Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area

Chiara Bocci, Leonardo Piccini, Patrizia Lattarulo (IRPET)
Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni (ISTI CNR)

Florence, 28-30 June 2017

Outline of the presentation

I. Introduction

II. Data sources

III. Methodology

IV. Results and future developments

Some introductory considerations

- ICT proliferation has led to the diffusion of sensors able to track human activity, as well as to the storage and computational capabilities needed to record and analyse them

- Academic literature and institutions recognize the intrinsic value of this kind of data but lack a systematic approach in its use

- We need to define shared and rigorous procedures to estimate and validate the data so that we can use them to program and evaluate public policies

Big Data sources

- Under the “Big Data” umbrella we can find heterogeneous data sources and analytic tools, all sharing some common features:

- PROS: the ability to timely capture and measure phenomena that escape traditional data sources and methods, with higher spatial and temporal disaggregation.

- CONS: Since they are collected without any kind of filter or correction, they may be subject to an unquantified bias and are usually not publicly (freely) available.

- Socio-economic analysis with Big Data has thus far focused mainly on:
  1. social network analysis
  2. machine learning
  3. unstructured data ← IRPET projects are mainly concentrated in this area

IRPET – ISTI/CNR joint research

- ISTI (Istituto di Scienza e Tecnologie dell'Informazione) is a CNR institute that jointly manages the Knowledge Discovery and Data Mining Laboratory (KDD Lab) with the University of Pisa Informatics Department, working on this topics since 1999.

- In November 2015 IRPET and ISTI subscribed a research agreement to work on big data sources from a public policy perspective.

- PILOT PROJECTS: Initially we wanted to investigate the use of GPS and GSM data for mobility analysis. Later we expanded to other research topics.

Long term project plan

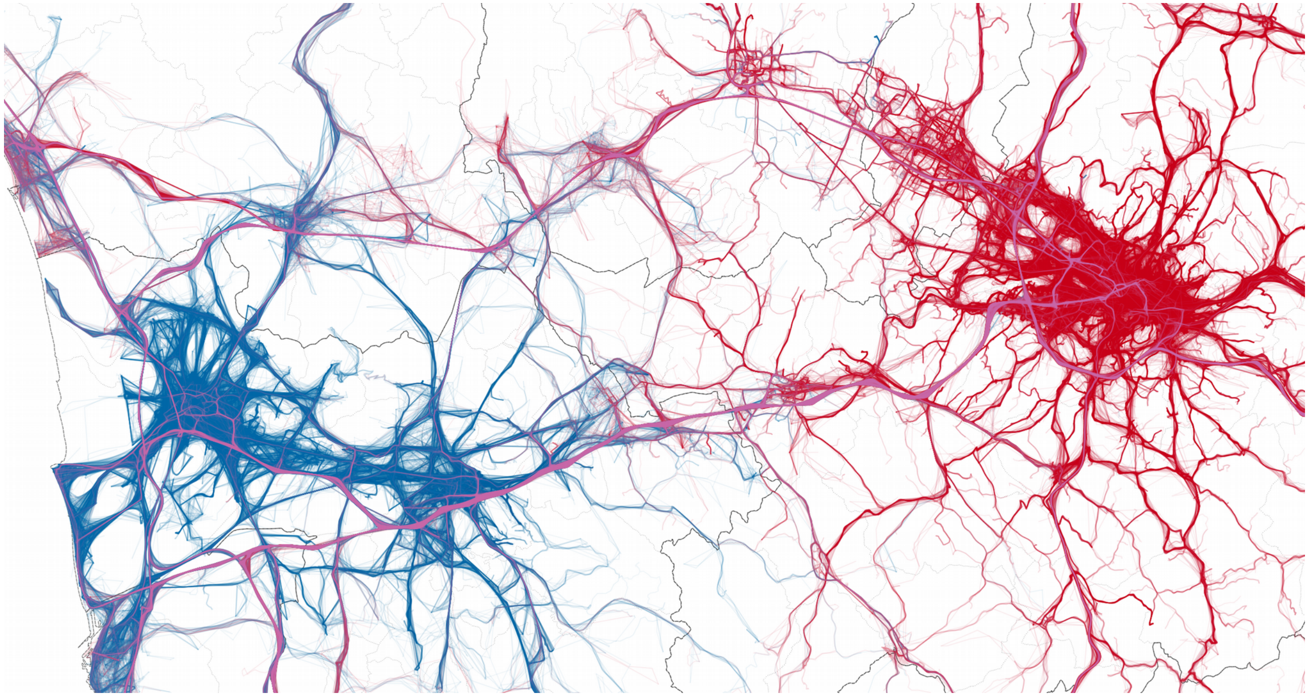
OBJECTIVE:

Joint use of traditional and innovative socio-economical data sources and analytical tools to implement and calibrate evaluation models for public policies at the regional and urban level.

STEPS:

- Data estimation and validation
- Pilot projects for tool testing
- Integration and systematization of all available data
- Regional and urban level analysis

GPS data source (I)



- The **GPS device** captures the position approximately every 30 seconds with a 10 meter precision. **Sample size** is nearly 240k vehicles (about 10% of actual Tuscan vehicles population), totalling about 12 million trips.

GPS data source (II)

Available data allows us to capture not only systematic (i.e. home-to-work and home-to-school) flows, but also non-systematic activities.

Since we cannot cross-reference mobility data with personal characteristics (age, sex, residency, etc.), we need to infer home and work location from the data itself.

Home is basically the most frequent visited location and work is the second most frequent one. This allows us to separate systematic flows from the rest.

Next step is to “expand” the GPS car flows in order to estimate the population flows, correcting for possible biases, both for cars and other transportations systems.

The estimation methodology

$$\widehat{Flow}_{i,j}^{CAR} = Flow_{i,j}^{GPS} * car.pen_i * avg.occ$$

$$\boxed{\widehat{Flow}_{i,j} = \widehat{Flow}_{i,j}^{CAR} + \widehat{Flow}_{i,j}^{MOTO} + \widehat{Flow}_{i,j}^{LPT}}$$

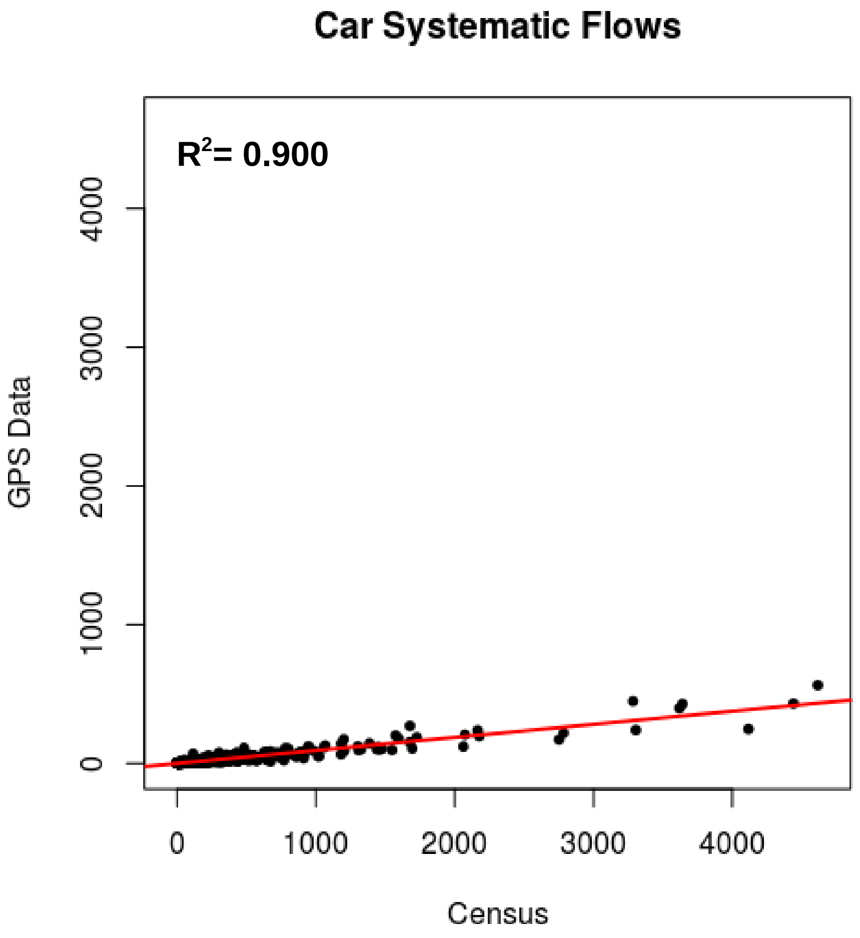
$$\widehat{Flow}_{i,j}^{MOTO} = Flow_{i,j}^{GPS} * car.pen_i * moto.ratio.i$$

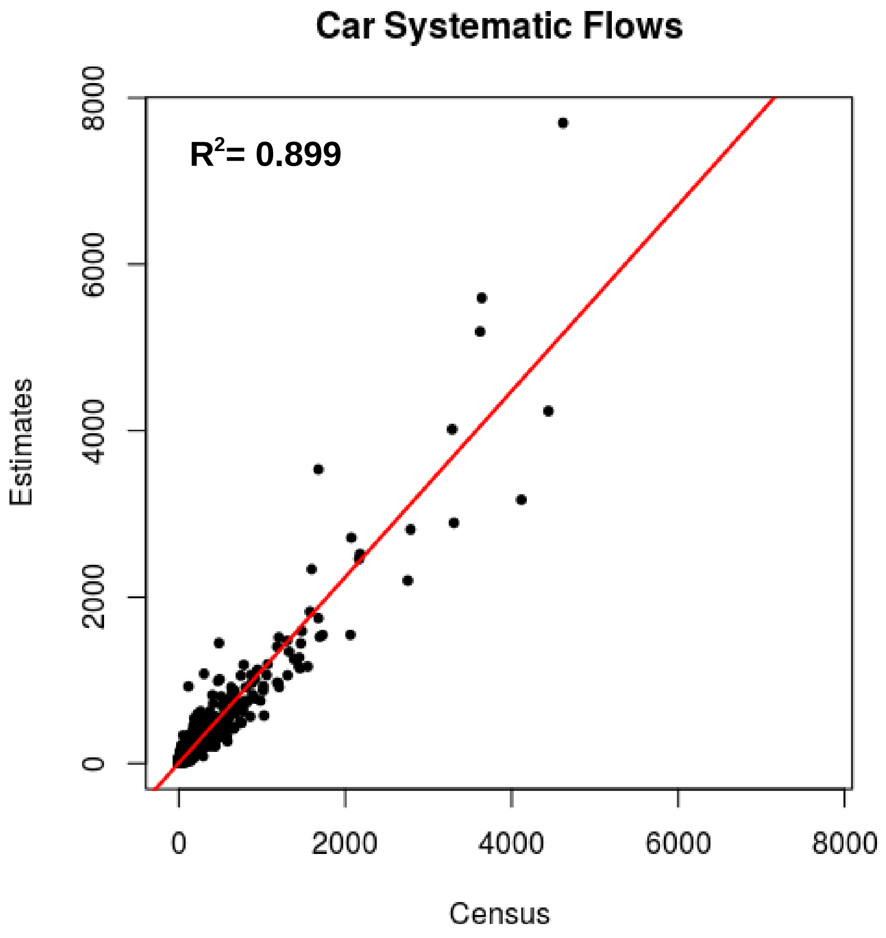
$$\widehat{Flow}_{i,j}^{LPT} = \frac{(\widehat{Flow}_{i,j}^{CAR} + \widehat{Flow}_{i,j}^{MOTO}) * \frac{pt.ind_{i,j}}{(1-pt.ind_{i,j})}}{\sum_{i,j}\left((\widehat{Flow}_{i,j}^{CAR} + \widehat{Flow}_{i,j}^{MOTO}) * \frac{pt.ind_{i,j}}{(1-pt.ind_{i,j})}\right)} * \frac{\hat{Q}_{pt}}{1-\hat{Q}_{pt}} \sum_{i,j}(\widehat{Flow}_{i,j}^{CAR} + \widehat{Flow}_{i,j}^{MOTO})$$

With:

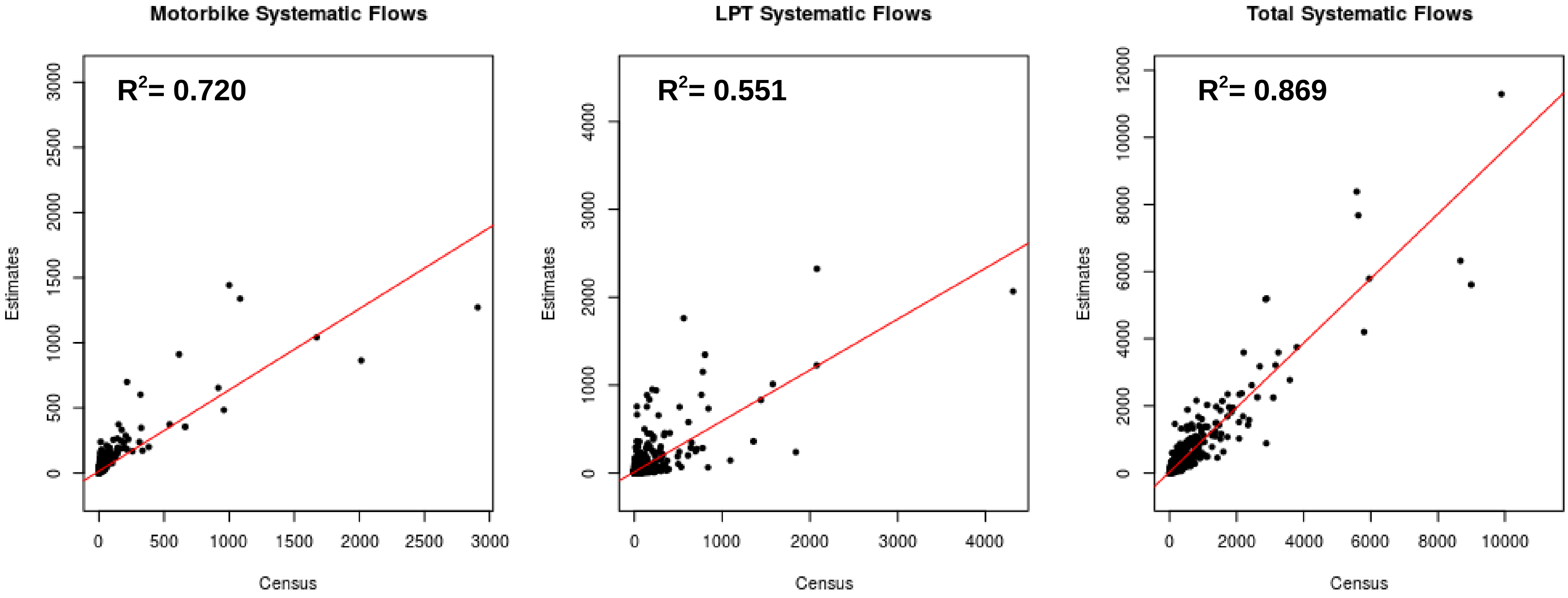
- **market penetration index:** $car.pen_i = \frac{\text{observed individuals per municipality } i}{\text{number of circulating cars in municipality } i}$
- **average car occupancy (from Census data):** $avg.occ$
- **motorbike/car ownership ratio:** $moto.ratio.i = \frac{\text{number of circulating motorbikes in municipality } i}{\text{number of circulating cars in municipality } i}$
- **LPT/car time travel ratio:** $time.ratio_{i,j} = \frac{\text{LPT time travel from } i \text{ to } j}{\text{car time travel from } i \text{ to } j}$
- **public transportation index:** $pt.ind_{i,j} = \frac{\max(time.ratio_{i,j}) - time.ratio_{i,j}}{\max(time.ratio_{i,j}) - \min(time.ratio_{i,j})}$
- **estimated regional quota of public transportation (from ISTAT MP):** $\hat{Q}_{pt}$

Goodness of fit (I)


Car Systematic Flows


Car Systematic Flows

Goodness of fit (II)



Goodness of fit (III)

We test the performance of our estimations applying different goodness of fit measures and using the 2011 census O/D matrix as benchmark.

Systematic Flows	TOTAL Estimates	Census	corr.	Relative RMSE	Theil's Indexes Um	Us	Uc	U	GEH$_k$ k=5	k=2	k=1
GPS Data	18,702	196,115	0.949	2.676	0.114	0.875	0.011	0.825	0.825	0.520	0.300
Car	227,714	196,115	0.948	1.104	0.021	0.203	0.776	0.172	0.987	0.813	0.432
Motorbike	35,353	23,051	0.848	4.174	0.016	0.237	0.748	0.308	0.995	0.887	0.649
LPT	51,231	57,261	0.742	3.139	0.000	0.106	0.893	0.368	0.959	0.854	0.700
Total	314,281	280,022	0.932	1.243	0.010	0.007	0.984	0.177	0.956	0.720	0.342

Bias proportion: =0 Ok, =1 No

$$U_m(x,y) = \frac{N \cdot (\bar{x} - \bar{y})^2}{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Variance proportion: =0 Ok, =1 No

$$U_s(x,y) = \frac{N \cdot \left(\sigma_x - \sigma_y\right)^2}{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Covariance proportion: =0 No, =1 Ok

$$U_c(x,y) = \frac{2 \cdot (1 - r) \cdot N \cdot \sigma_x \sigma_y}{\sum_{i=1}^{N}(x_i - y_i)^2}$$

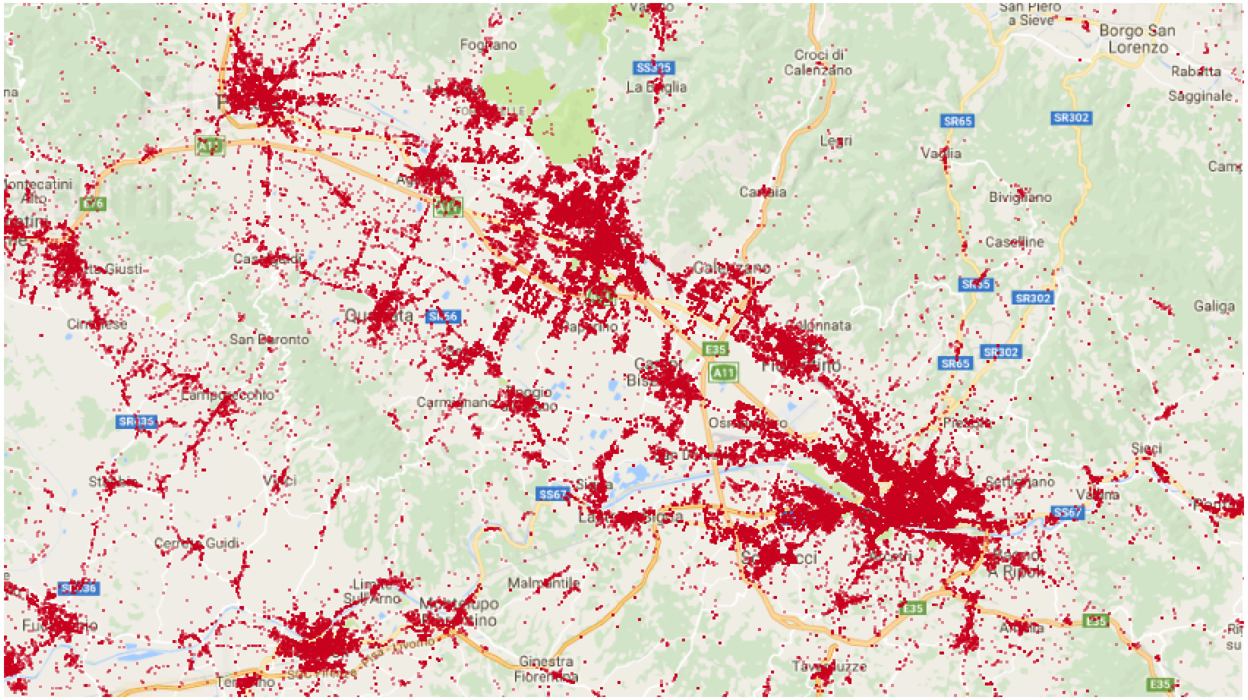
Inequality coeff: =0 Ok, =1 No

$$U(x,y) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}{x_i}^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}{y_i}^2}}$$

$$GEH_i(x,y) = \sqrt{\frac{2 \cdot (x_i - y_i)^2}{x_i + y_i}} \qquad GEH_k(x,y) = \frac{1}{N}\sum_{i=1}^{N}\delta_i \quad \text{with} \quad \delta_i = \begin{bmatrix} 1 \; if \; GEH_i(x,y) \leq k \\ 0 \qquad\qquad elsewhere \end{bmatrix}$$

GEH: 85% <1 Excellent
85% <2 Good
85% <5 Ok

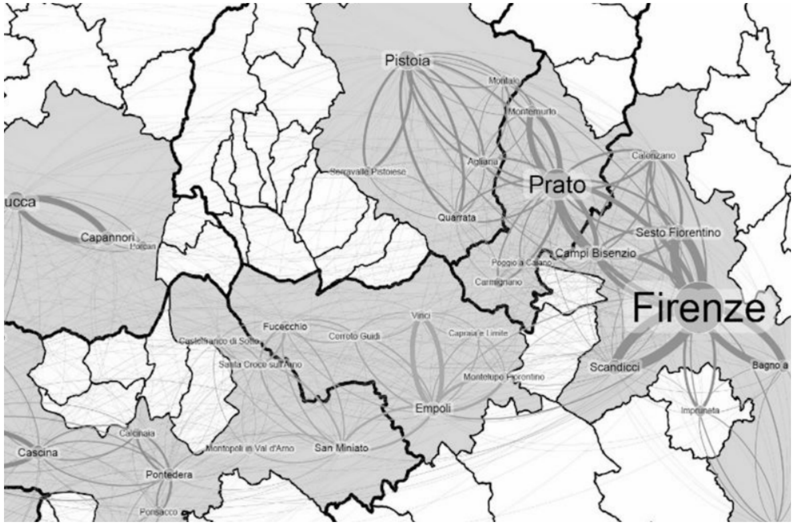
Preliminary applications

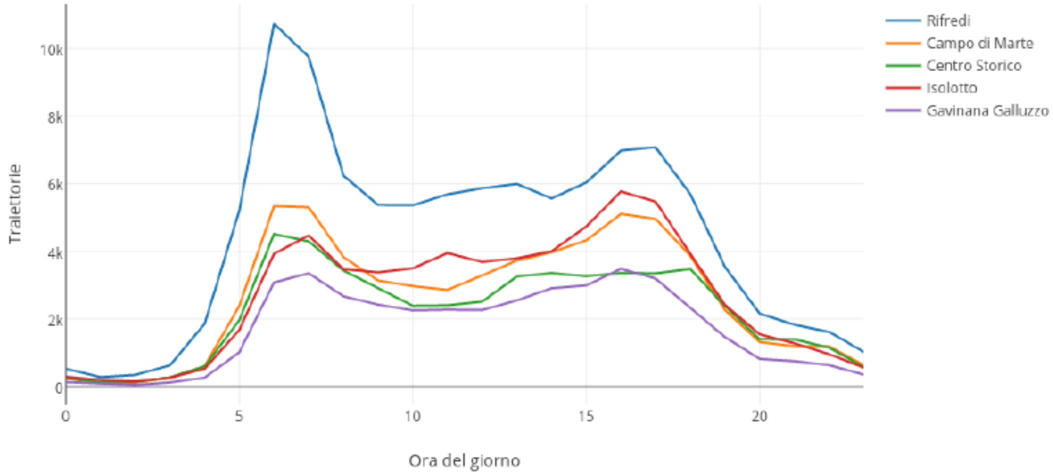


The **Florence Metropolitan Area** **is an high-density zone with a** **high concentration** **of economic activities and services that is crucial for the economic development of the whole region. Its** **functional form does not correspond** **with any precise administrative boundary.**

Additional analytical capacity

We might use GPS validated data to **assess the functional relations** within the whole area using **both systematic and non-systematic** mobility.





Since the data is very micro by nature, **spatial and temporal disaggregation** of mobility patterns and accessibility indexes is feasible.

Future lines of research

- Increase the performance of LPT estimations by improving the accessibility index and the estimation process

- Validation of a more disaggregated matrix (i.e. sub-municipal areas such as OMI areas)

- Generalize the estimation approach to the non-systematic GPS flows

- Further estimation and validation using GSM data (call records)

THANK YOU FOR YOUR ATTENTION

chiara.bocci@irpet.it
leonardo.piccini@irpet.it