whether and at what cost the information is actually available.
- Should have continuity and reliability of provision, especially for the construction of new indicators.
- Should have suitable time span
- Control of the stability of the relationship with the target variable is important, especially for nowcasting.

Recommendations

- Prioritizing data providers which are able to give sufficient guarantee of the continuity of the data process and of the availability of a good and regularly updated metadata associated to the Big Data.
- Prioritizing Big Data sources which ensure sufficient time and cross-sectional coverage to properly building up exercises such as nowcasting, construction of high frequency indicators and of new ones.
- In order to deal with possible instabilities of the Big Data either in relation with the target variables or over the time, models should be re-specified on a regular basis and occasionally in presence of unexpected events.

Step 6: Big data Conversion

Description

- Often substantial costs in making Big Data suitable for statistical or econometric exercises, given unstructured nature.
- There is no unique way to transform unstructured into structured data, as the specific transformation depends on the type of Big Data.
- Transformations can be treated in a unified analytical context, considered as functions that map the Big Data into the real space.

Recommendations

- Ensure the availability of an exhaustive and well documented Big Data conversion process,
- Identify the most suitable structuring and conversion rule to be applied to the Big Data under investigation,
- For originally non-numerical big data identify and apply a sound, transparent and well documented quantification method that preserves the information content of the original Big Data.

Step 7: Big Data Cleaning

Description

- Big data can include much more anomalies than traditional statistics because it is less strictly monitored so it and often has irregular data observations.
- By consequence, especially numerical big data, can be, even after the structuring and conversion phase, still affected by a high number of anomalies such as:
  - missing data,
  - presence of outliers,
  - presence of transitory or permanent shifts,
  - inconsistency of pattern over the time.

Recommendations

- Conduct an in-depth and possibly multivariate analysis of the anomalies and identify those to be corrected,
- Define a statistically and economically sound approach for dealing with all kinds of anomalies to be corrected.

Step 8: Removing Seasonal and other High Frequency Periodicities

Description

- Since most big data are collected and recorded at high or even very-high frequency, it is likely that they are characterized by seasonal or other very short run periodicities, as well as calendar effects.
- When dealing with nowcasting exercises, big data should be subject to a similar filtering process as the target variable, and the same also applies when constructing high frequency indicators.
- When constructing a new indicator, the decision involves:
  - Defining the frequency of computing and publishing the new indicator.
  - Testing for presence of calendar and seasonal effects.
  - If present, deciding if the new indicator should be presented raw, seasonally or seasonally and calendar adjusted, or both.

Recommendations

- Filter out seasonal and other high frequency periodicities as well as calendar effects, if any, in a consistent way with the reference variable in the nowcasting exercise or in the construction of high frequency indicators.
- For new indicators, if seasonal and other high frequency periodicities, together with calendar effects, have to be removed, then follow standard approaches properly extended to deal with high frequency data.
- Do not develop ad hoc solutions but invest in generalizing already existing seasonal adjustment methods.

Step 9: Bias Correction

Description

- Bias can be due to the so-called “digital divide” or the tendency of individuals and businesses not to report truthfully their experiences, assessments and opinions.
- Internet based big data subject to the ”digital divide” where not all age classes are equally represented in the internet population.
- Furthermore, internet of things based big data differs based on coverage of automatic sensors and detectors in different regions or countries.
- If not detected, such biases can potentially influence not only nowcasting exercises but also the construction of high frequency indicators and new indicators.

Recommendations

- If a bias in Big Data results is observed, provided that it has been reasonably stable in the last few years, a bias correction can be included in the modelling strategy,
- If a bias in the Big Data results is very unstable or too large, then Big Data should be considered not reliable enough to be used.

Step 10: Designing a Big Data Modelling Strategy

Description

Big Data prevents the use of standard econometric methods. Approaches to Big data econometrics can be classified as follows:

- **Machine Learning Methods** which start by regularizing OLS estimation to make it feasible also when N is very large.
- **Heuristic Optimization** which uses information criteria to reach a good balance between model fit and parsimony by assigning a penalty dependent on the number of model parameters
- **Dimensionality Reduction Techniques** based on reducing the dimension of the dataset by producing a much smaller set of generated regressors, which can then be used in a second step in standard econometric models to produce nowcasts and forecasts in common ways.
- **Shrinkage Estimators and Bayesian Methods;** Shrinkage estimators typically regularize OLS estimation, making it feasible also when N is very large and larger than T , by adding a set of a priori constraints on the model parameters.
- **Nowcast Pooling,** using a (possibly very large) set of small econometric models to produce nowcasts, one model for each of the N available indicator or small subset of them, and then to combine the resulting many nowcasts or forecasts into a single prediction.

Recommendations

- In the absence of any a priori information on the relative performance of various techniques, as many methods as possible should be evaluated and compared in order to select the best performing one.
- Alternative modelling strategies should be compared also by looking at the balance between their complexity in computational terms and their empirical performance.
- In case of mixed frequency data, linear methods such as UMIDAS and, as a second best, Bridge, should be prioritized.
- Forecast combination and model averaging techniques, also when the mixed frequency aspect is present, can be used as an alternative to a large-scale comparison among competing techniques.
- When the target is to estimate high frequency indicators, regression based temporal dis- aggregation methods should be privileged, possibly in combination with variable reduction techniques such as principal components or partial least square regressions.
- When constructing a new indicator for which a target or benchmarking variable is unavailable, methods such as principal components or partial least square are mostly recommended.

Step 11: Modelling Unstructured Big Data

Description

- When the conversion and structuring of big data into time-series of panel data is problematic or even unfeasible, it is still possible to work on the original unstructured data and try to extract some relevant signals and commonalities, especially useful for the construction of new indicators, mainly qualitative ones.
- By contrast, working on unstructured big data does not seem promising when dealing with nowcasting exercises or with the calculation of high frequency indicators.
- Working on unstructured data can be very time consuming and it also requires powerful IT equipment. Mainly exploratory techniques, such as data mining (and its extensions such as text mining) or data analytics can be used to find patterns, commonalities and to extract signals from the original Big Data.
- Based on the results obtained, it is then possible to construct a new indicator, mainly qualitative, which synthesizes the outcomes of the techniques mentioned above.

Recommendations

- Using either data mining or data analytics methods, in a judgmental approach, to avoid the identification of unrealistic patterns and relationships should be a focus.

Step 12: Evaluation of Big Data Based Analyses

Description

Assessing the contribution of Big Data in a critical and comprehensive way can be done by answering a set of questions:

- Is there a "Big Data hubris"? I.e. is there an implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis
- Is there a risk of "False positives"? Big Data based indicators may be producing good results just due to data snooping.
- Are correlations mistaken for causes when interpreting the results?
- Is there instability in the nowcasting ability of specific Big Data based indicators? Or in the relationships between certain big data and a target variable if any

Recommendations

- Conducting an in-depth real-time or pseudo real-time simulation of competing models in order to evaluate their relative performance.
- Models including Big Data should be preferred when they significantly lead to an improvement of the reliability and accuracy of the nowcasting at the same point in time.
- Models including Big Data should also be preferred when they allow for timelier nowcasting without any significant loss in terms of reliability and accuracy.

Step 13: Implementation of Big Data Based Nowcasting and Indicators

Description

Big Data can improve the situation from the view of different aspects:

- For nowcasting, either in terms of more precise results with a given timeliness or timelier results with a non-significant loss of precision.
- For the construction of high frequency indicators, a smooth pattern delivering non-contradictory signals with respect to the corresponding low frequency benchmarking variable.
- For the development of a new indicator, the possibility of obtaining reliable information on a phenomenon not previously measured or monitored.

Recommendations

- Implementing and publishing the most reliable nowcasts available either at the end of the reference period or at the beginning of the following one.
- Moving towards a daily or weekly update on nowcasting already during the reference period, only after detailed pros and cons analysis and a consultation of the most relevant stakeholders.
- Implementing and publishing high frequency indicators at the frequency for which the reliability and the precision with respect to the low frequency reference variable are the best possible
- Implementing and publishing new indicators choosing their release on the basis of statistical considerations as well as on user need and the frequency of similar indicators
- The new Big Data based nowcasting and indicators should be accompanied by clear metadata and widely available reference and methodological documentation.

General remarks on the Step by Step approach

- We have proposed an operational step by step approach for using Big Data as a complement to traditional information sources.
- We have considered three main cases: the production of early estimates of macroeconomic indicators (nowcasting), of high frequency indicators, and of new indicators not previously available.
- The proposed step-by-step approach and associated recommendations aim to facilitate the activity of experts and official statisticians involved in the evaluation of the big data usefulness for official statistics.
- Further extension of the step by step approach
  - Taking into account all big data typologies
  - Incorporating SDGs related issues

Possible integration of big data into official statistics

- Several tangible progresses towards the integration of big data into official statistics in the recent years
  - Promising empirical results
  - adapting existing modelling methods to big data specificities
  - developing new modelling strategies
  - creating IT infrastructure to host and treat big data

- big data integration still a complex issue
  - Medium term objective
- Several aspects still to be clarified and problems to be fixed

Possible integration of big data into official statistics 2

- how far would we go with the integration?
  - limiting to specific areas
    - nowcasting
  - providing support to fill gaps in official statistics
    - SDGs
  - more structural integration
    - smart statistics

- how to adapt official statistics paradigms and standards
  - GSBPM, GSIM, GANSO
  - adapting quality framework and quality dimensions

Possible integration of big data into official statistics 3

- National statistical authorities the best placed to manage the change
  - deciding on the degree of the integration
    - taking into account national specificities
  - defining a realistic timetable for the integration
    - considering available resources
  - Steering the whole process in a coordinated manner with other statistical institutions
    - supervisory role of international institutions
- But National statistical authorities facing several constraints:
  - Lack of internal resources
  - in some cases lack of specific skills in the big data area
  - need of prioritizing activities
    - privileging core activities: production and dissemination of official statistics

Private public partnership (PPP)

GOPA Consulting Group

- PPP has already proven its effectiveness in the statistical cooperation activities
  - several successful cases all over the world
- Possible role of private consulting companies active in the statistical domain
  - Providing advice on why, when, how and what for using big data
  - supporting specific methodological development
  - Providing methodological and technical support during all big data integration process
- If PPP has already provided excellent results in the cooperation domain, why it should not work in this specific case?

Making PPP more effective

- Creating and maintaining common private/public platform
  - a real-time exchange of information, results and for solving problems
- Promoting the exchange of ideas and experiences among private and public statisticians
  - organizing regular brainstorming sessions and workshops on specific statistical fields
- fostering the cooperation and the exchange of information on the development and updating or official statistics standards
- Promoting the participation of private companies into methodological activities organized by public statistical authorities
- Creating Center of Excellence to monitor and evaluate PPP activities

What can private companies bring to the PPP

- Modern and effective project management
- Effective support to official statistical authorities both for development and daily activities
- Carrying on advanced methodological and innovation project on behalf of statistical authorities
- Ability of organizing, developing and managing large high level networks of experts

What can private companies bring to the PPP 2

- Fast and easy deployment of highly qualified experts on new projects in a world wide scale
- high capacity of transforming the outcomes of methodological and innovation projects in concretely usable output
  - new products supporting daily production and dissemination activities
  - new tools and methods ready to be used
- Transforming projects into processes

A roadmap for tasks sharing within a PPP

- Public: identification of policy needs and priorities. Ex-ante evaluation
  - How big data could contribute in fill gaps in official statistics?
- Public: projects conception and description
  - Using big Data to improve timeliness
  - Smart statistics
- Private: projects implementation and realization
  - Designing new methods and tools dealing with big data
- Public: project supervision
- Private: delivering projects outcomes and results
  - quality insurance

A roadmap for tasks sharing within a PPP 2

- Public: ex-post evaluation of the project
  - identification of possible ways forward and suggesting future actions
- Private: post project follow up based on suggested actions
  - proposal for a testing phase
- Public-Private: jointly designing the testing framework
- Private: carrying out the texting exercise
  - delivering results
    - Comparing alternative big data based nowcasting schemes
    - Testing new smart statistics based indicators

A roadmap for tasks sharing within a PPP 3

- Public: evaluation of testing results and decision on the transformation of the project into a process
- Private: providing specific capacity building and training activities to the official statisticians
  - sensitization to the new data statistical challenges
  - familiarization to new methods and tools
  - Ensuring the availability of internal staff with appropriate skills to manage the new process
- Private: preparing a draft process consistent with official statistics standards
- Public: Validating the new process
  - launching the implementation phase

A roadmap for tasks sharing within a PPP 4

- Public-Private: Implementing testing and finalizing the process
- Public: final decision on putting the process into regular production
- Private: Providing, upon request, help support and advice on the new process

Thank you for your attention