

The Virtuous Cycle of Big Data and Big Cities: a Case Study from Milan

Il ciclo virtuoso tra Big Data e Big Cities: un caso studio milanese

Piercesare Secchi, Simone Vantini, Paolo Zanini

Abstract We here present the analysis of mobile network data from the city of Milan. In particular we aim at identifying spatiotemporal patterns characterizing specific locations and/or specific periods possibly associated to different human activities taking place within the city. The analysis is carried out by means of a new dimensional reduction technique, named Hierarchical Independent Component Analysis, providing a low-dimension and sparse representation of the phenomenon.

Abstract *In questo lavoro viene presentata un'analisi di dati di telefonia mobile relativi alla città di Milano. L'obiettivo primario del lavoro è l'individuazione di profili spazio-temporali in grado di caratterizzare luoghi e/o periodi specifici associati a diverse attività che prendono luogo all'interno della città. L'analisi viene effettuata mediante una nuova tecnica di riduzione dimensionale (analisi gerarchica delle componenti indipendenti gerarchica) in grado di fornire una rappresentazione sia di dimensione ridotta che sparsa del fenomeno oggetto di studio.*

Key words: Big Data, Mobile Data, Independent Component Analysis, Wavelets.

Piercesare Secchi

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy e-mail: simone.vantini@polimi.it

Simone Vantini

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy e-mail: piercesare.secchi@polimi.it

Paolo Zanini

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy e-mail: paolo.zanini@polimi.it

1 Big Data and Big Cities

New technologies are providing the scientific community with large amount of data pertaining to different aspects of every day life (e.g., consumer preferences, mobility habits, reader tastes, crime distribution, weather evolution). These are usually called Big Data since - from a statistical perspective - they are “big” in both their dimension (i.e., the number of features that are recorded for each sample unit) and their sample size (i.e., the number of subjects appearing in the data base). One could think that the massiveness of these data is just a matter of improving storing and quick access technologies and thus an issue for computer scientists and engineers. The growing literature pertaining to big data in both the machine learning and the statistical community proves instead that big data provide challenges that go far beyond data managing skills. A big question that always comes with big data is indeed: “How big are these data? Are they really big or just look like they are big?”. Big data of course contain a large amount of information, but with respect to some specific questions, much of it can be useless and acting just as a Baroque curtain covering a simple picture of the phenomenon. These questions urge the current research in Statistics and Machine Learning investigating for new dimensional reduction techniques able to deal with large amount of data in a quick way and providing easily interpretable results.

Large urban communities are natural concentrations of people, services, and activities, and are thus natural big-data factories. Some of these data could be used to improve the quality of life and the efficiency of the city itself creating a virtuous cycle between the data generated by the city and its development. In detail, in terms of mobility, everyday private and public firms - thanks to modern technologies - record vehicle gps-tracks, and mobile phone data, bike- and car-sharing hires, bus, underground, train, and airplane ticketing data, accesses to congestion-charge or low-emission areas. These data if suitably analyzed can give to policy makers (and CEOs) useful insights for planning the future development of the city: projects of new roads, location of docking station of the bike- and car-sharing networks, redesign of public transport routes, optimal time scheduling of train and airplane departures, distribution of police patrols and ambulances in the city.

In the following, we present a paradigmatic example of the virtuous cycle between big data and big cities: an analysis of mobile-phone data collected in the city of Milan. In details, we will introduce a new dimensional reduction technique (i.e., Hierarchical Independent Component Analysis) providing a low-dimension and sparse representation of the spatiotemporal evolution of the density of people in the city of Milan within a typical non-holiday week. This research is part of the Green Move Project, a research project held at Politecnico di Milano and financed by Regione Lombardia investigating the potential of a third-generation car-sharing system within the city of Milan ([3]).

2 A Case Study from Milan

The Telecom Italia database ([2]) contains 13.8 million records referenced in space over a regular lattice of 10573 pixels at a spatial resolution of nearly 250 m covering an area of 757 km², and referenced in time over a regular grid of 1308 intervals at a temporal resolution of 15 minutes covering a period of two weeks. Each record is an intensity measure of the use of the mobile network in a specific site at a specific time. In particular we aim at identifying spatiotemporal patterns characterizing specific locations and/or specific periods possibly associated to different human activities taking place within the city.

Hierarchical Independent Component Analysis (HICA [4]) is based on a recursive hierarchical application of Independent Component Analysis (ICA [1]) on pairs of variables. The final output of HICA is a multi-resolution, wavelet-inspired, and data-driven basis useful to represent data and to investigate their sources of variability and able to describe both macro and micro-scale features of these data. Differently from ICA and similarly to wavelets, the basis provided by HICA is naturally ordered according to the dimension of each basis element support. Similarly to ICA, the basis provided by HICA, is not orthogonal and driven by the search for independent components. Thus differently from all principal-component inspired methods no purely mathematical (and possibly unrealistic) constraint is imposed to the final representation.

Coherently with the geostatistical literature instants of times are assumed to index variables while sites to index instances. The declination of HICA in this case allows to impose temporal sparsity to the final representation. The analysis unveils interesting patterns interpretable in terms of working, residential, shopping, leisure, and commuting activities.

References

1. Hyvarinen, A., Oja, E. (2000): "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, 13, 411430.
2. Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., Vitelli, V. (2012): "Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region", *Mox Report 25/2012*, Dipartimento di Matematica, Politecnico di Milano.
3. Politecnico di Milano (2013): "Greenmove: Sintesi del progetto", http://gm.polimi.it/Download/Sintesi_Green_Move.pdf.
4. Secchi, P., Vantini, S., Zanini, P. (2014): "Hierarchical Independent Component Analysis: a multi-resolution non-orthogonal data-driven basis", *Mox Report 01/2014*, Dipartimento di Matematica, Politecnico di Milano.