International Conference on Big Data for Official Statistics
Organised by UNSD and NBS China

Beijing, China, 28-30 October 2014

Web scraping for Labour Statistics

Emanuele Baldacci

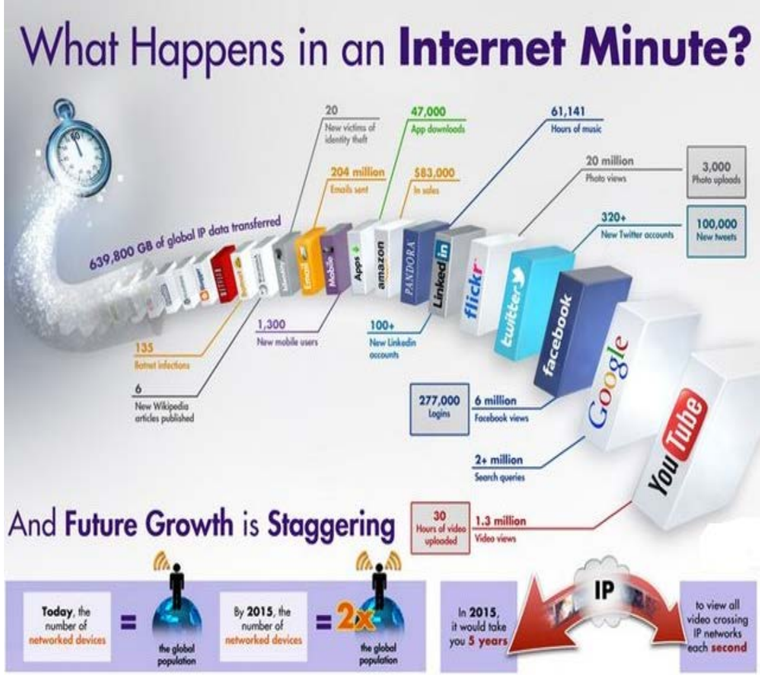
Italian National Institute of Statistics (Istat)

Head, Department for Integration, Quality, Research and Production Networks Development (DIQR)

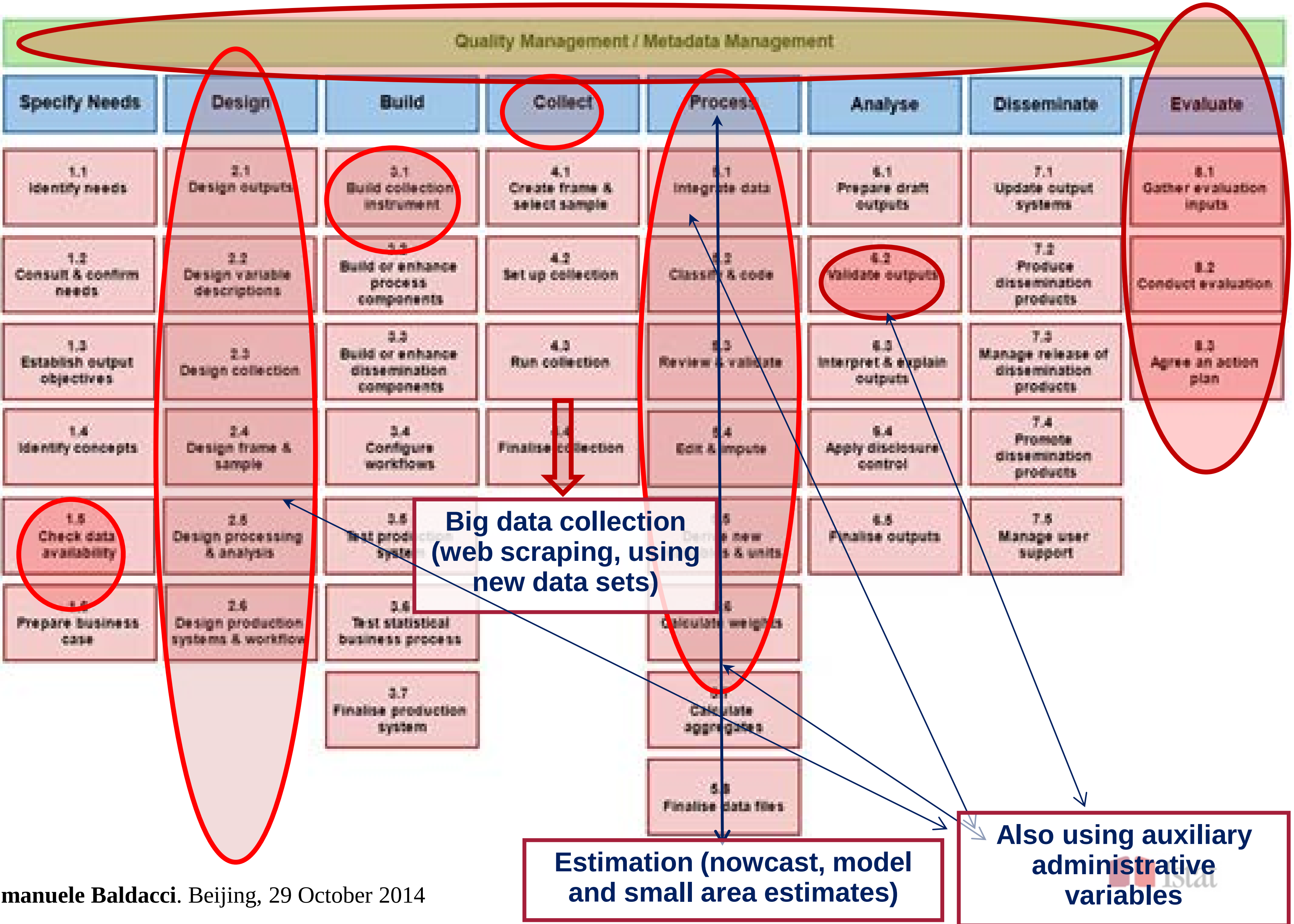
中华人民共和国国家统计局
National Bureau of Statistics of the People's Republic of China

Outline

- Big Data: what can change?
- Google trend capability
- Istat integrated research project
- Focus on Labour Market
- Main results
- Area of interest for the next future



Big Data: what can change?



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

Big data collection (web scraping, using new data sets)

Estimation (nowcast, model and small area estimates)

Also using auxiliary administrative variables

Istat ongoing experimentation

Different type of sources

		Google Trends
DATA SOURCE		*Human-sourced information*
ISSUES	IT	✓*Search records acquisition and processing*
	STATISTICAL	✓*Enhance prevision performances (e.g., root-mean-square error)*
	ORGANIZATIONAL	✓*Access to Web search results*
SCENARIO (IMPACT ON THE PRODUCTION PROCESS)		***Limited impact*** ***on the production process: complementing estimation phase***

Open questions

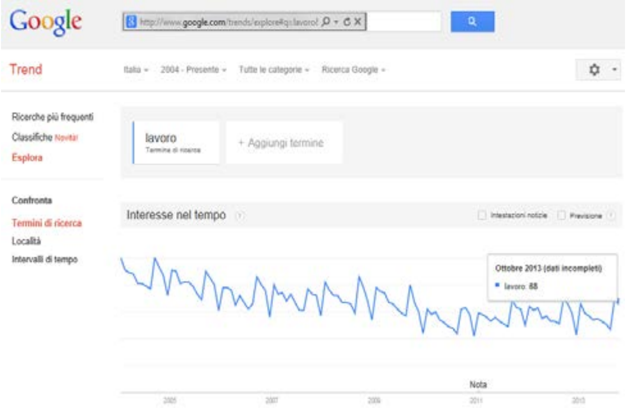
Different possible impacts on production scenarios

Google Trend capability

- It is possible **to exploit it for different statistical purposes**

- At national level it allows to exploit the **time series of query shares to improve the quality of estimates** of short-term (monthly or quarterly) socio-economic indicators

- It can be used **as external auxiliary information for improving the forecasting or nowcasting** of short term indicators (Labour Market Indicators)



lavoro
Termine di ricerca

Ottobre 2013 (dati incompleti)
lavoro: 88

Desires, opinions, sentiments

Istat integrated research project

- Aimed at evaluating the potential of Big Data for the production of preliminary estimates and small area estimates

✓ Modifying Istat methodology to introduce **Google Trend auxiliary variables** in the time-space model for provisional estimation

✓ Studying the **variables available on Google Trend** for the construction of **advanced estimators** of certain categories of products (related to retail, wholesale and PRODCOM survey), or of **small area estimates related to the Labour Market** (employed, unemployed, etc.), evaluating the predictive ability of these variables to produce estimates on a monthly, quarterly, provincial and regional level

✓ Analysing the time series of monthly ILO variables exploiting the **Google Trends weekly queries**

Google Cloud Platform

Google Developers

A focus on Labour Market

- **Purpose**:
  - ✓ Use Google Trends for forecasting and nowcasting purposes in the Labour Force domain:
    - Monthly forecasting, e.g. Release on February of (*i*) unemployment rate related to January (*ii*) prediction of the unemployment rate related to February
    - Nowcasting for small areas - improving territorial level estimates by accessing GT series at finer granularity (e.g. Provinces)
- **Actors involved**:
  - ✓ Istat, Central Methodology Sector and Labour Force Survey
- **Status of advancement**: Ongoing experimentations

Labour Market Estimation (I)

- **Methodology**:

  Benchmarking

  ✓Autoregressive model *vs* Usage of Google Trends results on the category «job» and on the search term «job offers» by adopting several prediction models (parametric and semi-parametric)

  ✓Comparison extended to macroeconomics prediction models

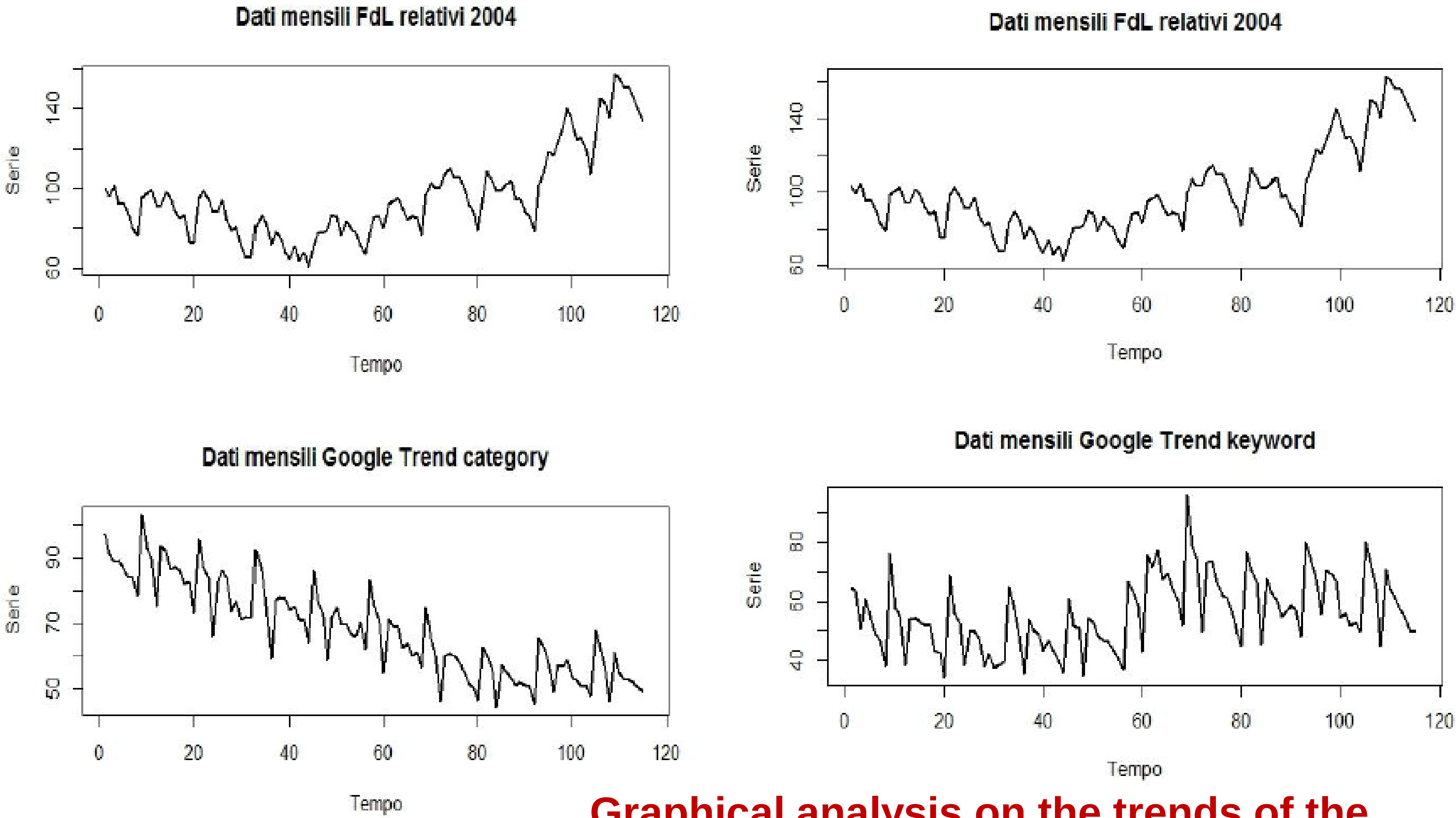
Labour Market Estimation (II)

- **Preliminary results**

✓ Cross correlation in preliminary tests indicates a potential use of Google Trends for the target

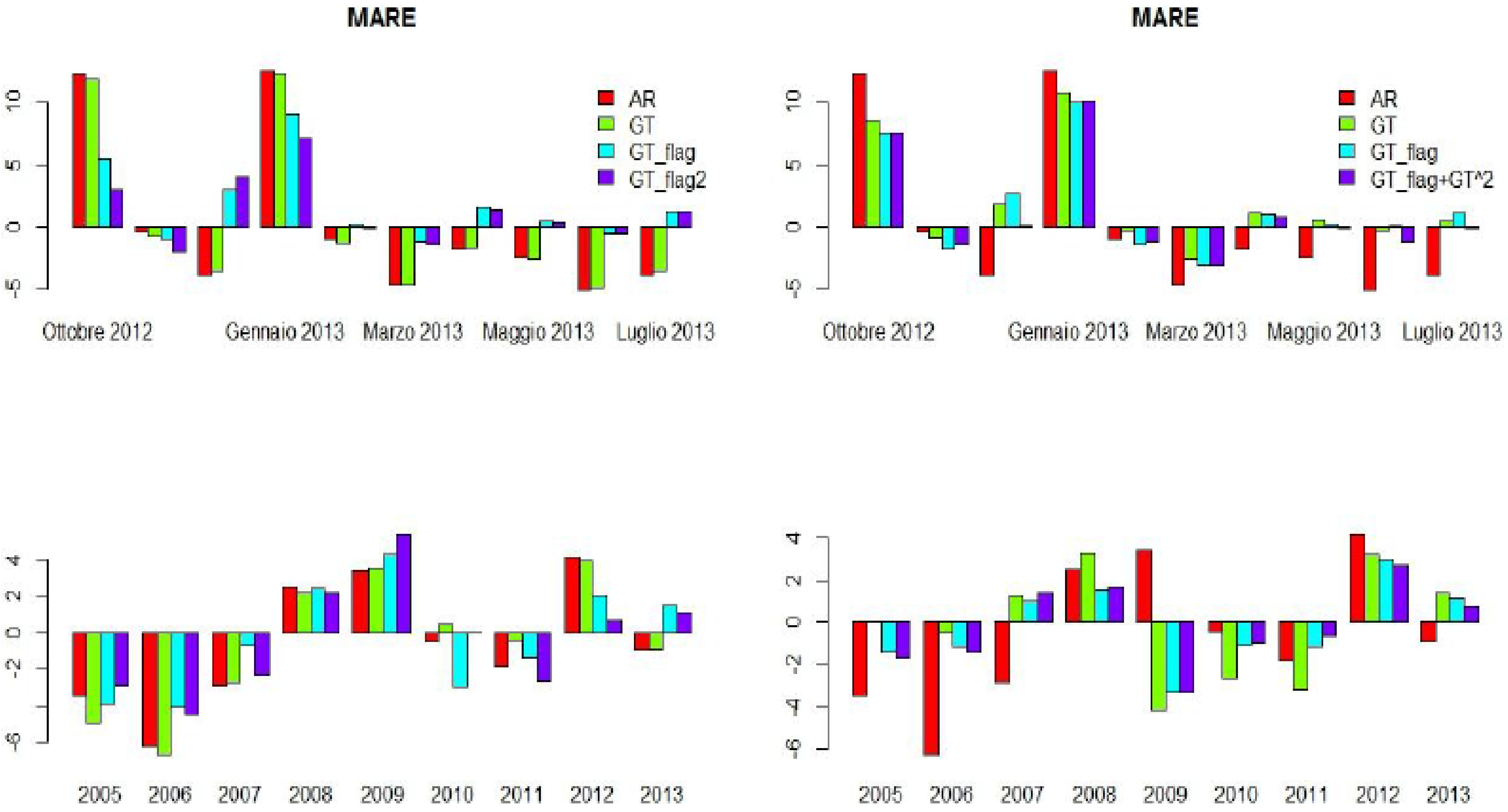
- **Outcome**:
  - ✓ Google Trends use on **Italian data** in the Labour Force domain
  - ✓ Monthly prediction capabilities
  - ✓ Finer territorial level series estimation

Main Results (I)



Graphical analysis on the trends of the time series: Google Trends and Istat Labour Force Survey

Main Results (II)



Analysis of alternative models and comparison with the benchmark model

Areas of interest for web scraping in the next future

- **Social media statistics**: messages on public social media are available to anyone with Internet access. The content of these messages should be investigated in order to understand their potentiality in terms of contribution to statistical indicators regarding spare time activities, media, politics, etc.. Text mining is the candidate tool for such an analysis

- **Wellbeing indicators**: to be calculated investigating the potential use in terms of attitude towards the economic situation. Messages on social networks like *Facebook* are difficult to obtain, while the ones left on *Twitter* are publicly available

- **Measuring and monitoring Smart Cities**: at the moment a set of indicators is under evaluation. This is a **multidimensional and complex area** requiring the availability of timely and low cost information that can be obtained through the integration of data coming from **official statistical sources**, the exploitation of **Administrative Archives**, the use of **Big Data**

Thank you for your attention

感谢您的关注

Contacts:

baldacci@istat.it

www.istat.it

Main References

- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. Detecting influenza epidemics using search engine query data. Nature 457: 1012–10155, 2008.
- Choi, H. and Varian, H., Predicting the Present with Google Trends. Economic Record, 88: 2–9. doi: 10.1111/j.1475-4932.2012.00809.x, 2012.
- H. Choi and Hal Varian (2009) , "Predicting the present", Google Inc., Draft Date April 10
- Askitas, N. and Zimmermann, K. F. Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly, 55: 107–120, 2009.
- Francesco D'Amuri, Juri Marcucci. The Predictive Power of Google Searches in Forecasting Unemployment. Banca D'Italia, Working Papers n. 891, 2012.
- Pratesi M. , Pedreschi D., Giannotti F., Marchetti S., Salvati N. , Maggino F. (2013), "Small area model-based estimators using big data sources", NTTS 2013; Brussels
- D'Alò M., Gismondi R., Solari F., Naccarato A. (2006), Estimation in Repeated Business Surveys using Preliminary Sample Data, Atti della XLIII Riunione Scientifica SIS, 14-16 giugno, Torino.
- D'Alò M., Di Consiglio L., Falorsi S. Solari F., (2008), Small Area Estimation Methods far Socio-Economie Indicators in Households Surveys, Rivista Internazionale di Scienze Sociali, Università Cattolica del Sacro Cuore di Milano, Anno CXVI, Ottobre-Dicembre, 2008.
- D'Alò M., Di Consiglio L., Falorsi S. Solari F., (2012), Use of spatial information in small area models for unemployment rate estimation at subprovincial areas in Italy, Journal of the Indian Society of Agricultural Statistic,Volume 66, December 2012, pp. 1-239.
- Bacchini F., D'Alò M., Falorsi S., Fasulo A., Pappalardo C.. Does Google index improve the forecast of Italian labour market? In the proceedings of 47*th* Scientific Meeting of the Italian Statistical Society, 2014.