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

Beijing, China, 28-30 October 2014

Mobile phone data for Mobility statistics

Emanuele Baldacci

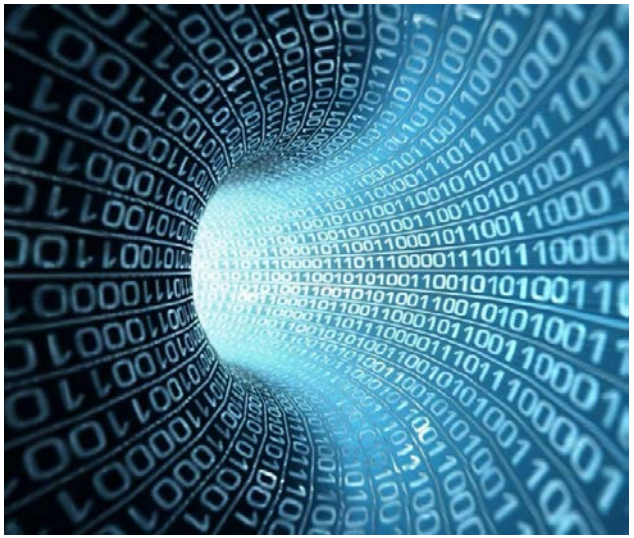
Italian National Institute of Statistics (Istat)

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 reference classification
- The methodology taxonomy
- Istat ongoing experimentation
- Persons and places
- Some experimentation details
- Main results
- Concluding remarks

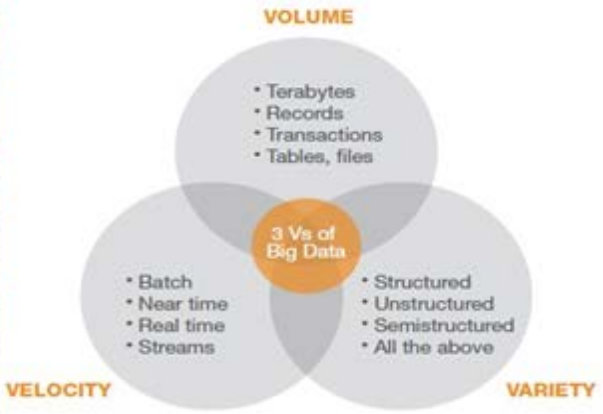


Big Data reference classification

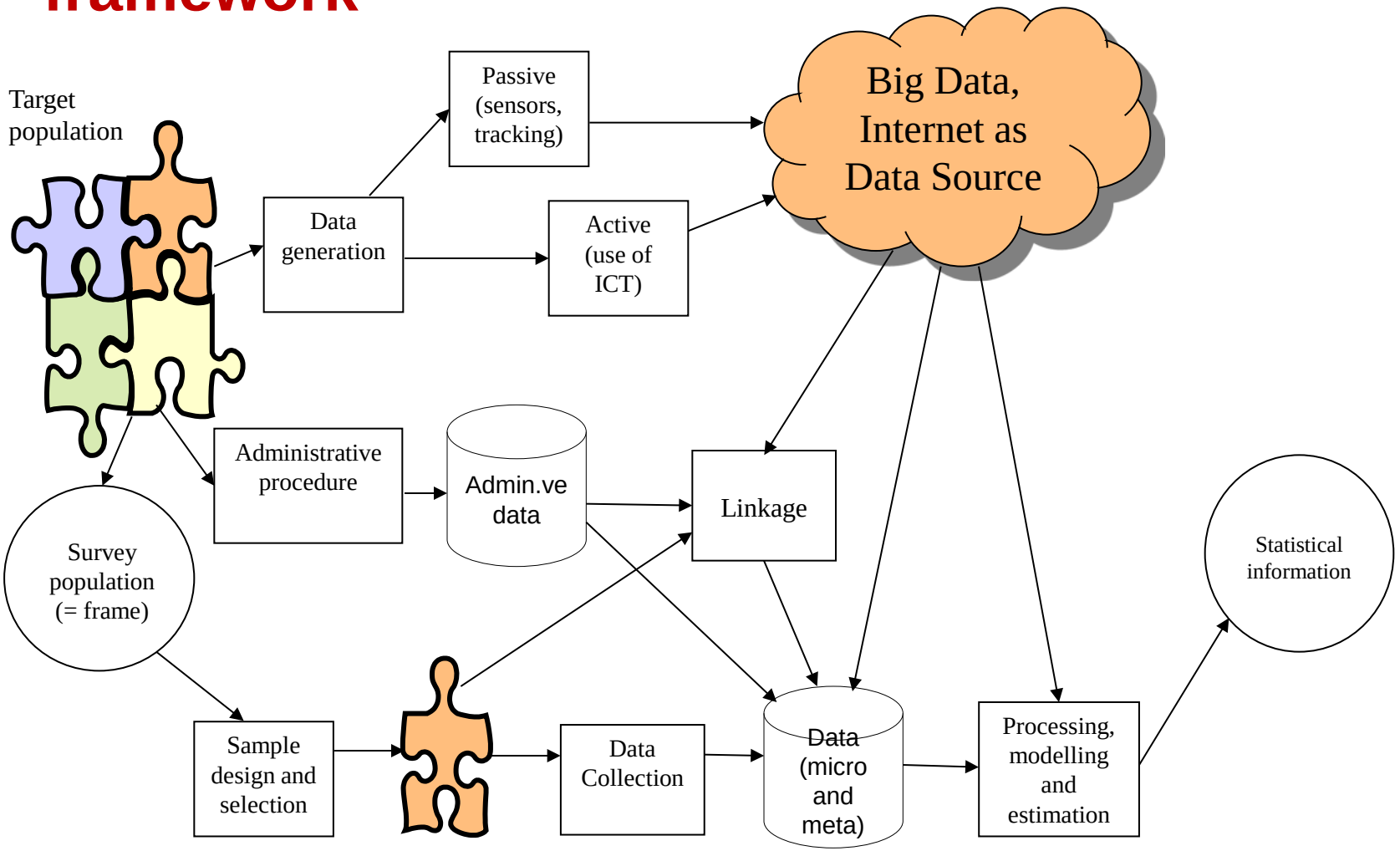
"*Data that is difficult to collect, store or process within the conventional systems of statistical organisations. Either their **volume**, **velocity**, structure or **variety** requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made*"

Big Data Project UNECE 2014

- Human-sourced information (Social Networks)
- Process-mediated data (Traditional Business systems and Websites)
- **Machine-generated data** (Automated Systems)



The methodological taxonomy: general framework



Istat ongoing experimentation

Different type of sources

Open questions

		Persons &Places
DATA SOURCE		***Machine-generated data***
ISSUES	IT	✓***Smart sensing application Pattern identification on tracking data***
	STATISTICAL	✓***Record linkage and Statistical matching Non homogeneous target populations Quality control on results***
	ORGANISATIONAL	✓***Privacy***
SCENARIO (IMPACT ON THE PRODUCTION PROCESS)		***Considerable impact on the production process : source replaces traditional sampling and collection***

Different possible impacts on production scenarios

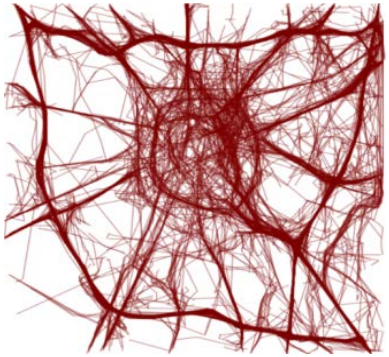
Persons and Places (I)

- **Purpose**:
  - ✓ Production of the origin/destination matrix of daily mobility for work and study at the spatial granularity of municipalities starting from mobile phone (tracking) data
- **Actors involved in the project**:
  - ✓ Istat (Central Methodology Sector, Directorate of Censuses, Administrative and Statistical Registers)
  - ✓ National Research Council (CNR)
  - ✓ University of Pisa
- **Status of advancement**: Ongoing implementation

Persons and Places (II)

- **Methodology**:
  - ✓ Inference of population mobility profile from GSM Call Data Records (CDR)

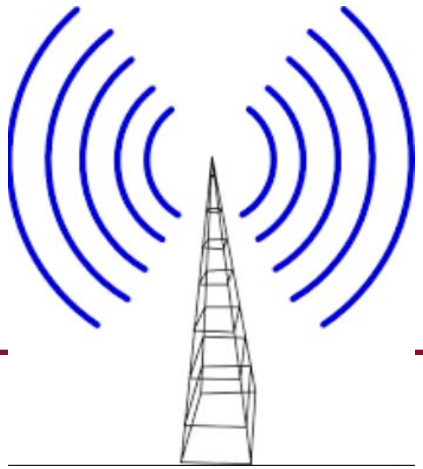
    Combination of pre-defined extraction patterns and unsupervised learning method (SOM - Self Organising Map)
  - ✓ Comparison with data derived from administrative sources



- **Outcome**:
  - ✓ Production of statistics on **city users** - Standing resident, Embedded city users, Daily city users (commuters)
  - ✓ Possible comparison of quality of statistics from a Big data source and from administrative sources

Some experimentation details

- The spatial granularity considered is the municipality level
- Focus on the **39 municipalities in the province of Pisa** (Tuscany, Italy)
- These municipalities host a **largely variable number of residents**, ranging from less than one thousand for the smaller ones, up to around 86,000 for the central municipality of Pisa, with an average of 10,000
- Each municipality is spatially covered by **an average of 3-4 GSM antennas**



The analysis process

- *Sociometer*, a **data mining tool for classifying users by means of their calls habits**, was extended to work on a larger territory and to include the flows of people between different territorial units (municipalities)
- The aim is **producing statistics that are comparable with those obtained by Istat**: residences and flows of people are studied using administrative data sources
- Achieving success along this direction means to be able to **safely integrate existing population and flow statistics with the continuously up-to-date estimates obtained from GSM data**: a further step towards exploiting Big Data in official statistics

Core objectives

- Correctly estimate, for each municipality, the population that belongs to each of the following categories, **already calculated by Istat using administrative data**:

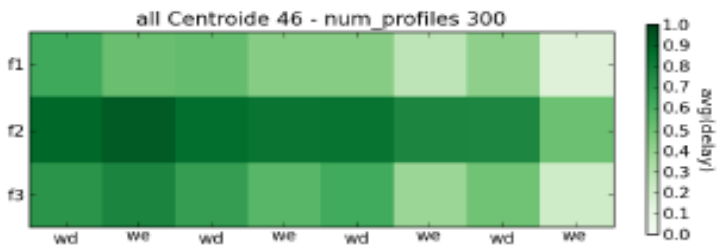
✓ **Standing residents in A**: persons who have formal residence and place of work (study) in the same municipality A, or who do not work (study)

✓ **Embedded city users in A**: people that spend long periods for working (studying) in a municipality A (e.g. most days of the week), while being formally resident in another municipality, different from A

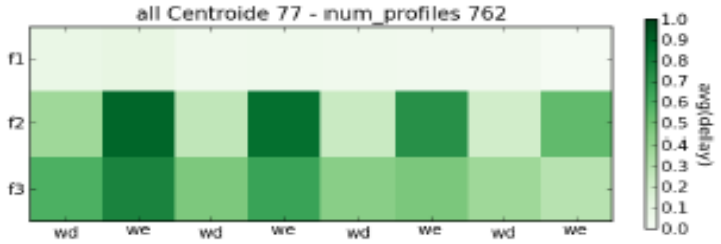
✓ **Daily city users in A**: people who commute to municipality A, having formal residence in another municipality, different from A

Main Results (I)

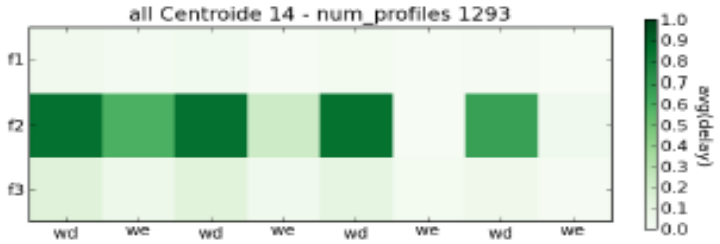
- The analysis process on GSM data allows to infer slightly different user categories: Standing residents and Embedded city users are not distinguished yet, due to **the lack of administrative information about the GSM users** (their physical presence tends to be identical)
- The physical presence of users allows to easily distinguish (at least in principle) **Dynamic vs. Static residents**



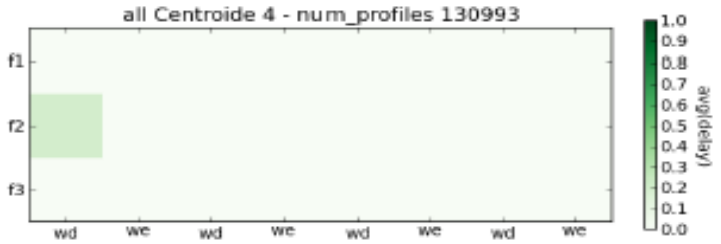
(a) Residents



(b) Dynamic residents



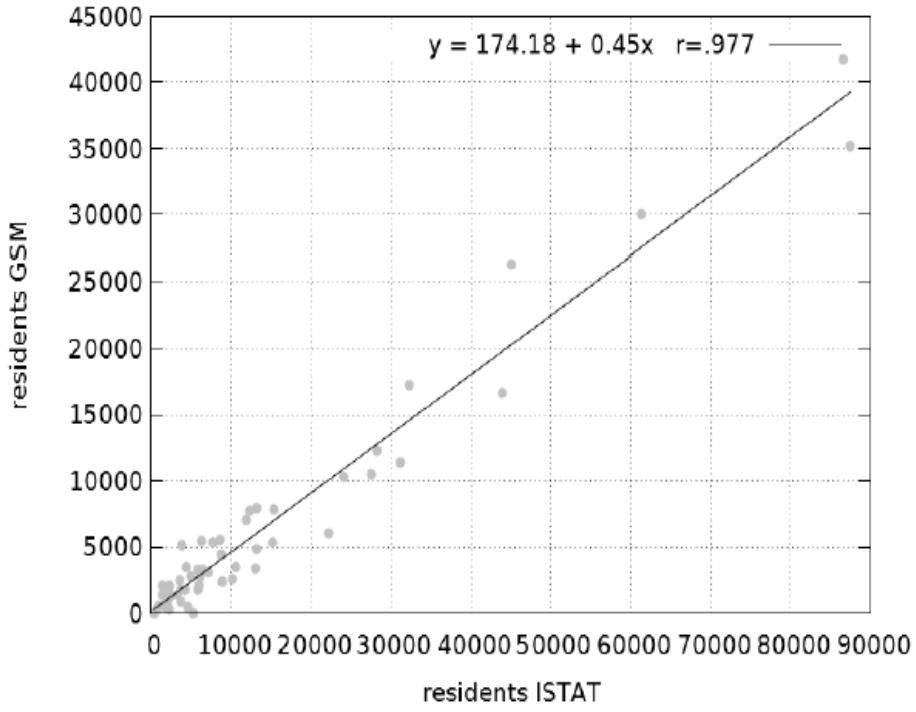
(c) Commuters

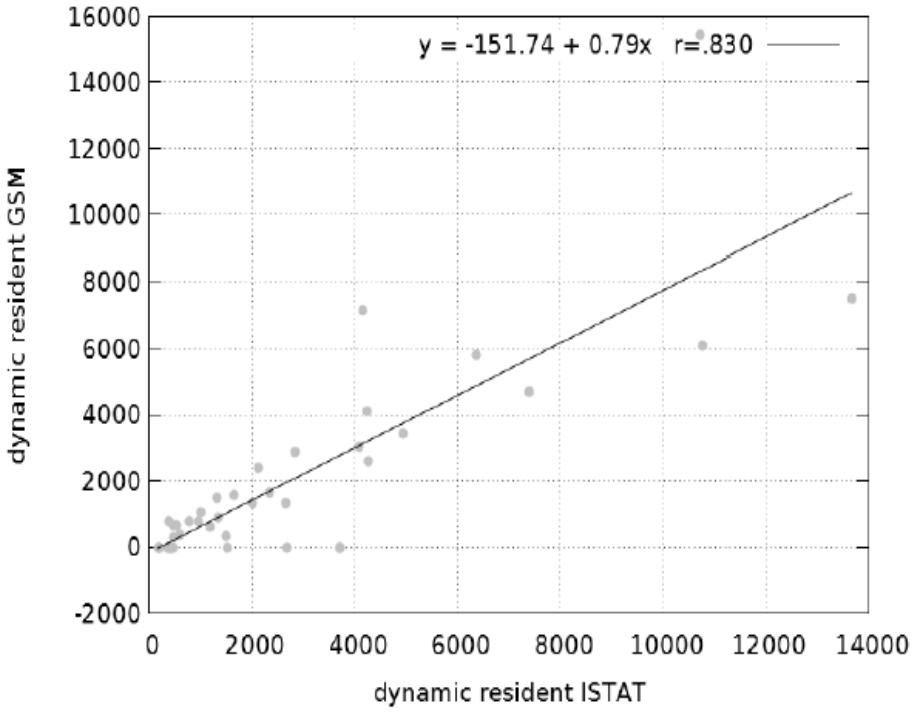


(d) Visitors

Istat

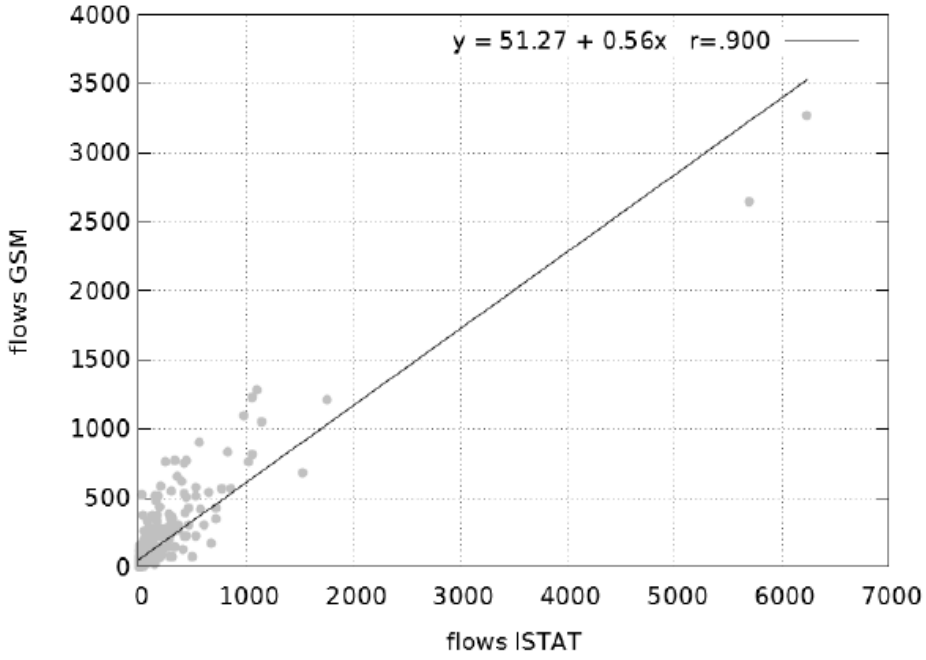
Main Results (II)



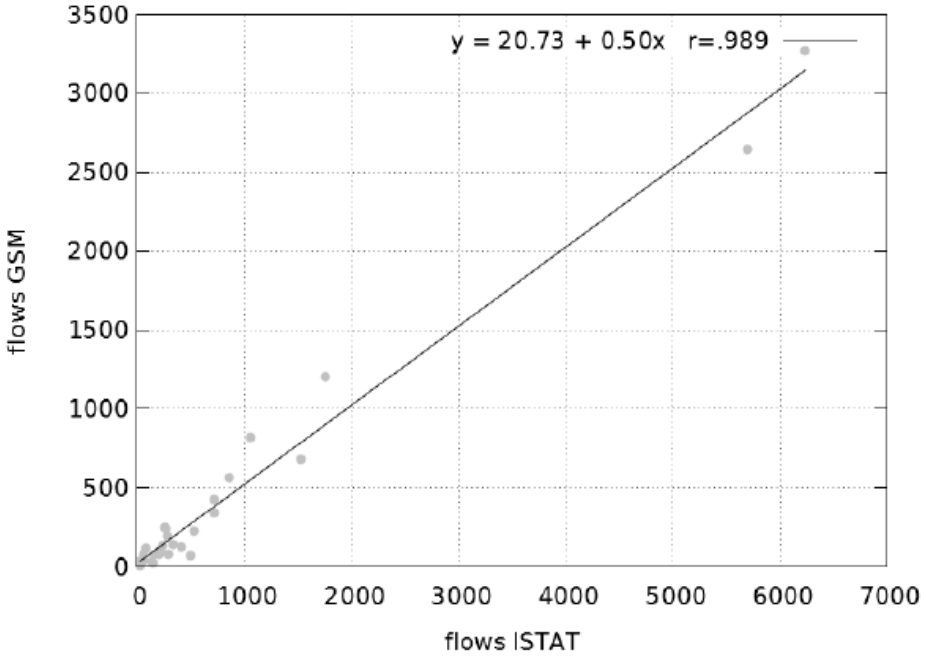


Correlation between GSM and ISTAT Resident and Dynamic resident

Main Results (III)



(a) All flows



(b) Pisa

Correlation between systematic flows measured by Istat and *Sociometer*

Concluding remarks

- Population and flow estimation **based on mobile phone**
- Big Data used as proxy of the **presence and mobility of individuals**
- The **results obtained are generally encouraging** and, for some specific statistics, **very accurate** in comparison to analogous statistics obtained with official data
- Several **improvements are planned for the future**, also extending the experimentation to larger areas, in order to both increase the sample of population covered and avoid border effects

Thank you for your attention

感谢您的关注

Contacts:
baldacci@istat.it

www.istat.it

Main References

- Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD 11. ACM, New York, NY. 2011.
- Nanni, M., Trasarti, R., Furletti, B., Gabrielli, L., Mede, P. V. D., Bruijn, J. D., Romph, E. D., and Bruil, G. MP4-A project: Mobility planning for Africa. In D4D Challenge @ 3rd Conf. on the Analysis of Mobile Phone datasets (NetMob 2013). 2013.
- Oltenau, A.-M., Trasarti, R., Couronne, T., Giannotti, F., Nanni, M., Smoreda, Z., and Ziemlicki, C. GSM data analysis for tourism application In Proceedings of 7th International Symposium on Spatial Data Quality (ISSDQ). 2011.
- F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal, 2011.
- B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo Turism fluxes observatory: deriving mobility indicators from GSM calls habits In the Book of Abstracts of NetMob 2013.
- B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo. Analysis of GSM calls data for understanding user mobility behaviour In the Proceedings of Big Data 2013.
- B. Furletti, L. Gabrielli, G. Garofalo, F. Giannotti, L. Milli, M. Nanni, D. Pedreschi, Roberta Vivio. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach. In the proceedings of 47*th* Scientific Meeting of the Italian Statistical Society, 2014.