Big Data Analytics

FOSCA GIANNOTTI AND LUCA PAPPALARDO

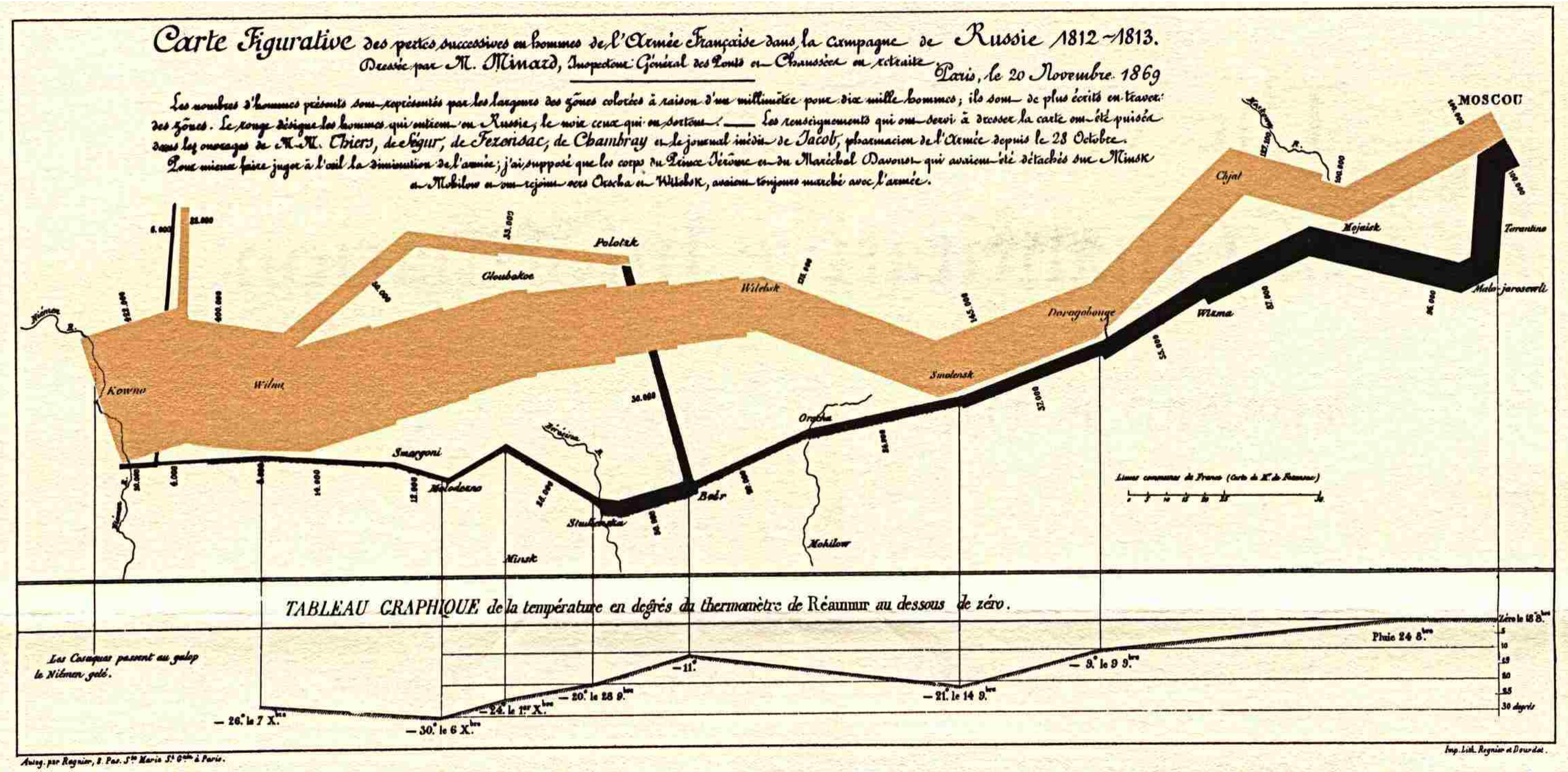
HTTP://DIDAWIKI.DI.UNIPI.IT/DOKU.PHP/BIGDATAANALYTICS/BDA/

DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2018/2019

Mobility Data Mining

MOBILITY DATA ANALYSIS FOUNDATIONS

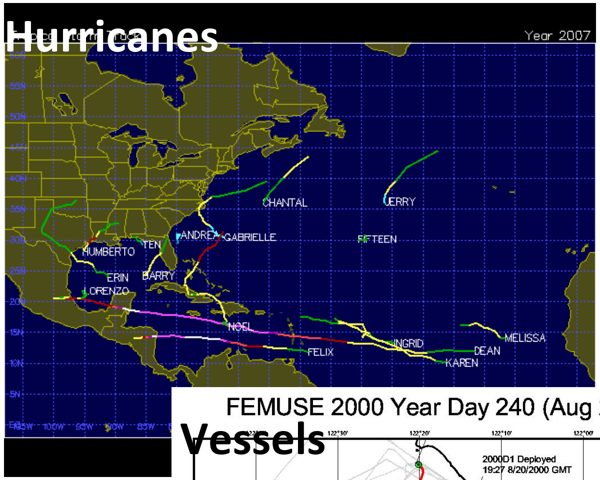
Understanding Human Mobility: a long path

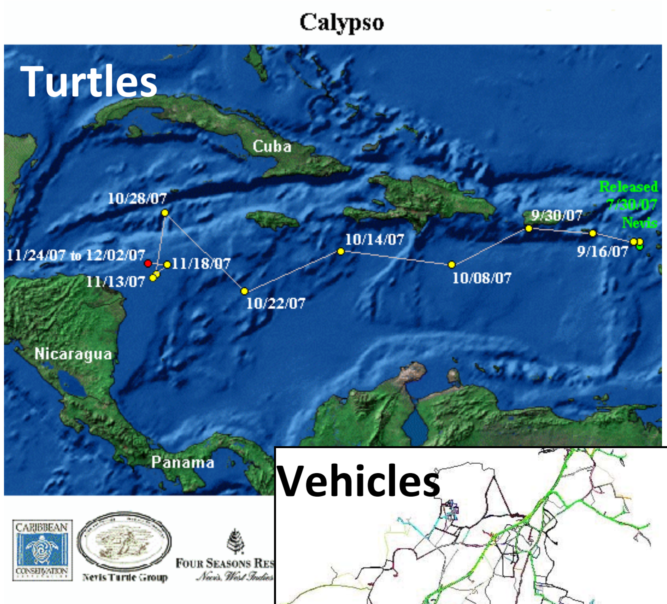


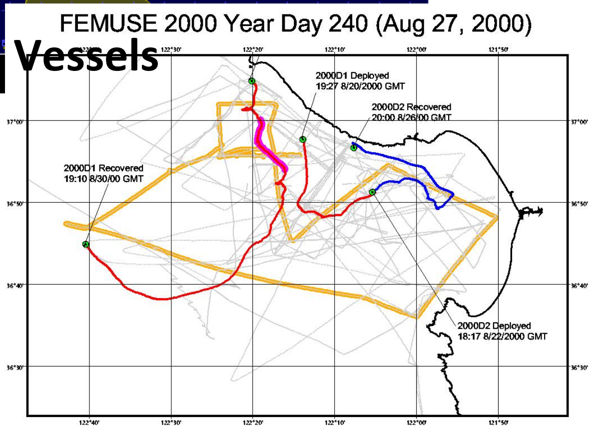
Charles Minard. "Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813", 1869.

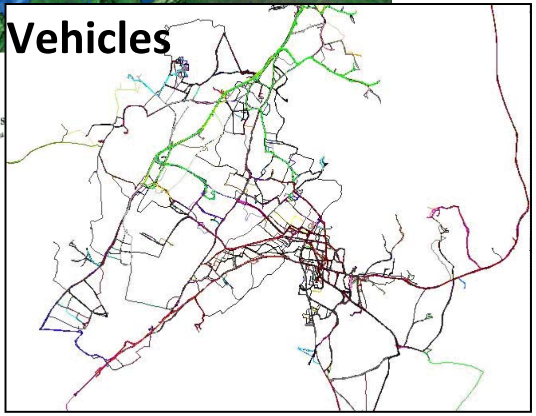
g Object Data

- Several domains:





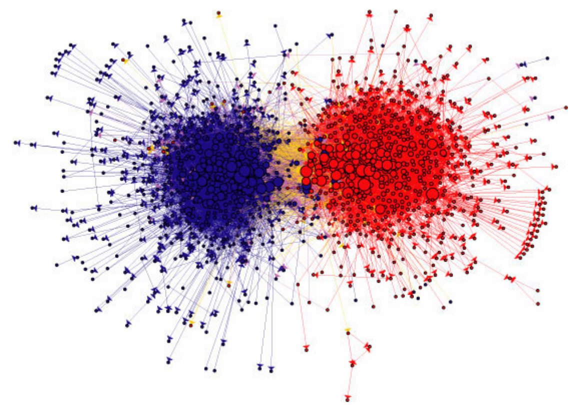




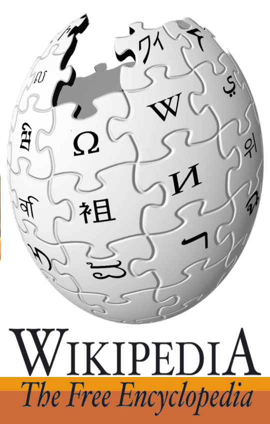
The novelty : BIG DATA

we buy

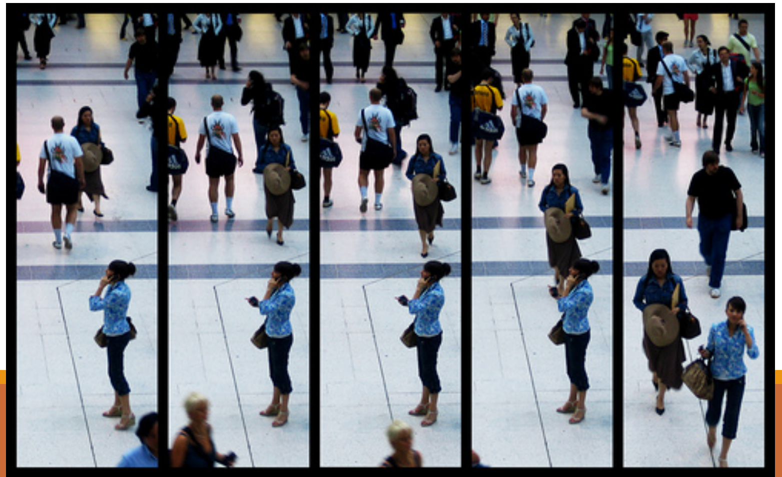
Whom we interact with



search for



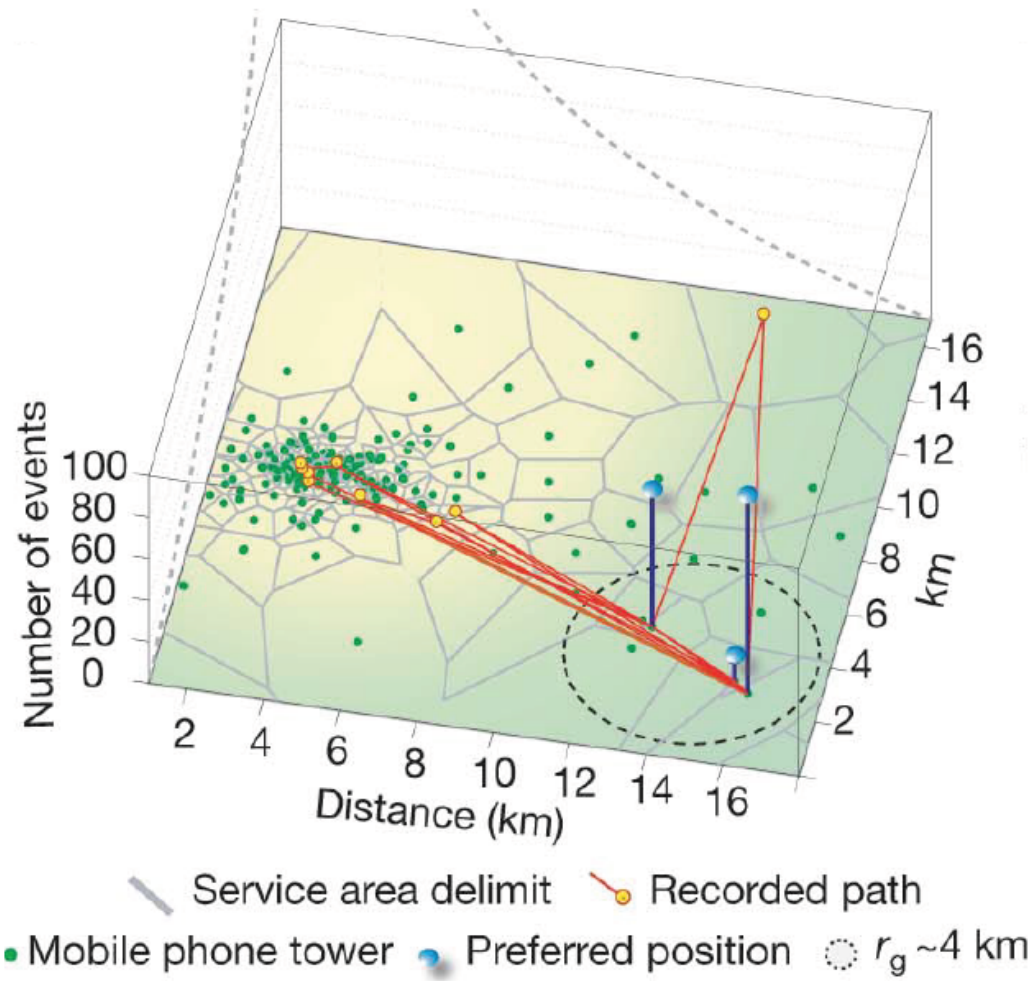
Where we go



Why Mining Moving Object Data?

- Large diffusion of mobile devices, mobile services and location-based services

Country-wide mobile phone data



when you call

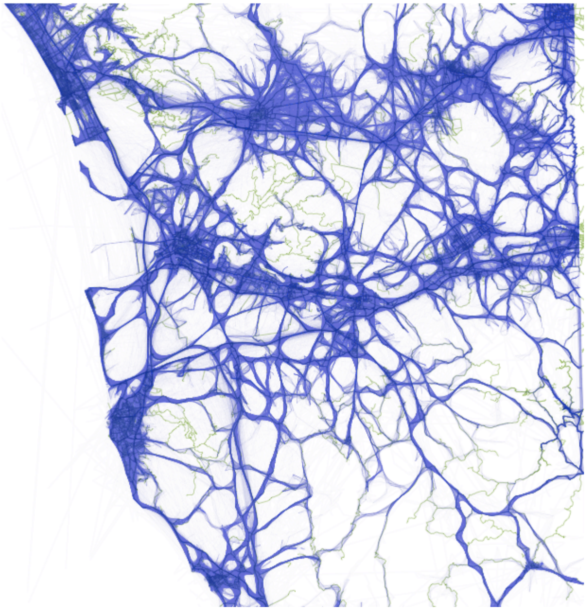
where you call

who you call

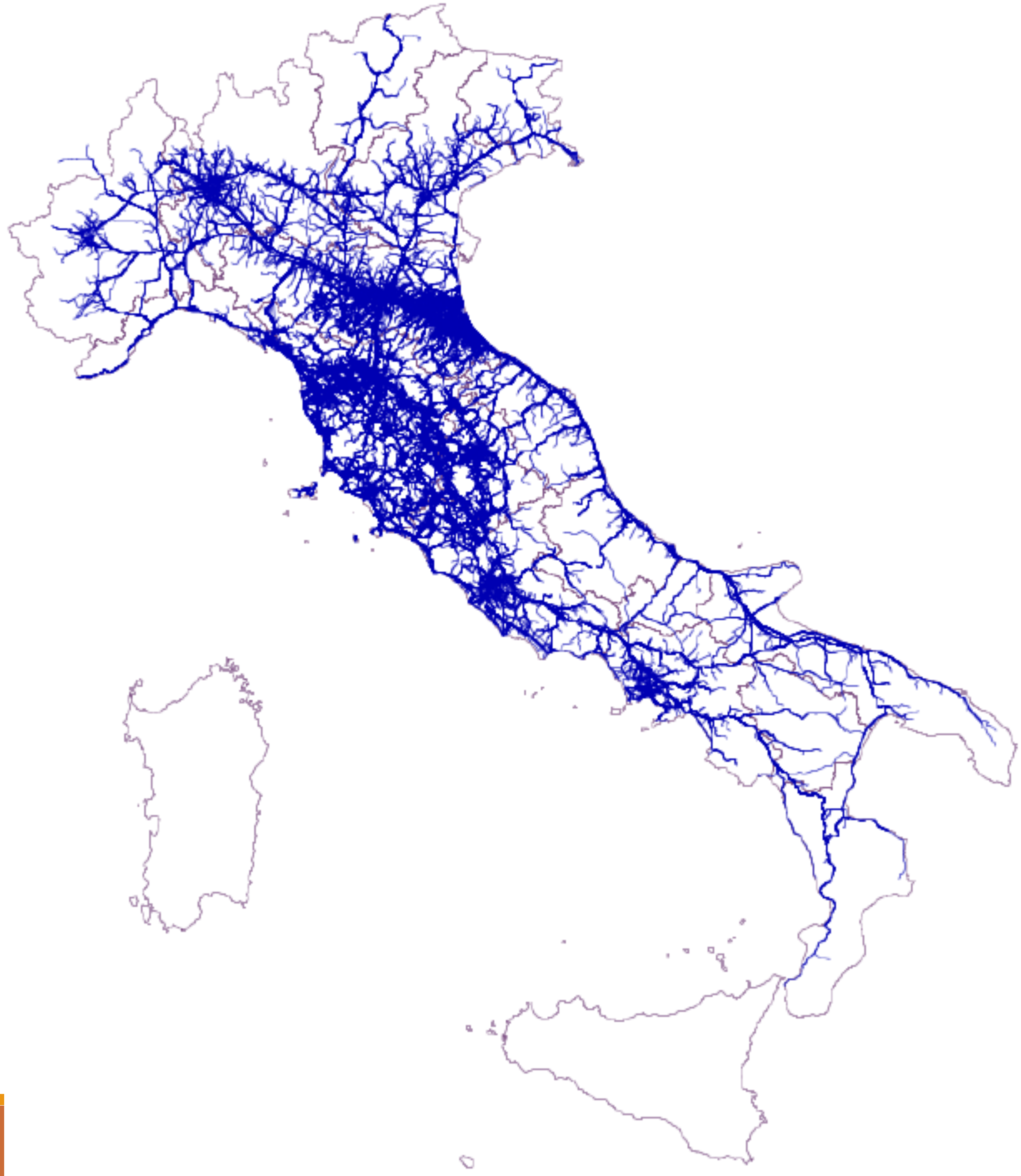
GPS tracks

Onboard navigation devices send GPS tracks to central servers

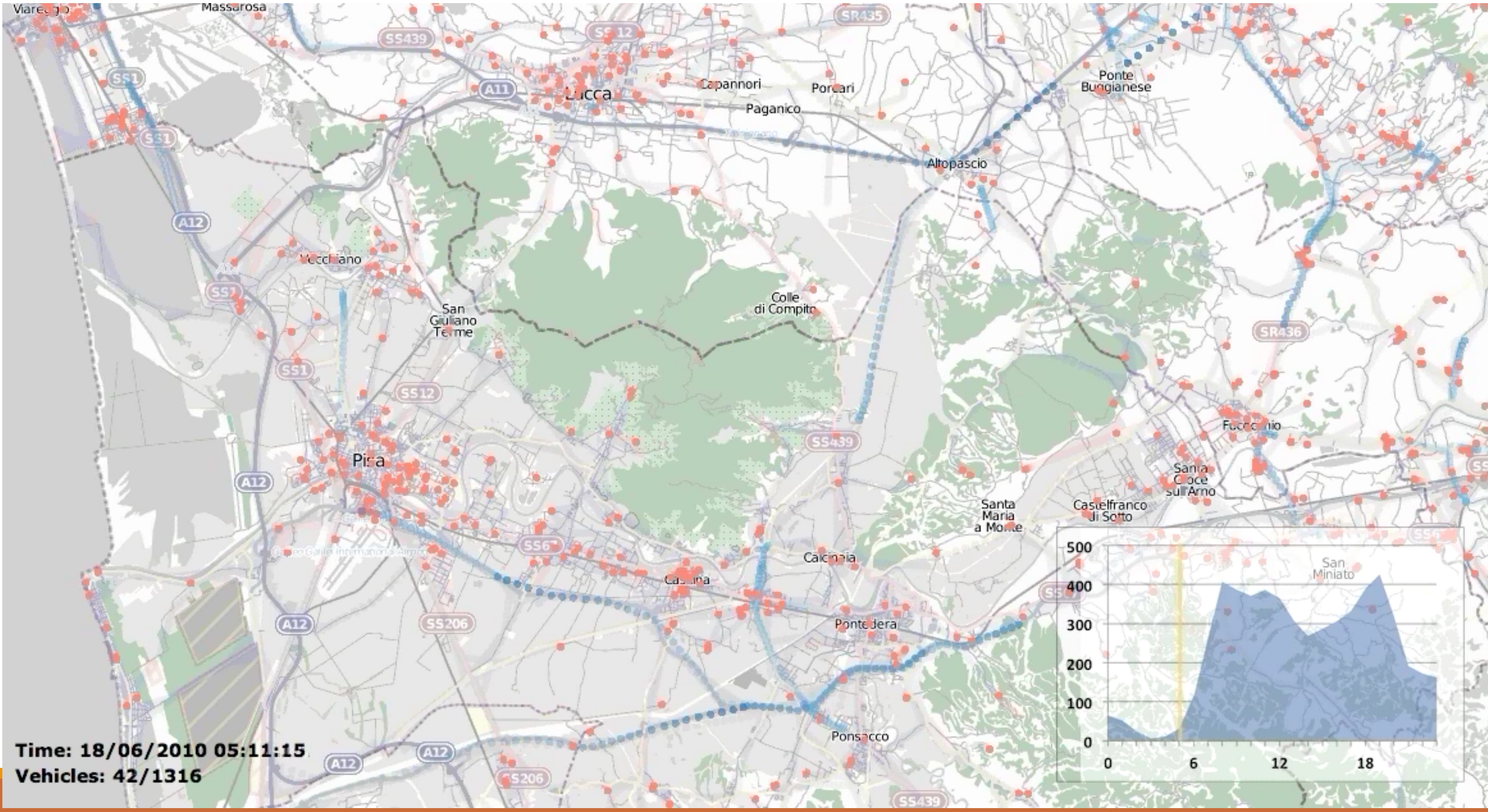
Ide;Time;Lat;Lon;Height;Course;Speed;PDOP;State;NSat
…
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
8;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
8;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
8;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
8;22/03/07 08:52:06;50.777185;7.206202; 68.1;124.1;30.058;3.8;1808;4
8;22/03/07 08:52:09;50.777057;7.206522; 67.9;117.7;34.003;3.8;1808;4
8;22/03/07 08:52:12;50.776925;7.206858; 66.9;117.5;37.151;3.8;1808;4
8;22/03/07 08:52:15;50.776813;7.207263; 67.0;99.2;39.188;3.8;1808;4
8;22/03/07 08:52:18;50.776780;7.207745; 68.8;90.6;41.170;3.8;1808;4
8;22/03/07 08:52:21;50.776803;7.208262; 71.1;82.0;35.058;3.8;1808;4
8;22/03/07 08:52:24;50.776832;7.208682; 68.6;117.1;11.371;3.8;1808;4
…



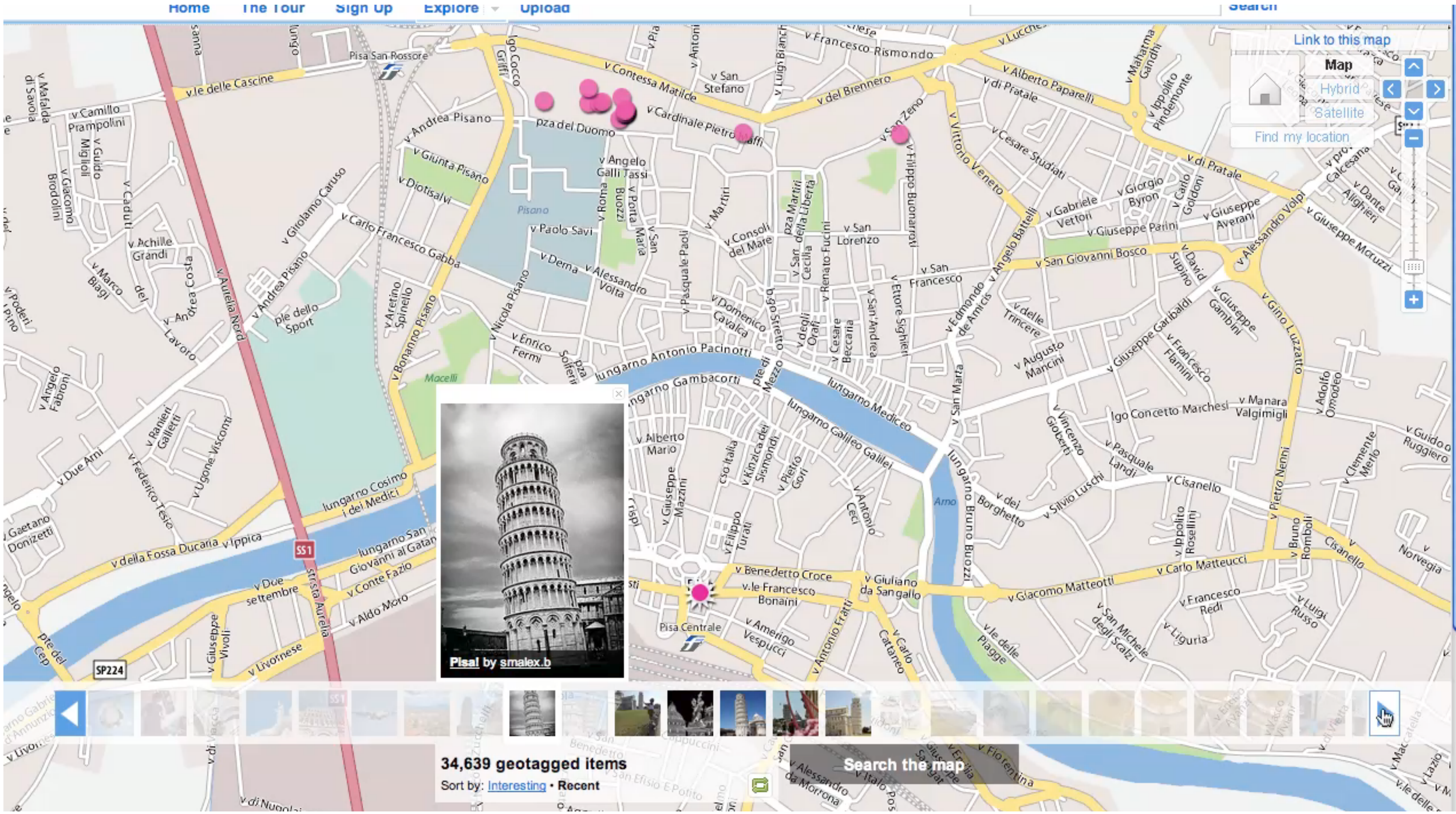
Sampling rate from few secs to 1-2 minutes



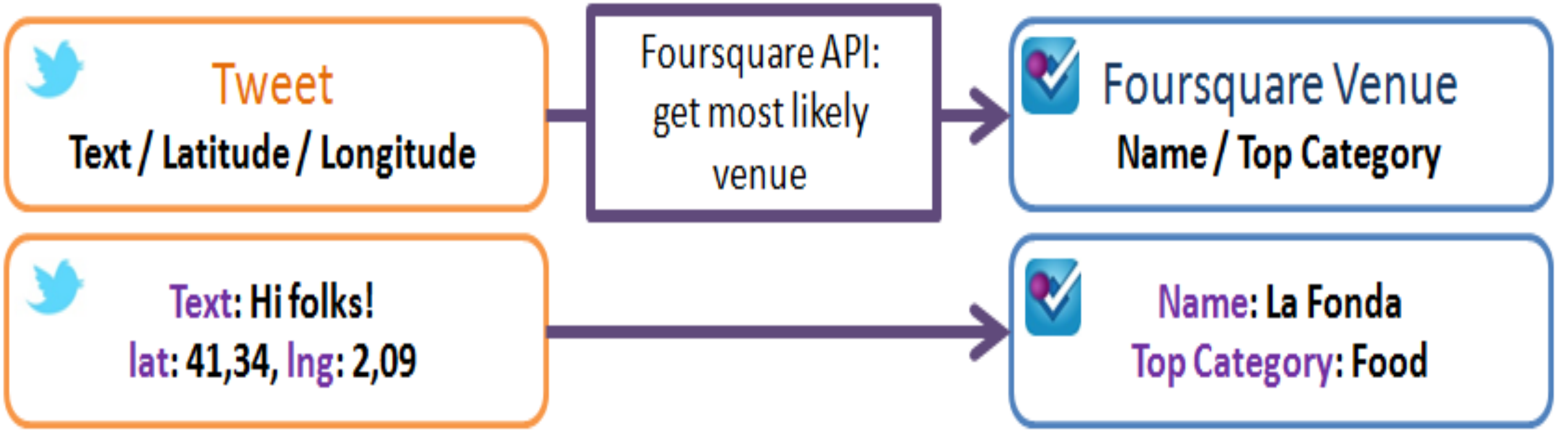
Urban Mobility Complexity: vehicles



Social networks



Twitter



Tweet
Text / Latitude / Longitude
Foursquare API: get most likely venue
Foursquare Venue
Name / Top Category
Text: Hi folks!
lat: 41,34, lng: 2,09
Name: La Fonda
Top Category: Food

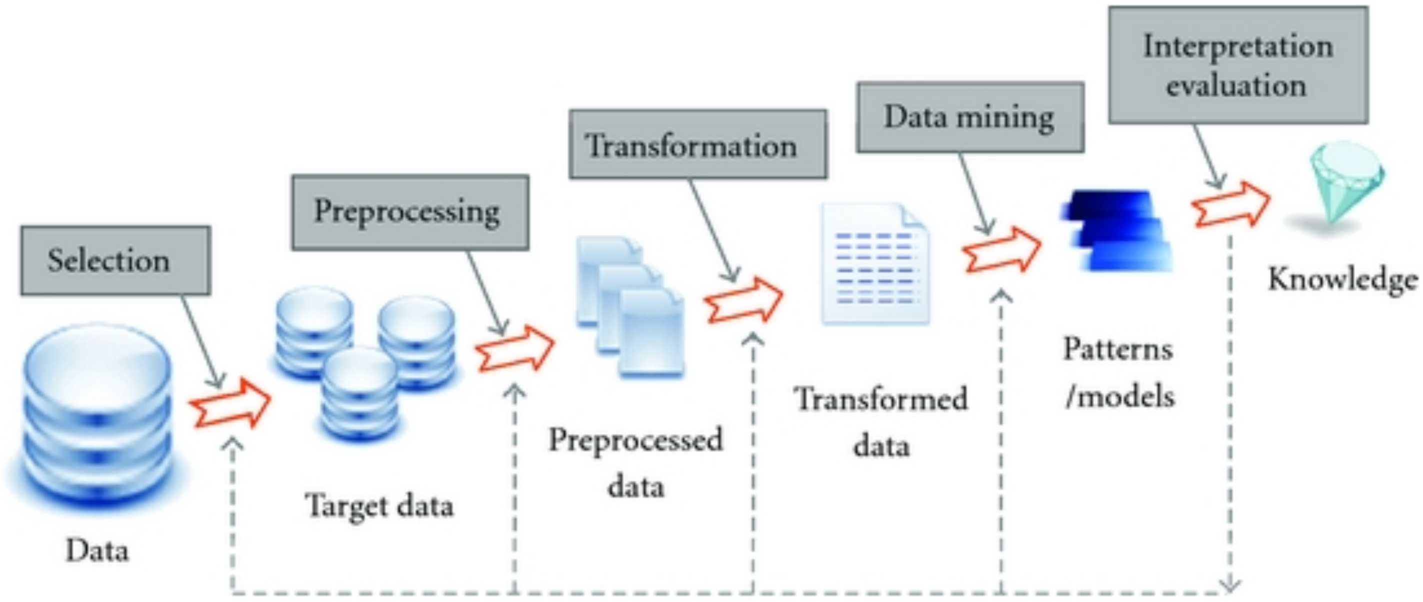
rch Impacts

- Moving object and trajectory data mining has many importan
  - Ecological analysis (e.g., animal scientists)
  - Weather forecast
  - Traffic control
  - Location-based services
  - Homeland security (*e.g.*, border monitoring)
  - Law enforcement (*e.g.*, video surveillance)
  - …

t Data

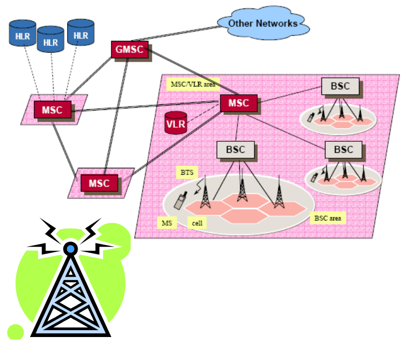
- Uncertainty
  - Sampling rate could be inconstant: From every few seconds tr
  - Data can be sparse: A recorded location every 3 days
- Noise
  - Erroneous points (e.g., a point in the ocean)
- Background
  - Cars follow underlying road network
  - Animals movements relate to mountains, lakes, ...
- Movement interactions
  - Affected by nearby moving objects

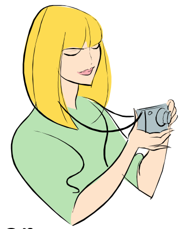
Knowledge Discovery process



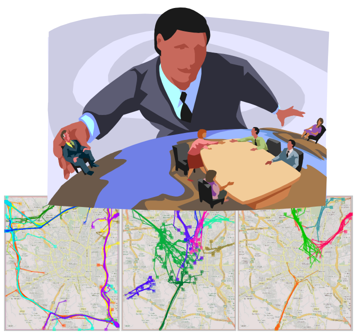
The KDD process for Mobility Data

Mobile phone data, GPS tracks

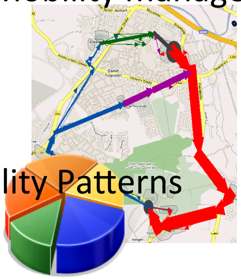




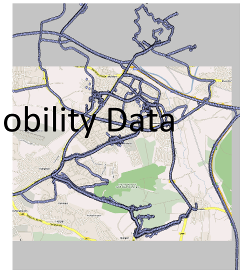
End user



Mobility manager



Mobility Patterns



Mobility Data

Raw data

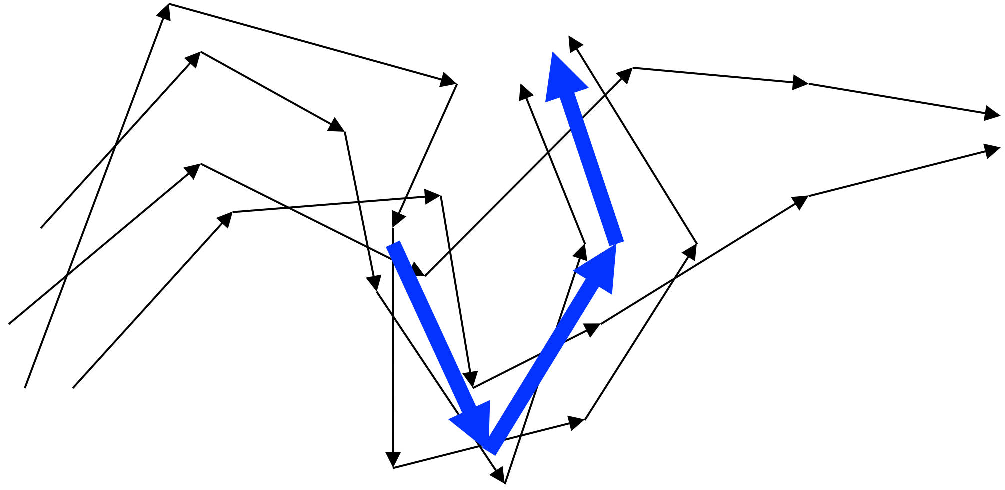
```
name|date|y|x
Prinzessin|08.20.1998|52.118|12.087
Prinzessin|08.23.1998|51.019|15.309
Prinzessin|08.26.1998|47.723|22.786
Prinzessin|08.29.1998|43.040|27.119
Prinzessin|08.31.1998|38.715|32.165
Prinzessin|[illegible].1998|37.195|35.255
Prinzessin|[illegible].1998|32.979|36.021
Prinzessin|09.05.1998|28.513|33.437
Prinzessin|09.06.1998|23.961|32.937
Prinzessin|09.07.1998|19.418|33.446
Prinzessin|09.12.1998|15.823|34.094
Prinzessin|10.11.1998|14.685|32.848
Prinzessin|11.03.1998|11.510|32.591
Prinzessin|11.24.1998|13.888|35.667
Prinzessin|12.08.1998|12.562|34.777
Prinzessin|12.10.1998|9.124|35.644
...
```

Data mining …

- … is about finding models that emerge directly from the data
  - Data-driven vs hypothesis-driven analysis
- Local models
  - **Patterns**: find groups of items/events that frequently co-occur in the data
- Global models
  - **Clustering**: find a natural partition of the data into groups of similar objects
  - **Classification**: find a function that predicts the value

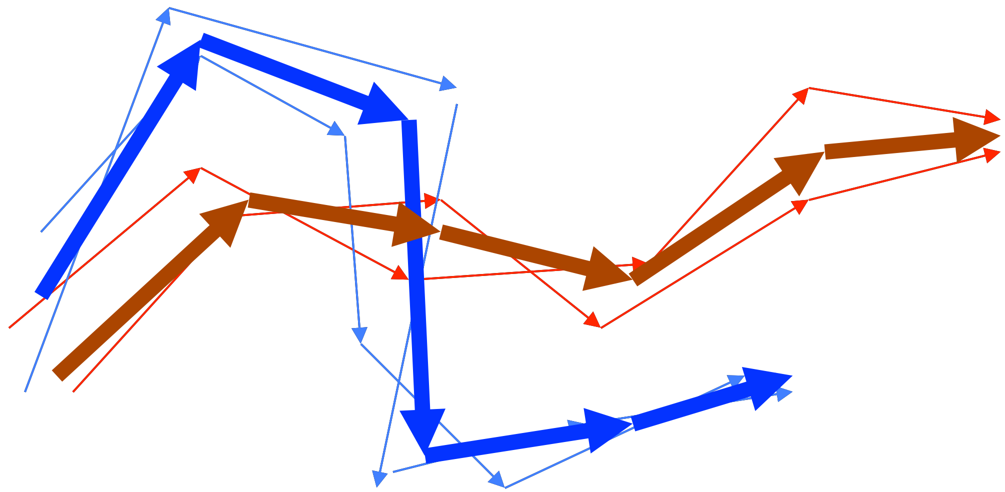
Trajectory patterns

- Discover frequently followed itineraries



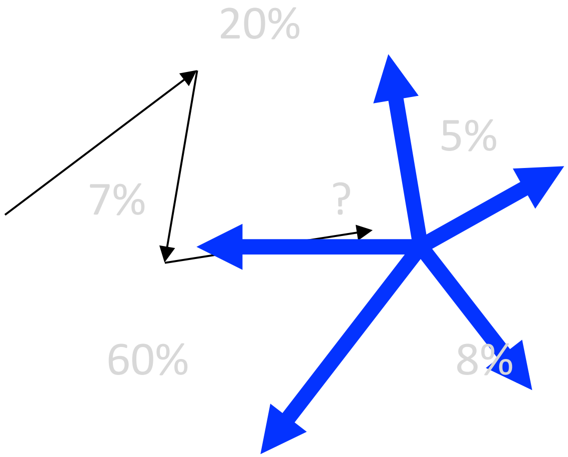
Trajectory Clustering

- Group together similar trajectories
- For each group produce a summary



prediction

- Extract behaviour rules from history
- Use rules to predict behaviour of future users



GPS processing and statistics

Raw GPS Data

ID Timestamp Latitude Longitude Others (optional)

946826,14/06/10 14:08:54,43964259,10216895,0,0,1,0,0

457380,13/06/10 22:05:27,43682201,10408320,0,0,3,0,0

457380,13/06/10 22:06:00,43682688,10408501,10,10,3,1,33

457380,13/06/10 22:06:34,43683609,10409146,14,24,3,1,115

457380,13/06/10 22:07:09,43685653,10410117,52,18,3,1,241

457380,13/06/10 22:07:43,43689775,10412032,50,18,3,1,484

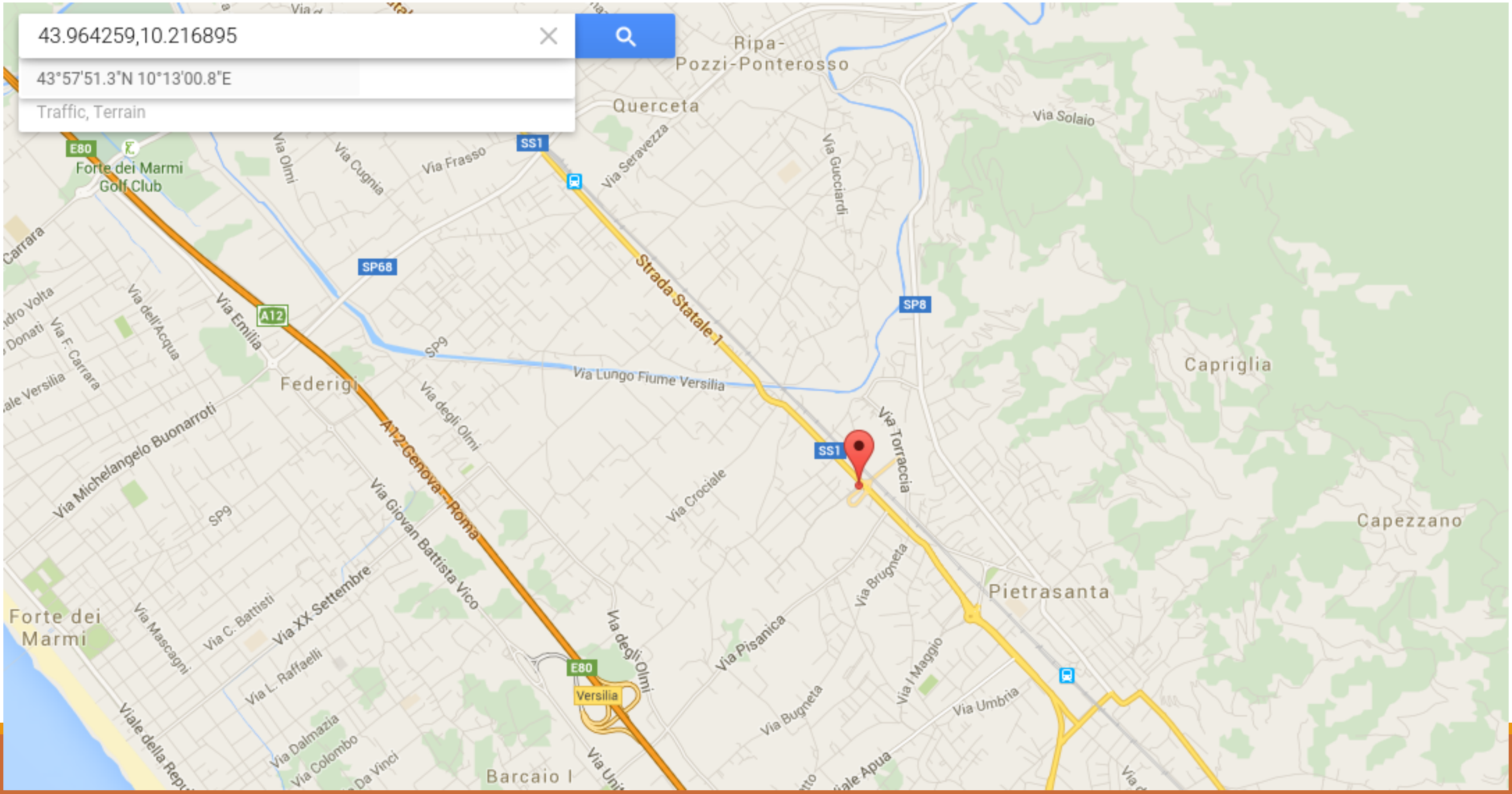
457380,13/06/10 22:08:19,43692906,10413910,32,356,3,1,401

457380,13/06/10 22:08:53,43690801,10415016,60,126,3,1,279

…

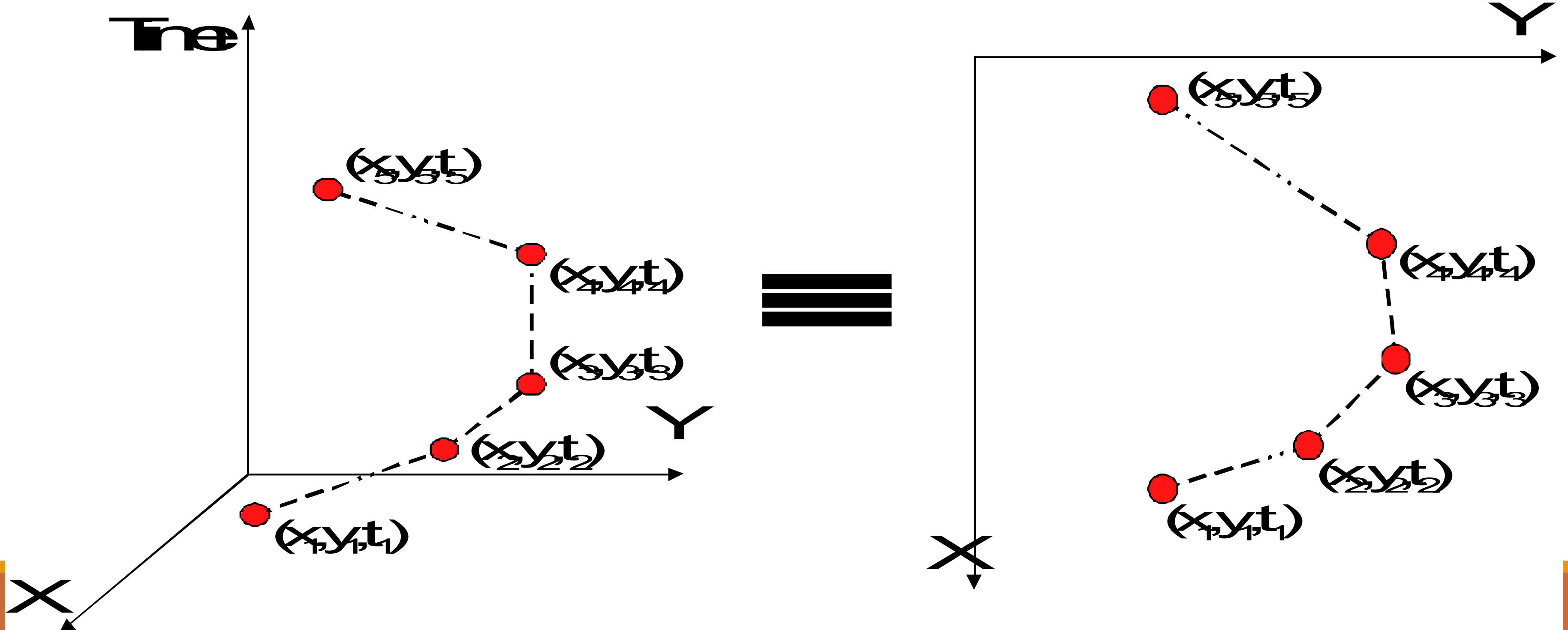
Sample point on the map

946826,14/06/10 14:08:54,43964259,10216895,0,0,1,0,0



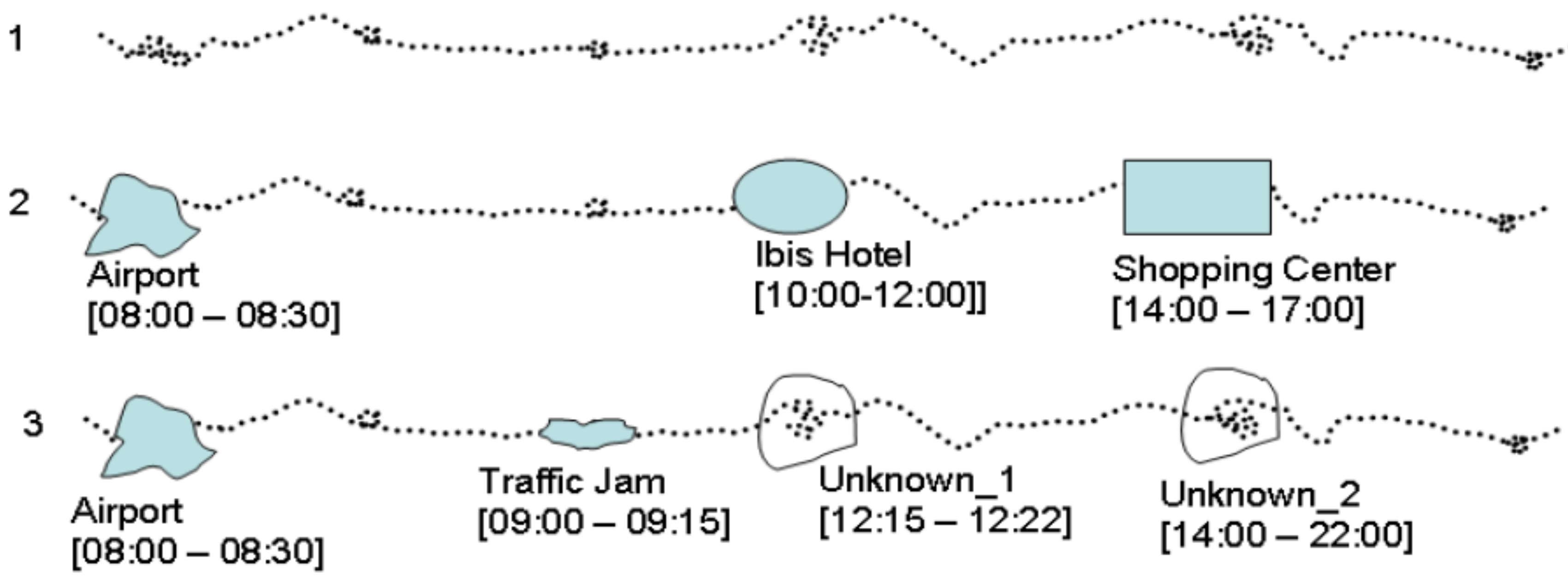
Trajectory data

- Mobility of an object is described by a set of trips
- Each trip is a trajectory, i.e. a sequence of time-stamped locations



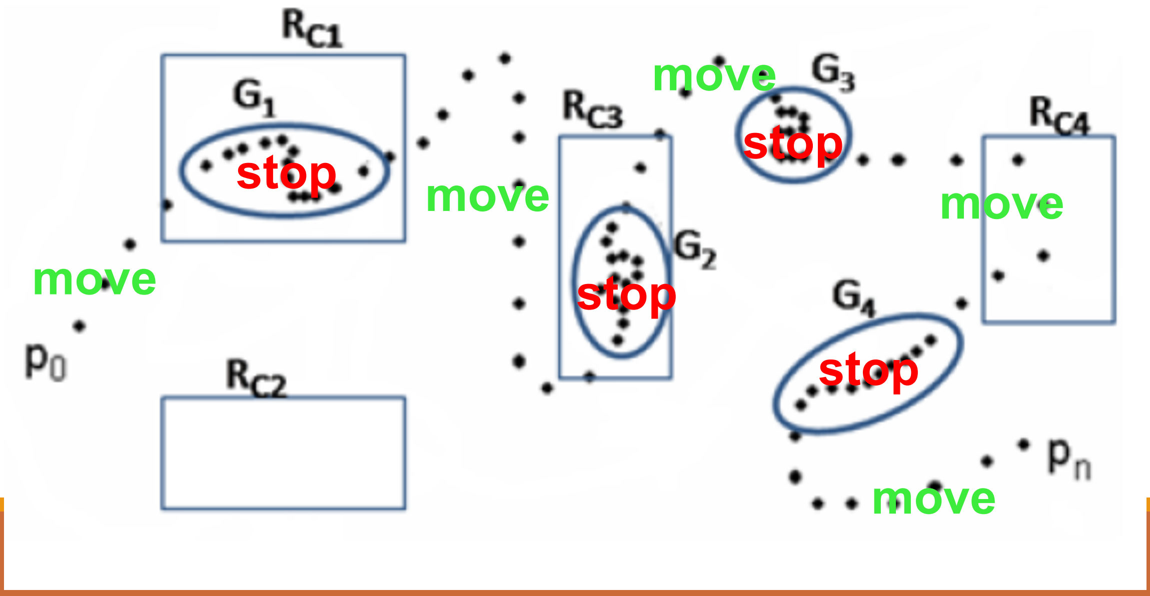
Trajectory reconstruction

- Raw data forms a continuous stream of points
- How to cut it into stops and trips?
  - Example on smart phone traces:



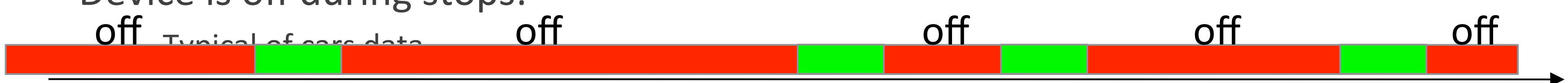
Trajectory reconstruction

- General criteria based on speed
  - If it moves very little (threshold Th_S) over a significant time interval (threshold Th_T) then it is practically a stop
  - Trajectory (trip) = contiguous sequence of points between two stops

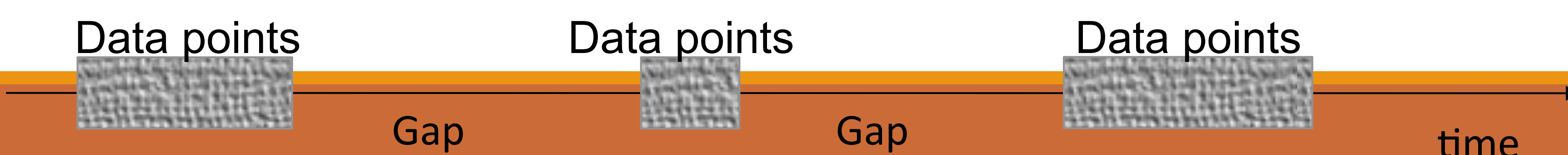


Trajectory reconstruction

- Special cases, easier to treat
  - Stop explicitly in the data: e.g. engine status on/off
    - Simply "cut" trajectories on status transitions
  - Device is off during stops:
    - Typical of cars data

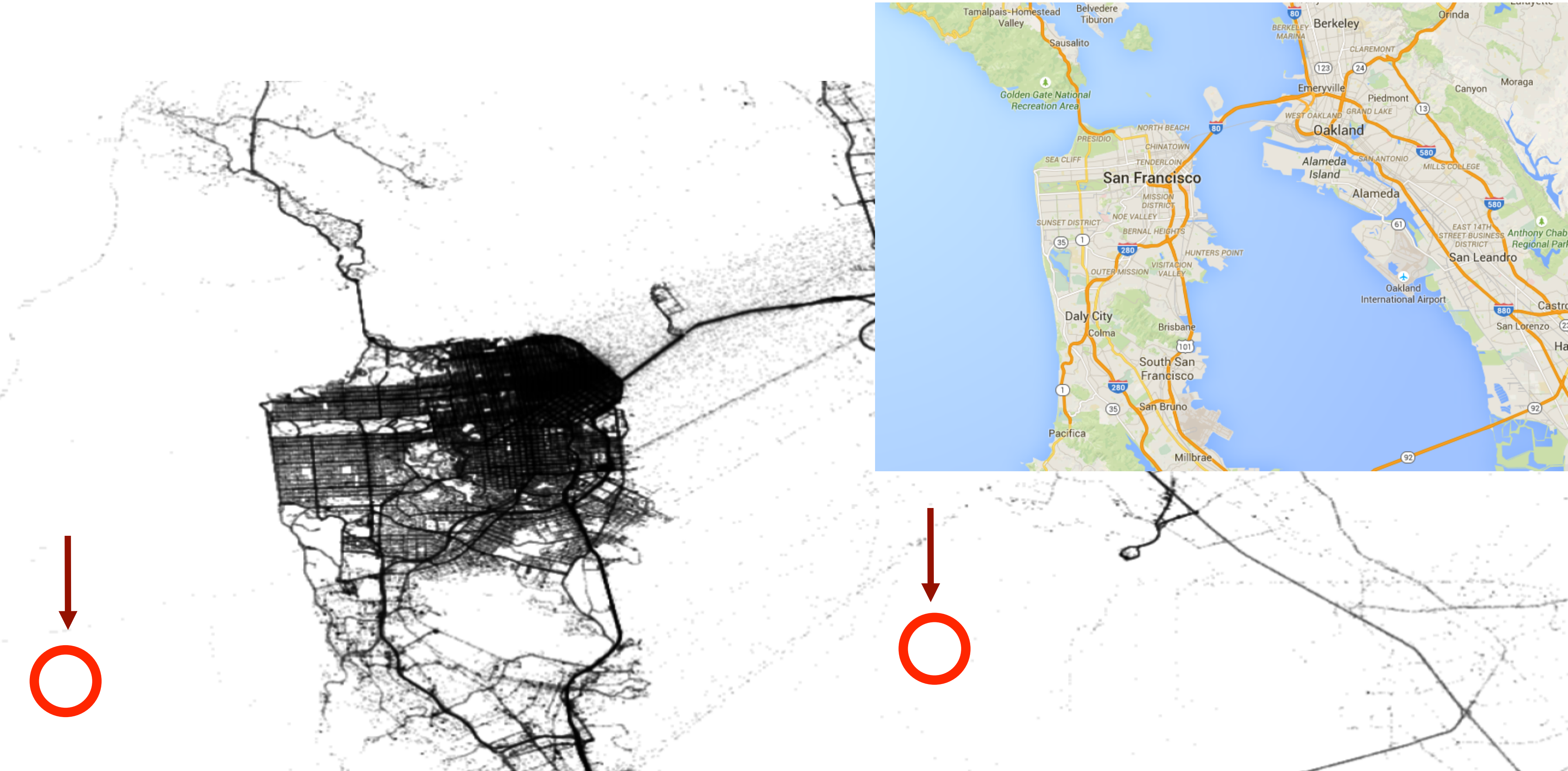


    - A stop results in a time gap in the data
    - Exceptions: short stops might remain undetected



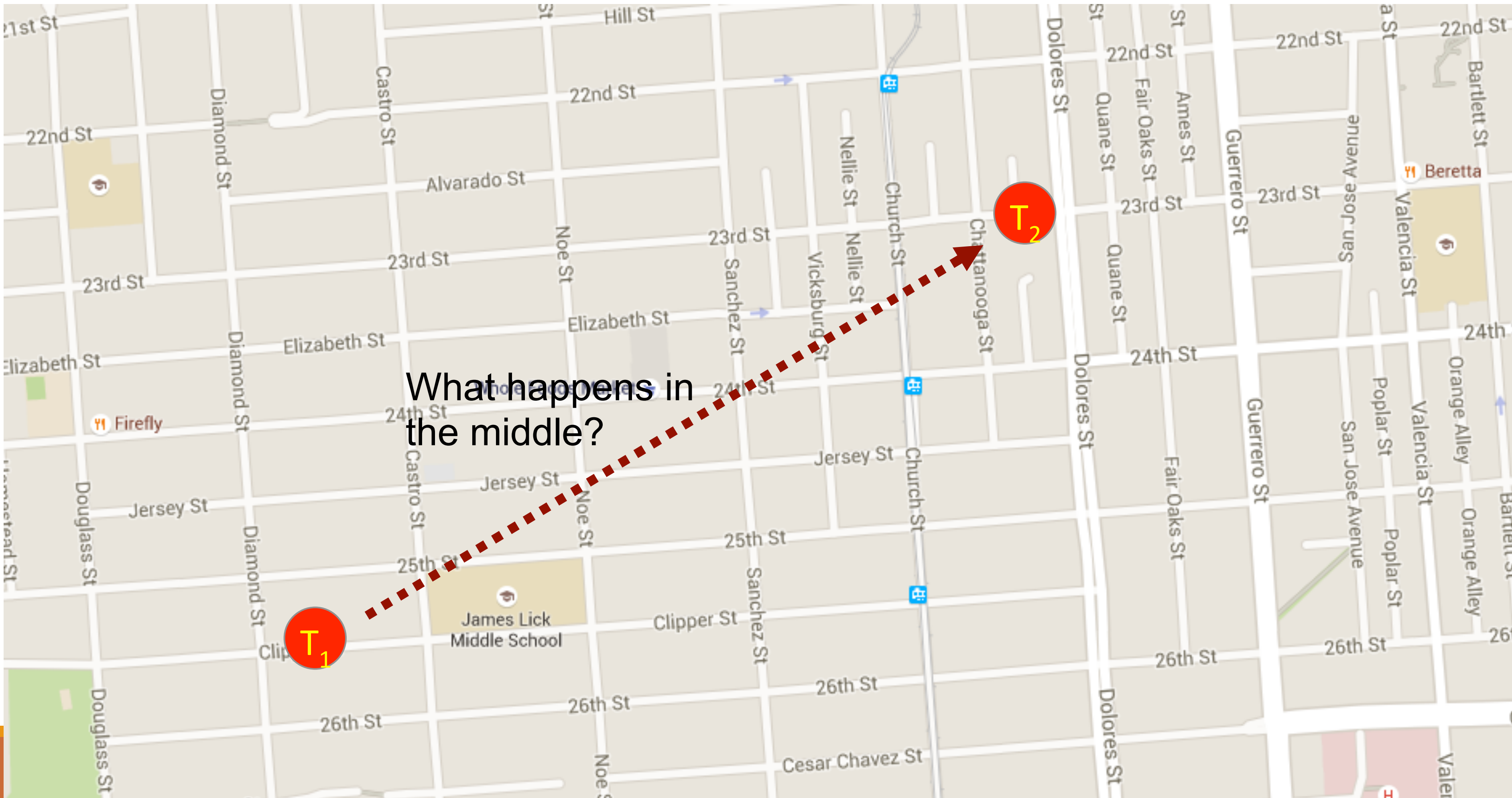
Outliers / noise

- Single points might contain errors of various kinds



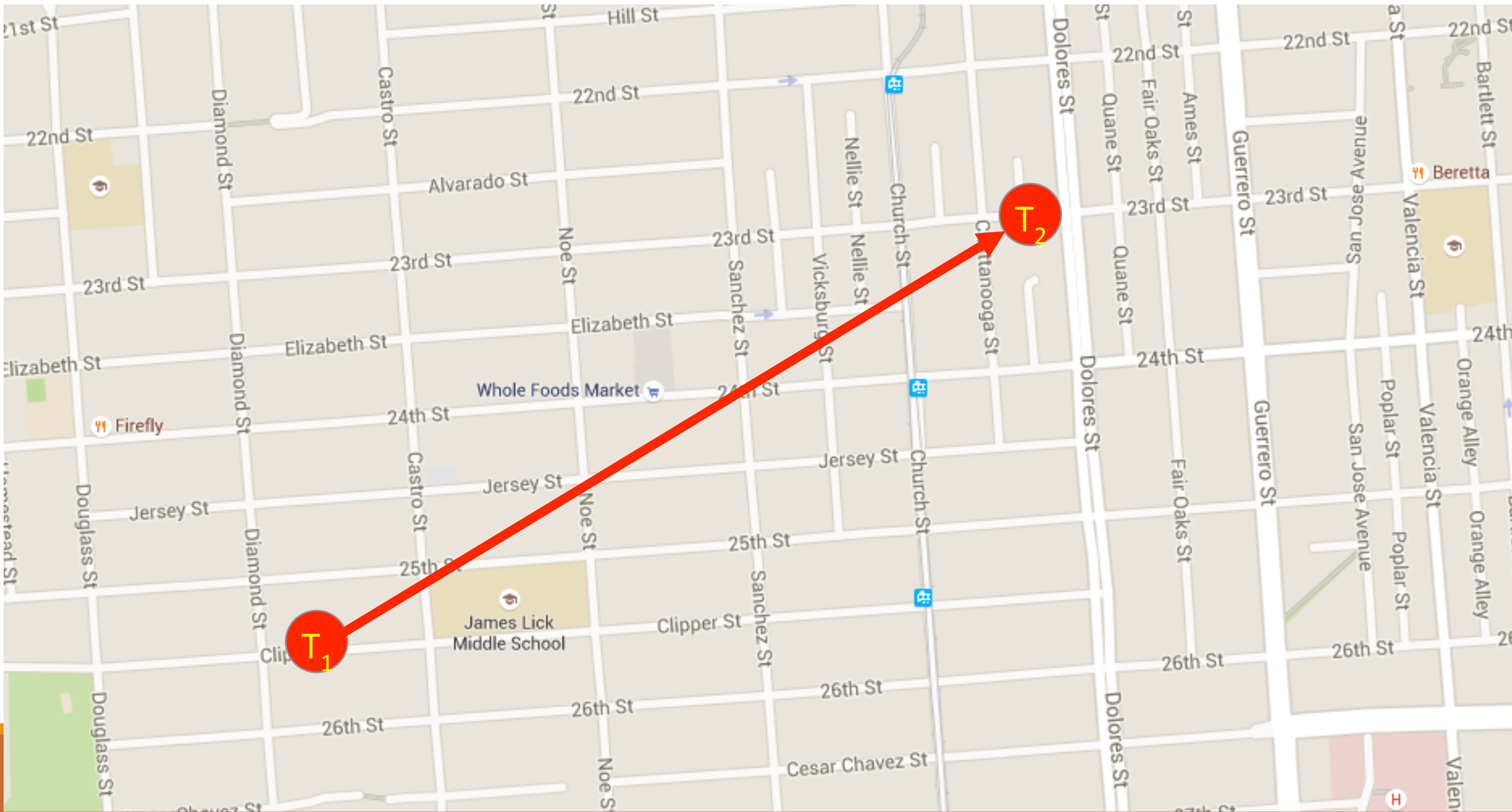
Gaps

- Sometimes the space/time gap between consecutive points is significant



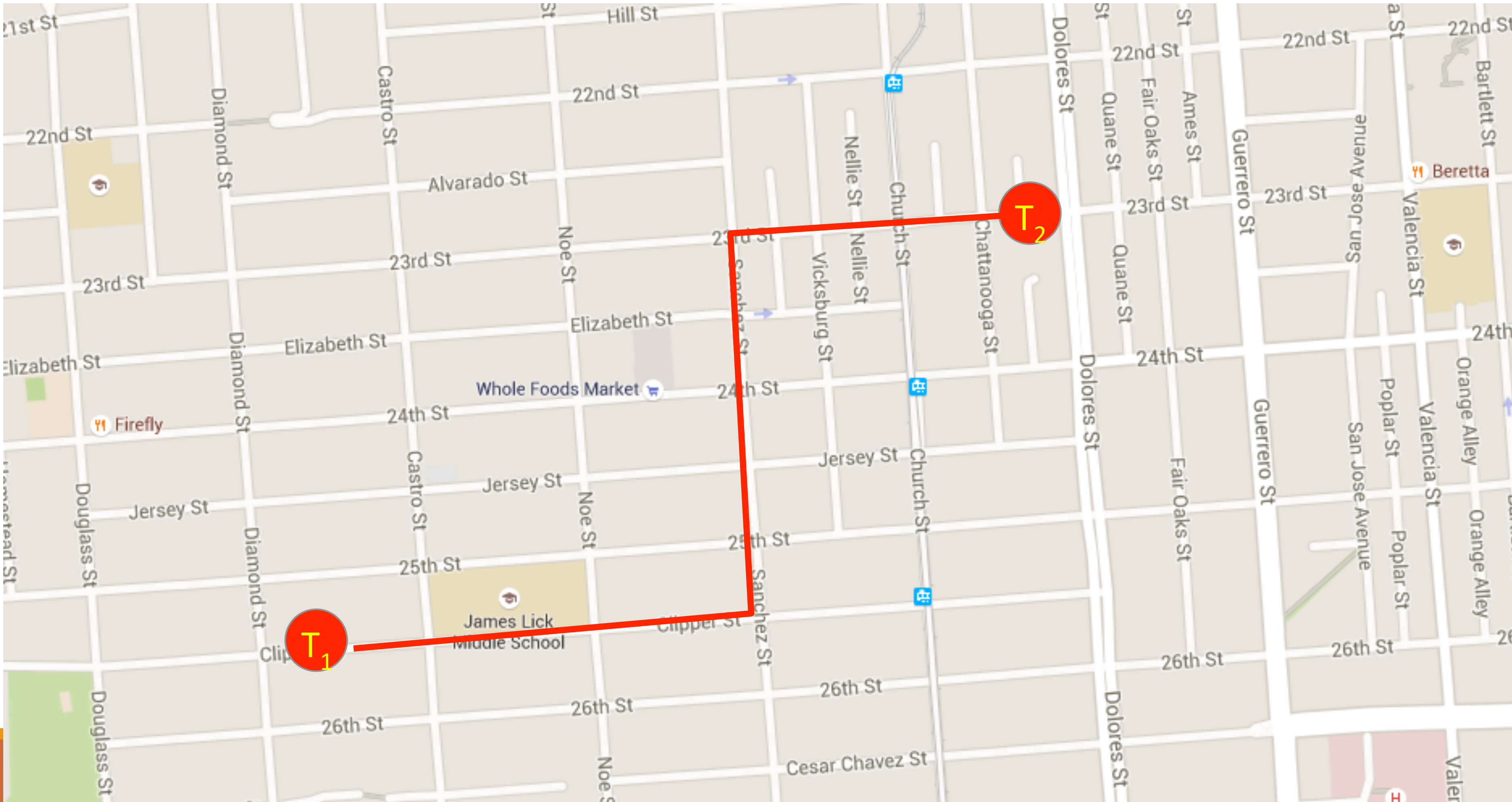
Free vs. constrained movement

- Typical solutions:
  - Free movement => straight line, uniform speed

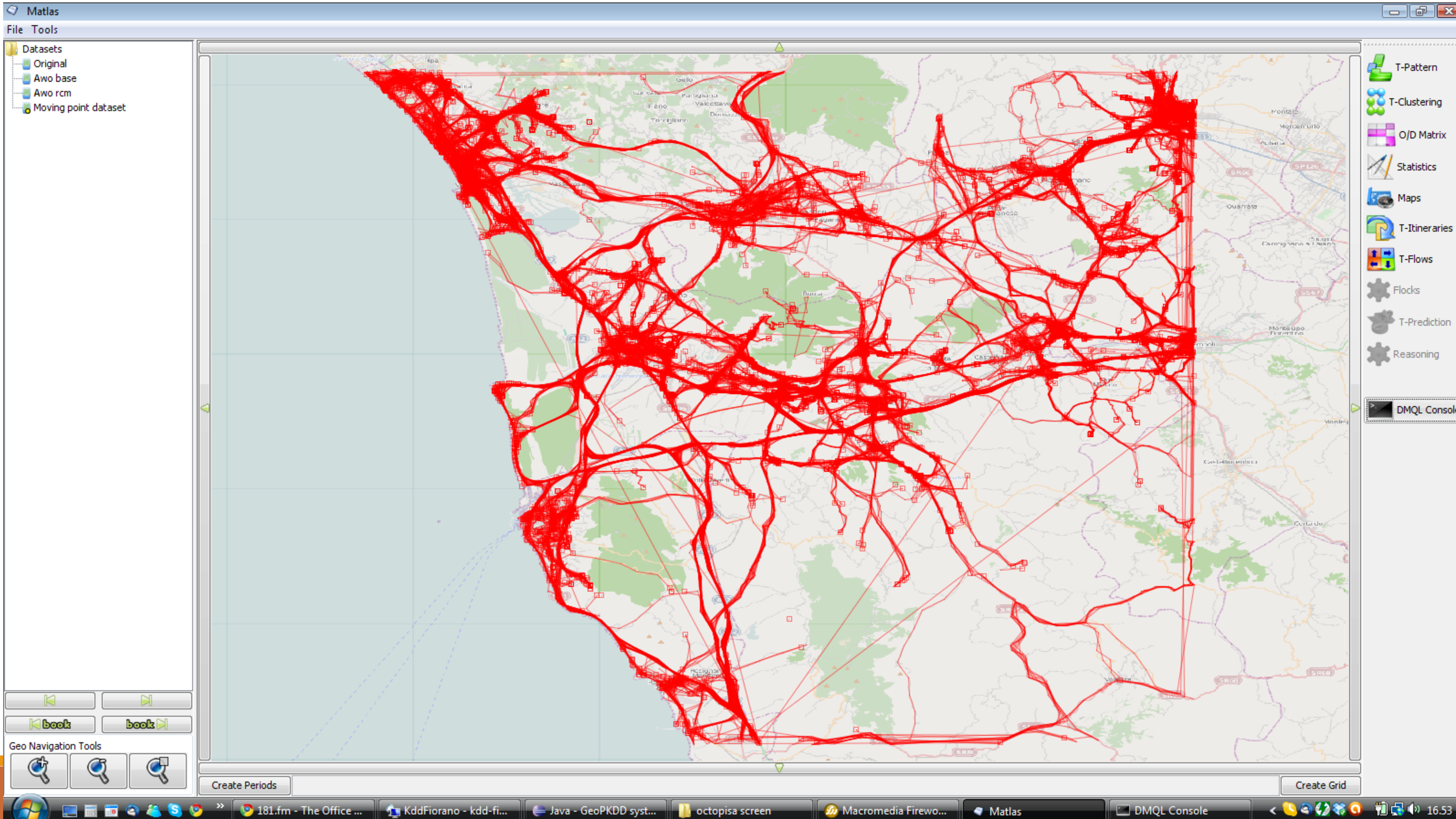


Free vs. constrained movement

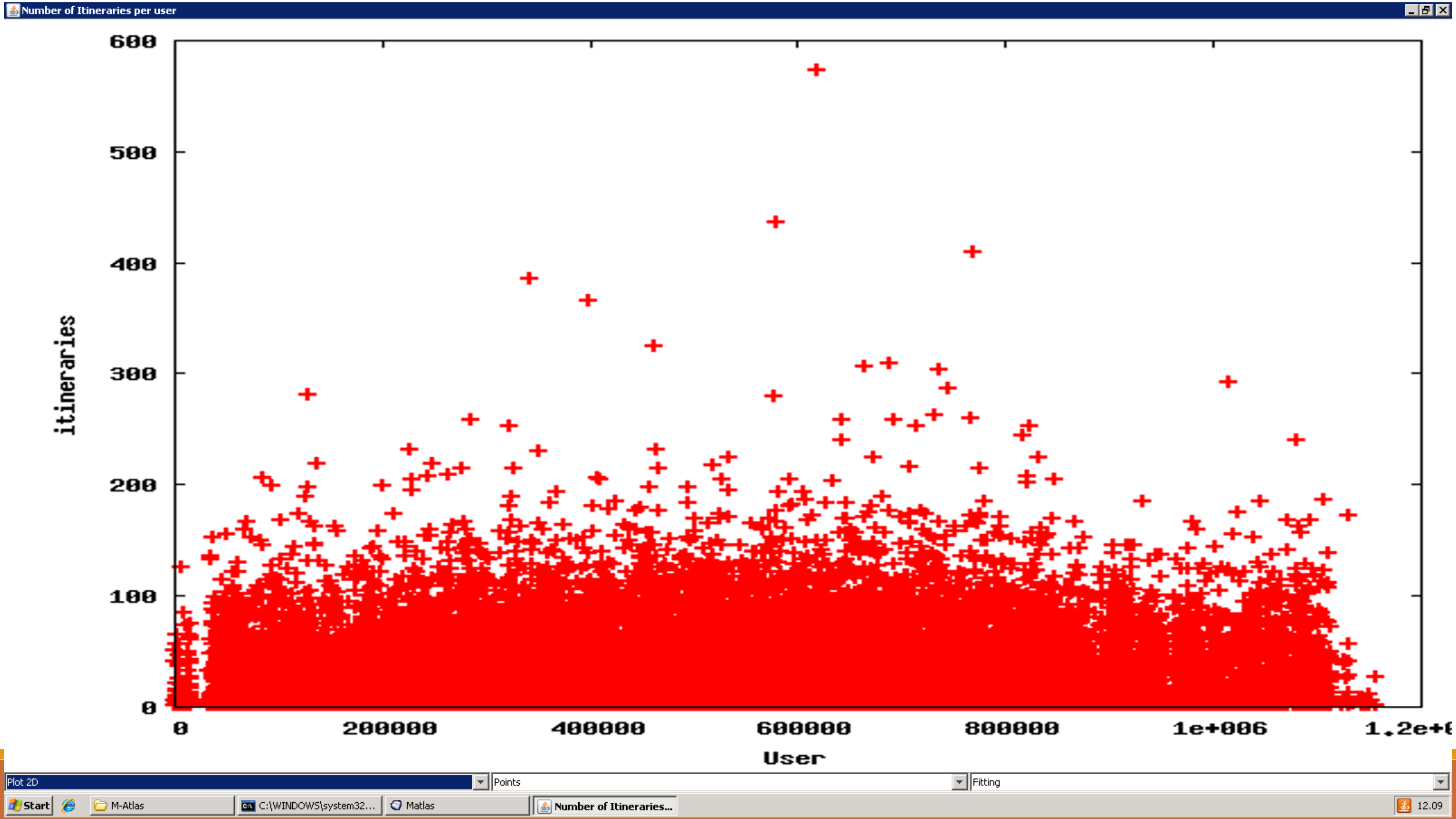
- Typical solutions:
  - Constrained movement => shortest path



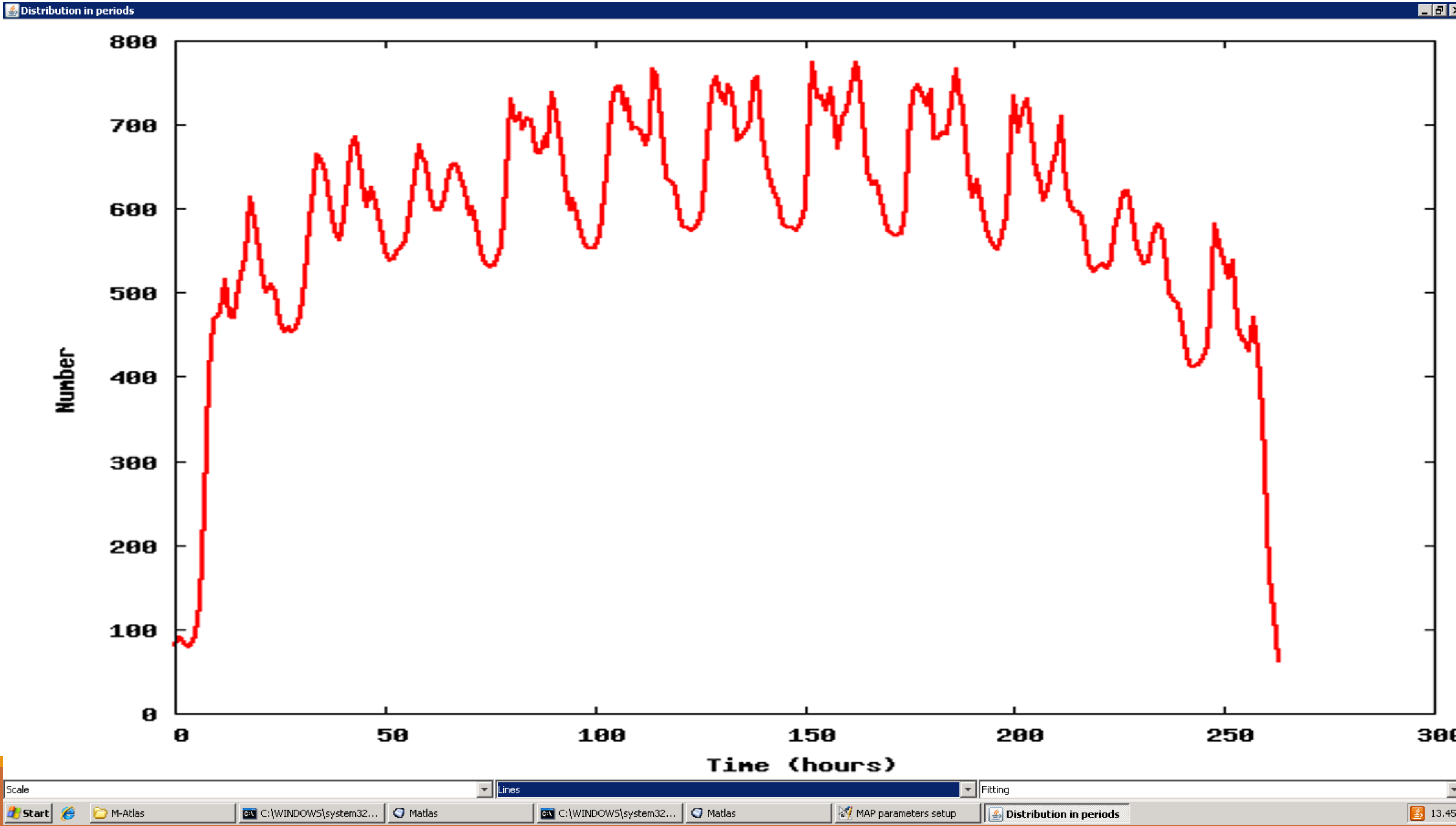
A Dataset (2/7 → 12/7)



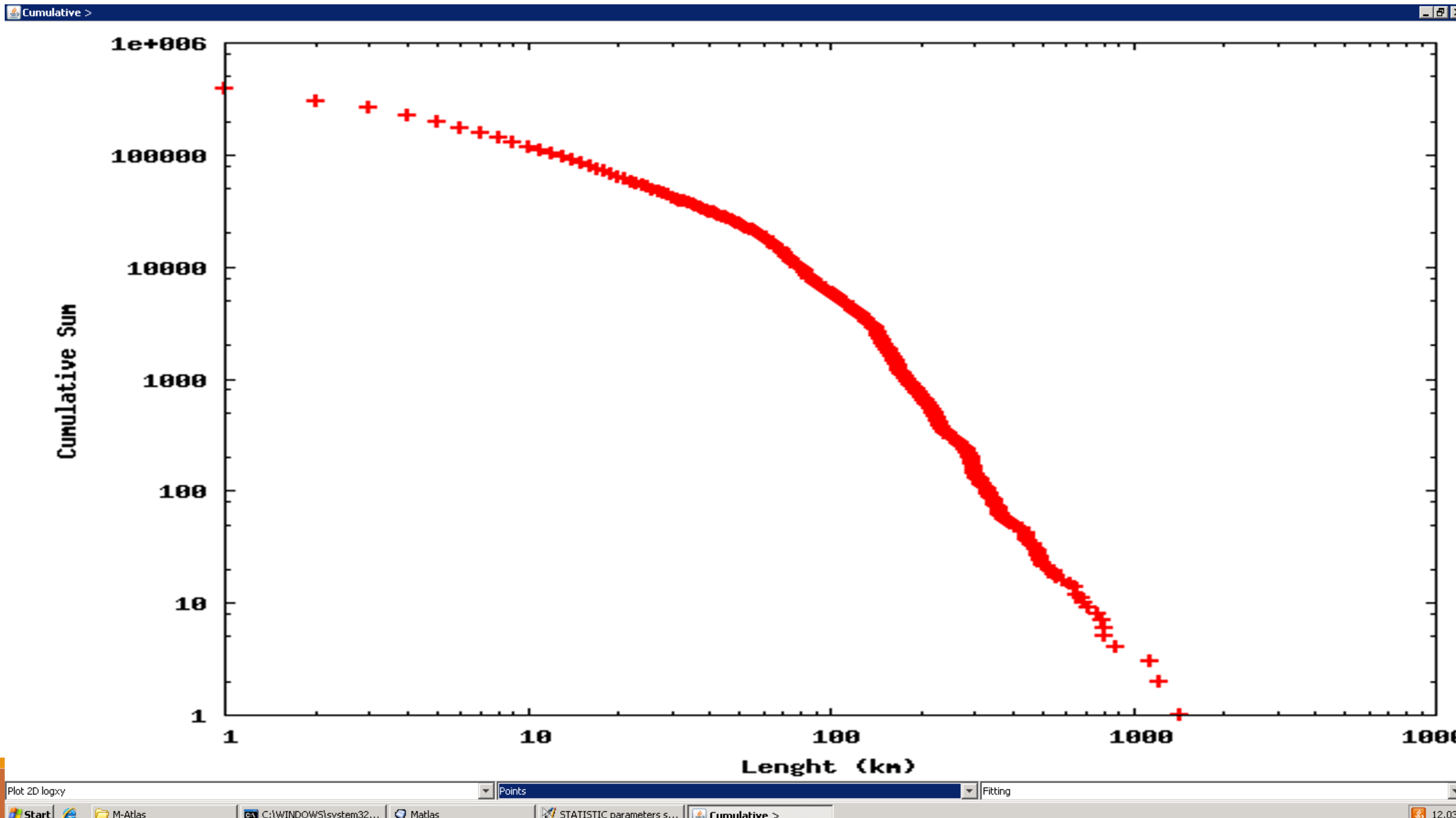
Number of trajectories per User



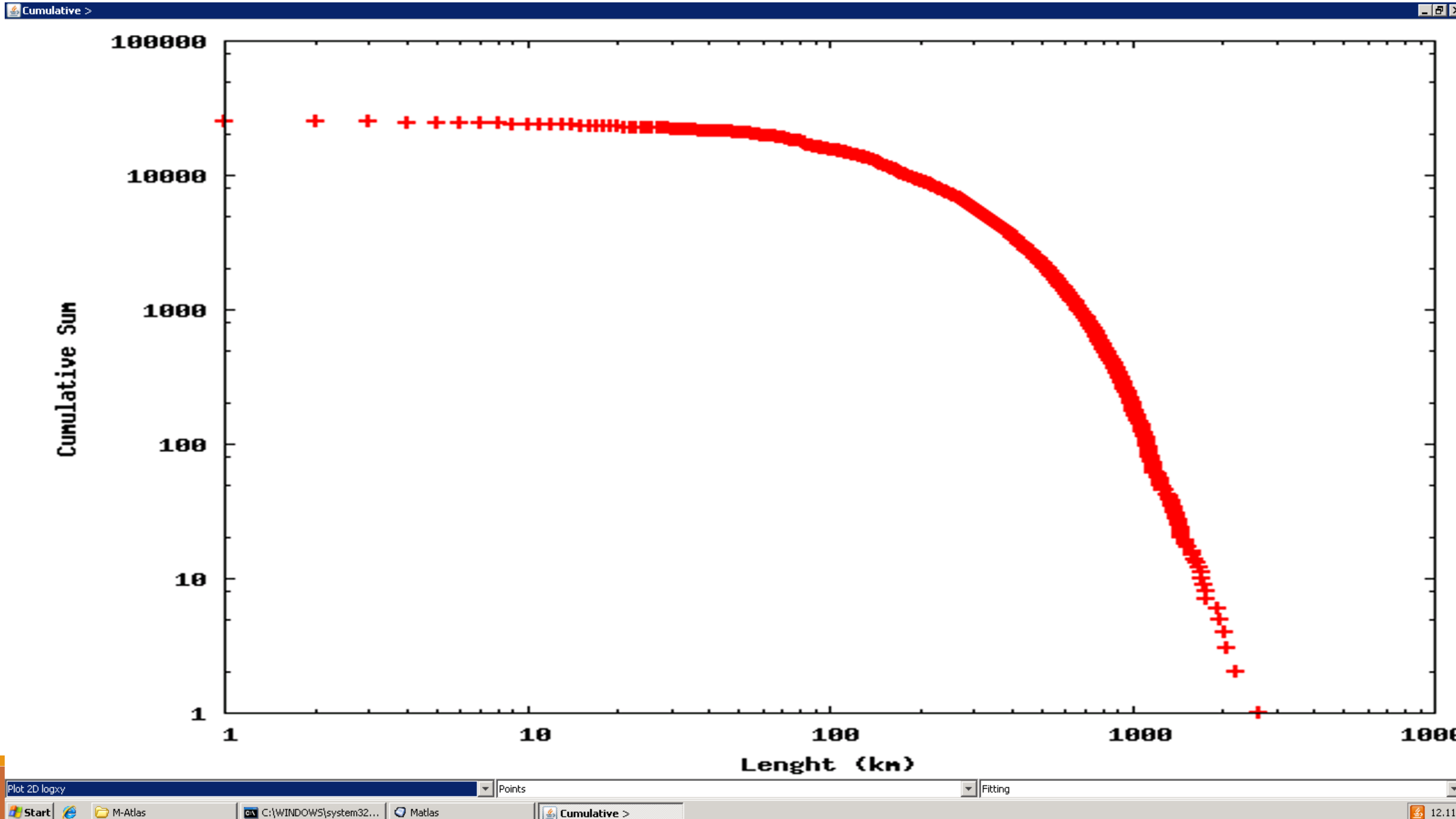
Distribution in periods (hours)



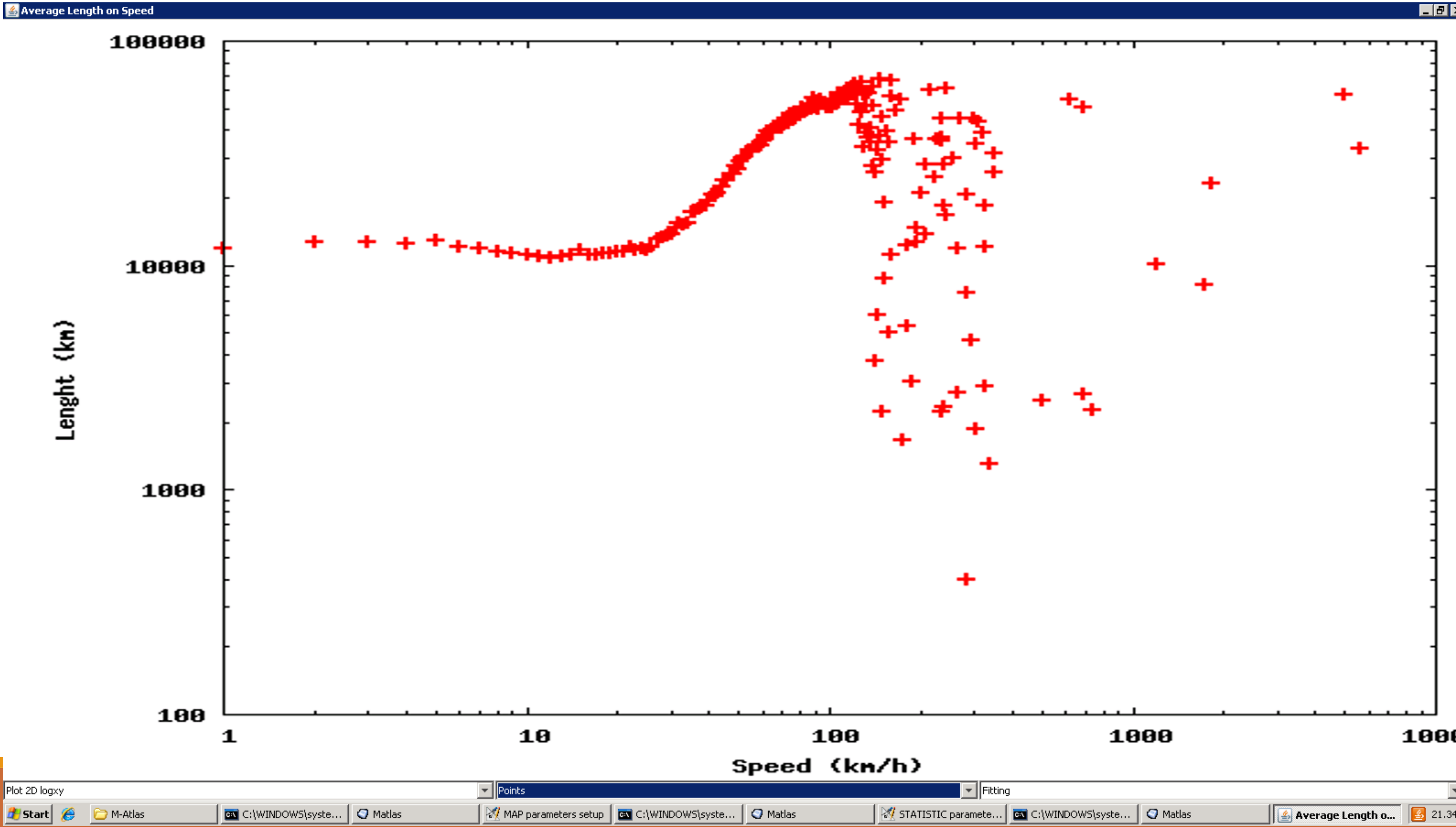
Distribution of lengths (Cumulative)



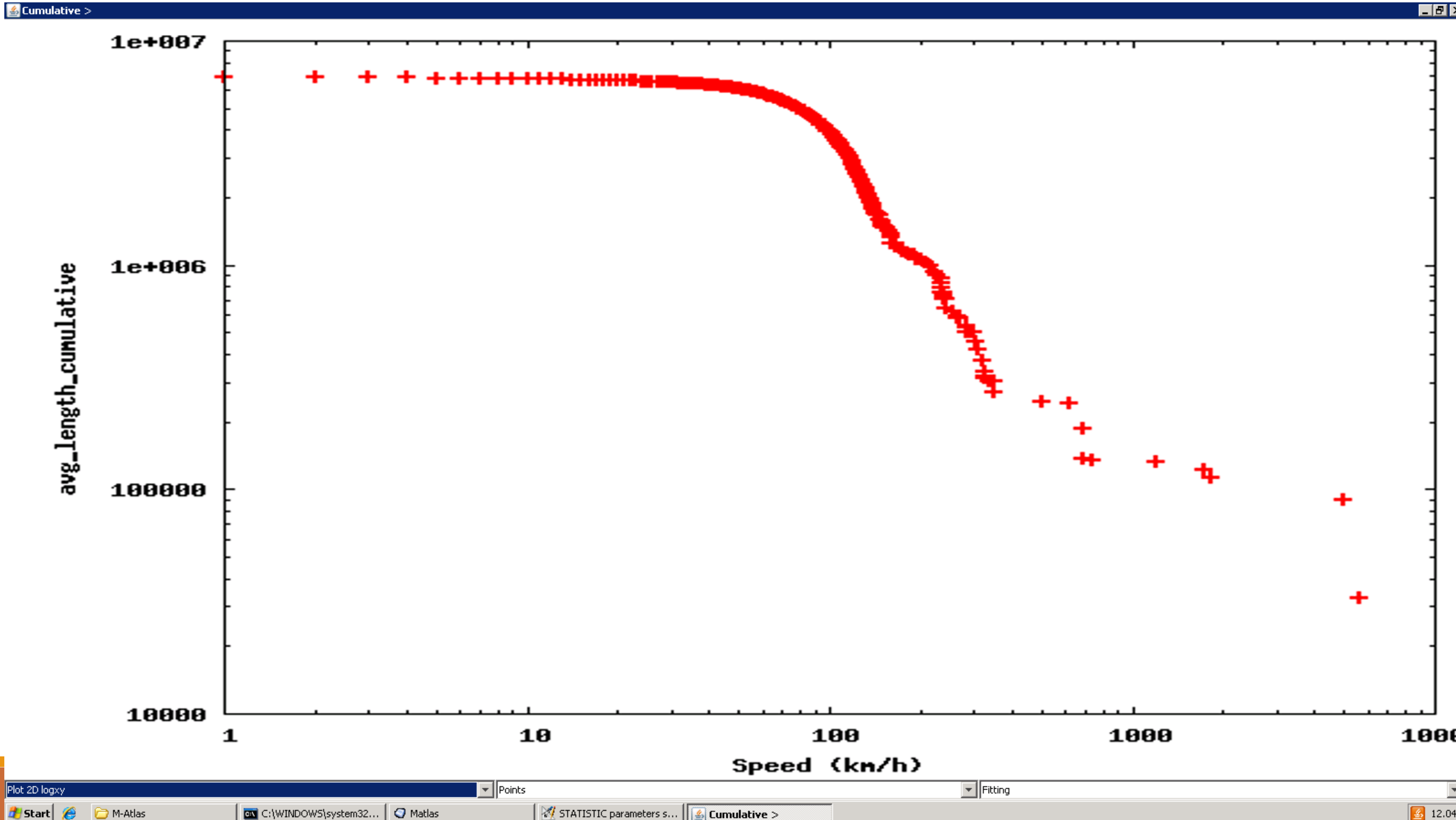
Distribution of Lengths per User (Cumulative)



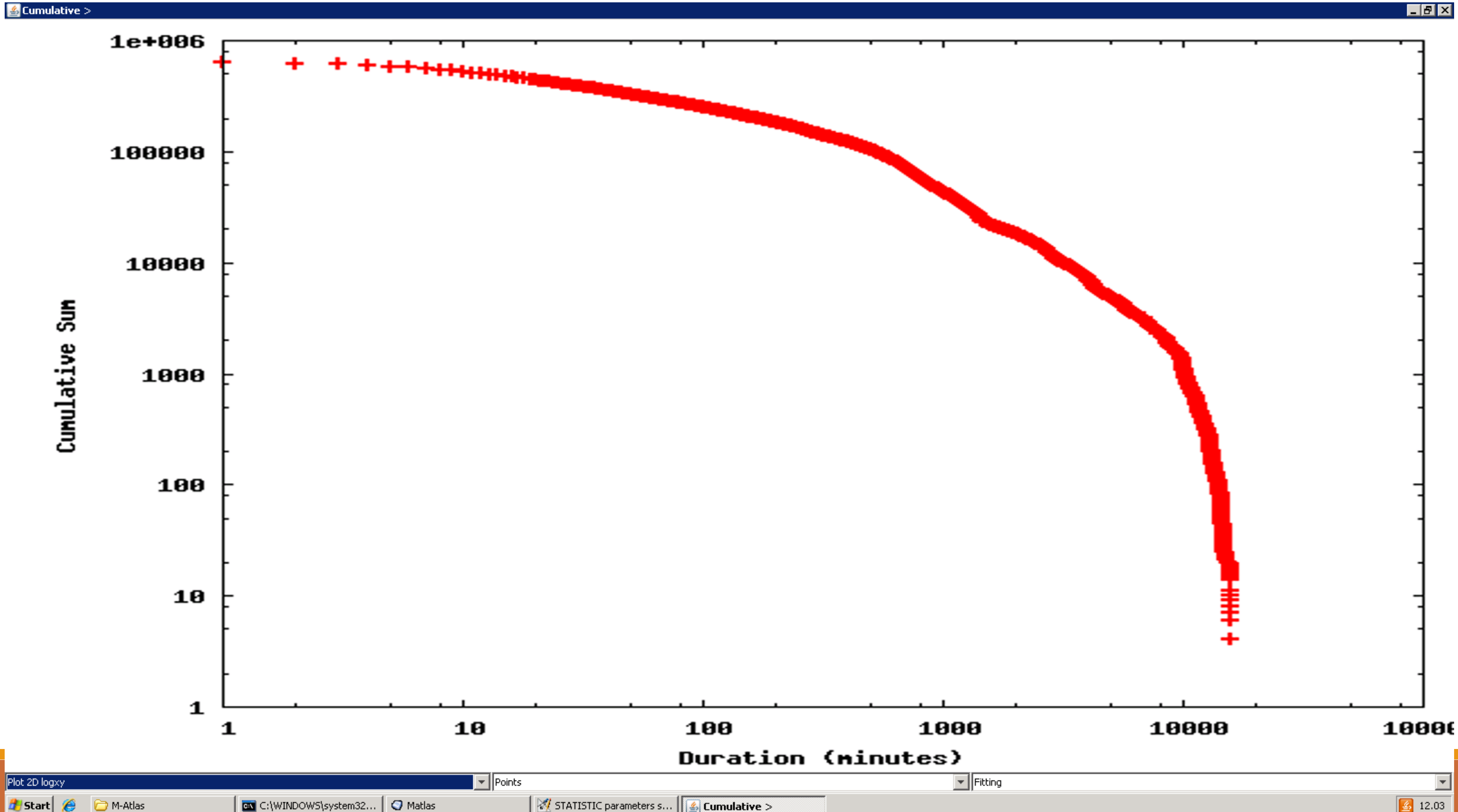
Average length on speed



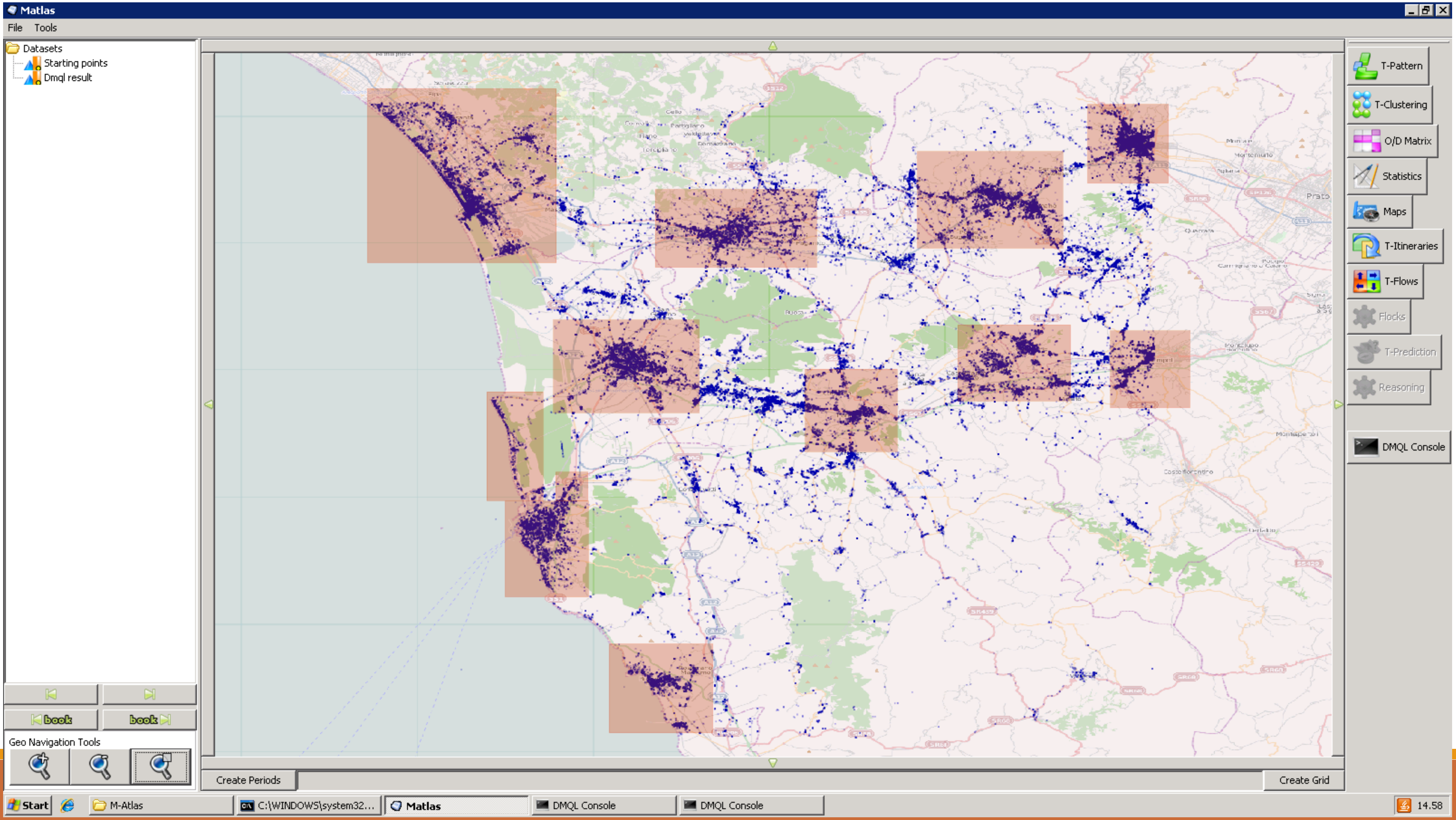
Average length on speed (Cumulative)



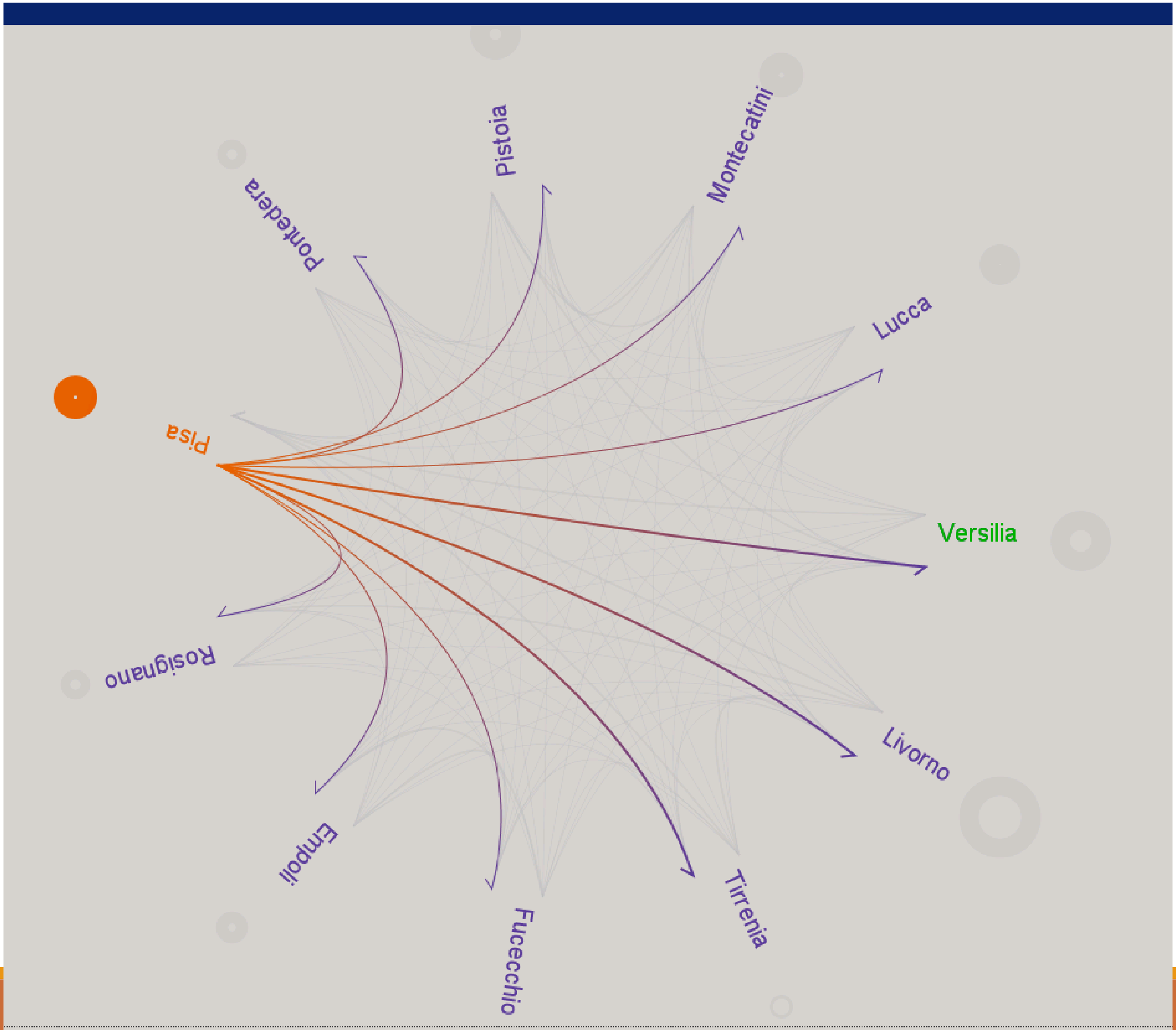
Distribution of Durations (Cumulative)



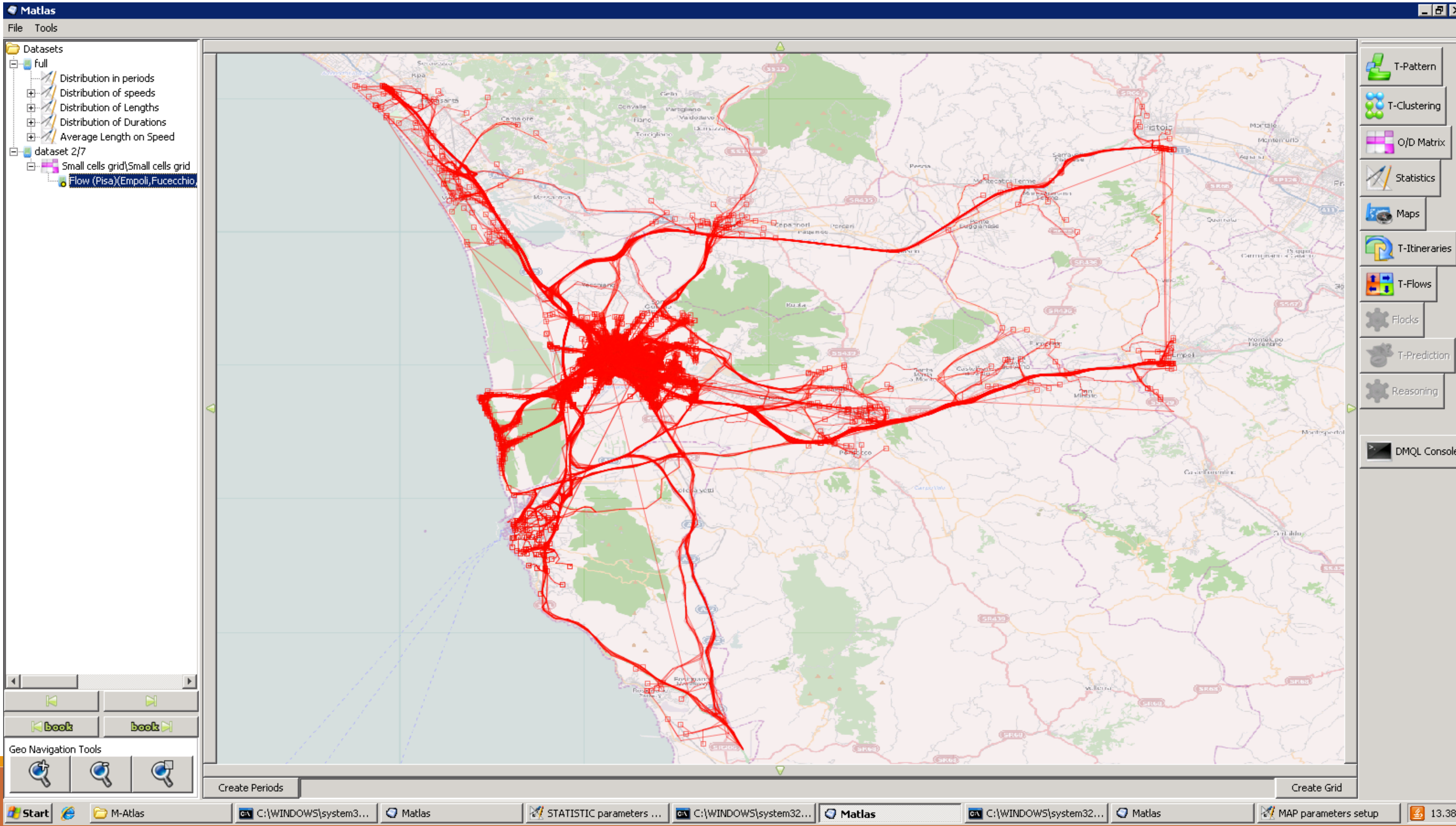
Cities (Approximation)



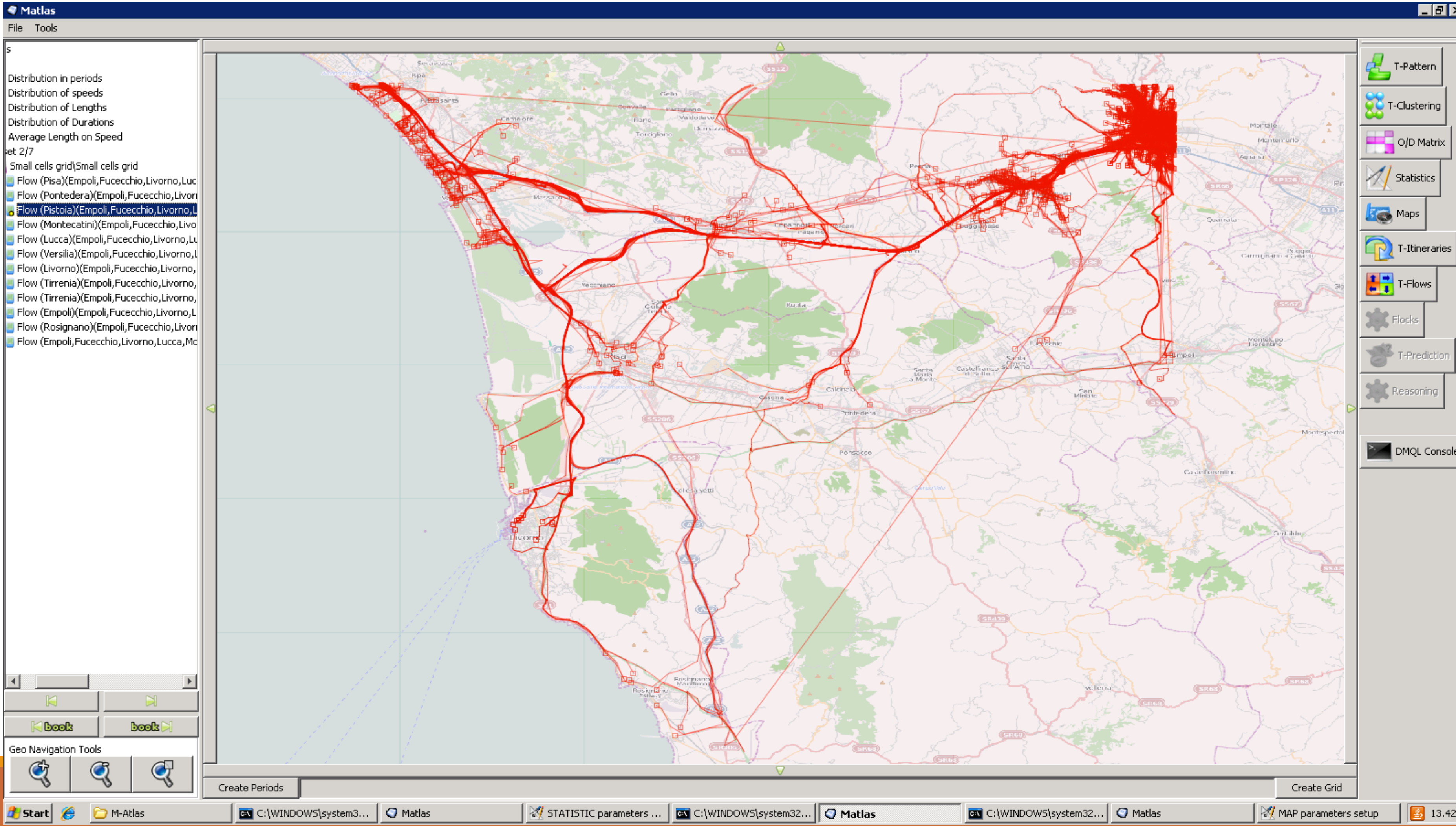
OD Matrix (Cities ←→ Cities)



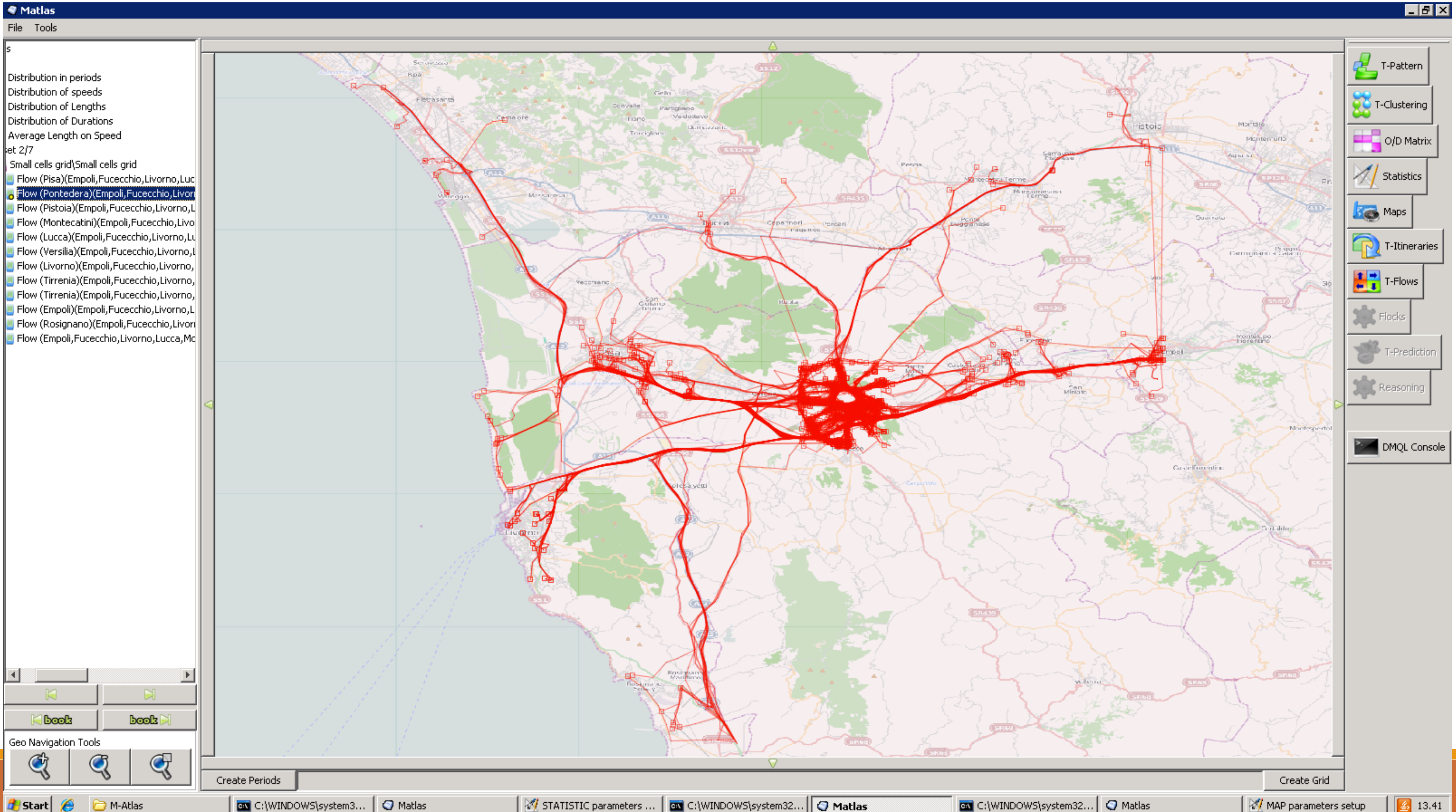
Flow from Pisa



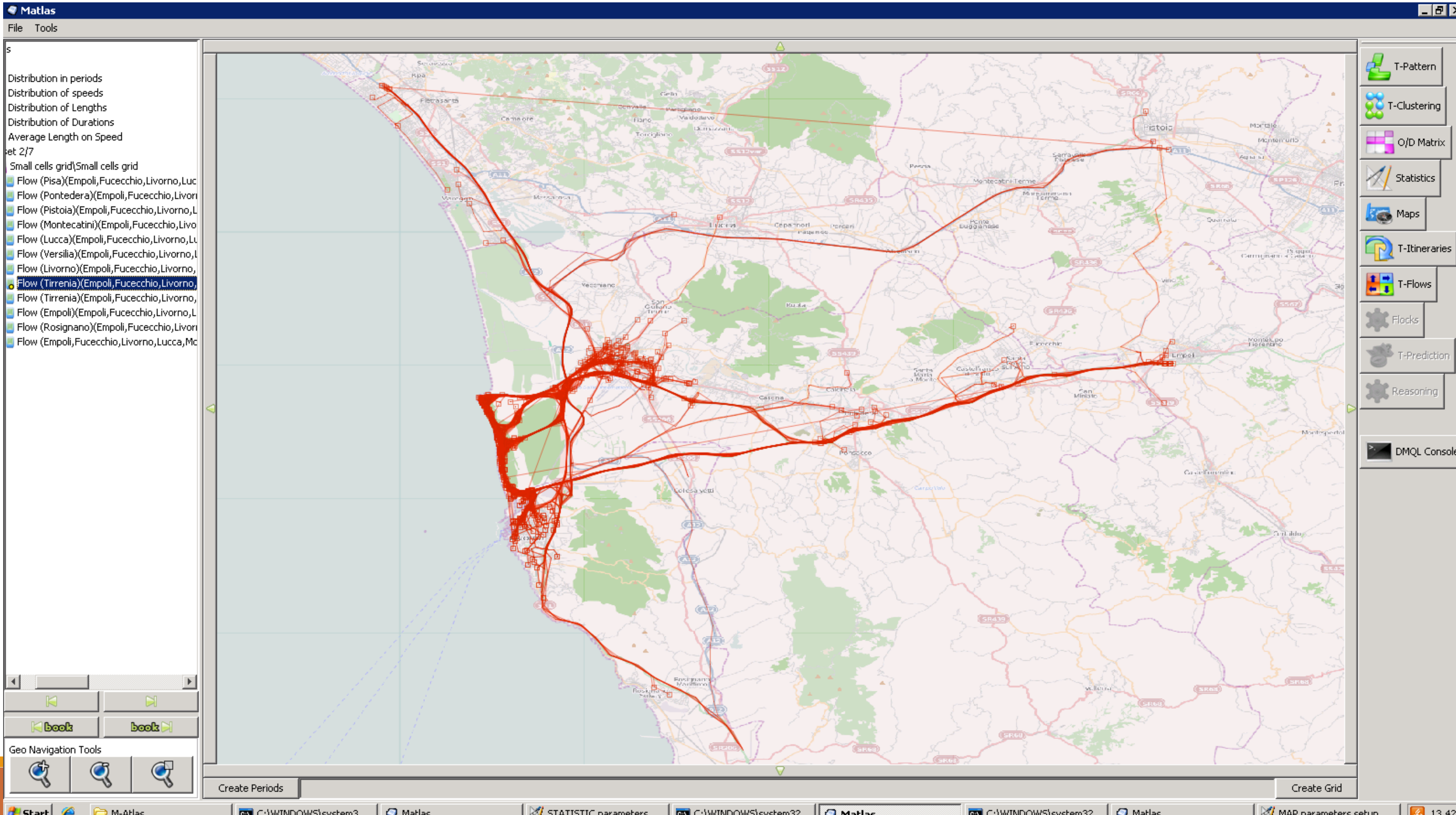
Flow From Pistoia



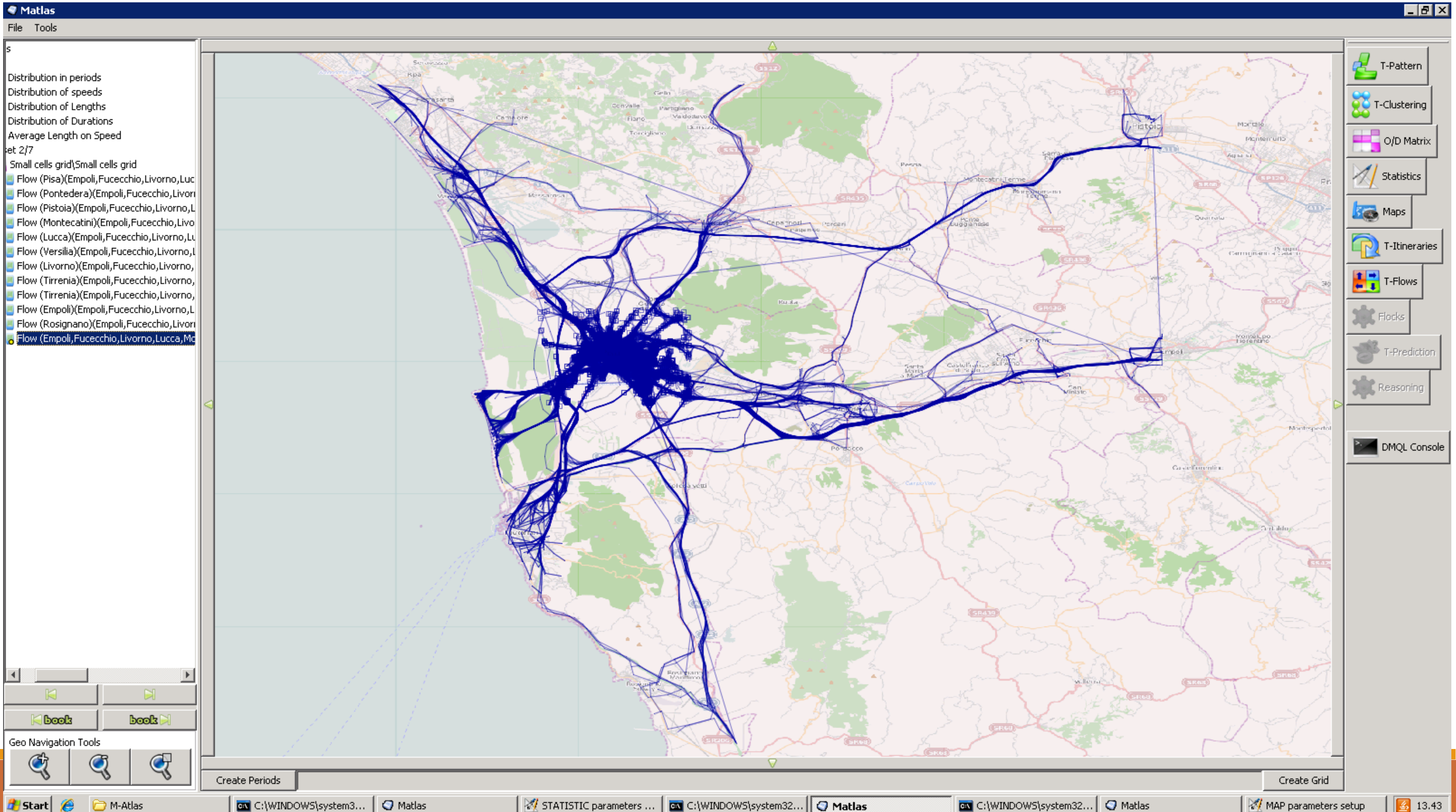
Flow From Pontedera



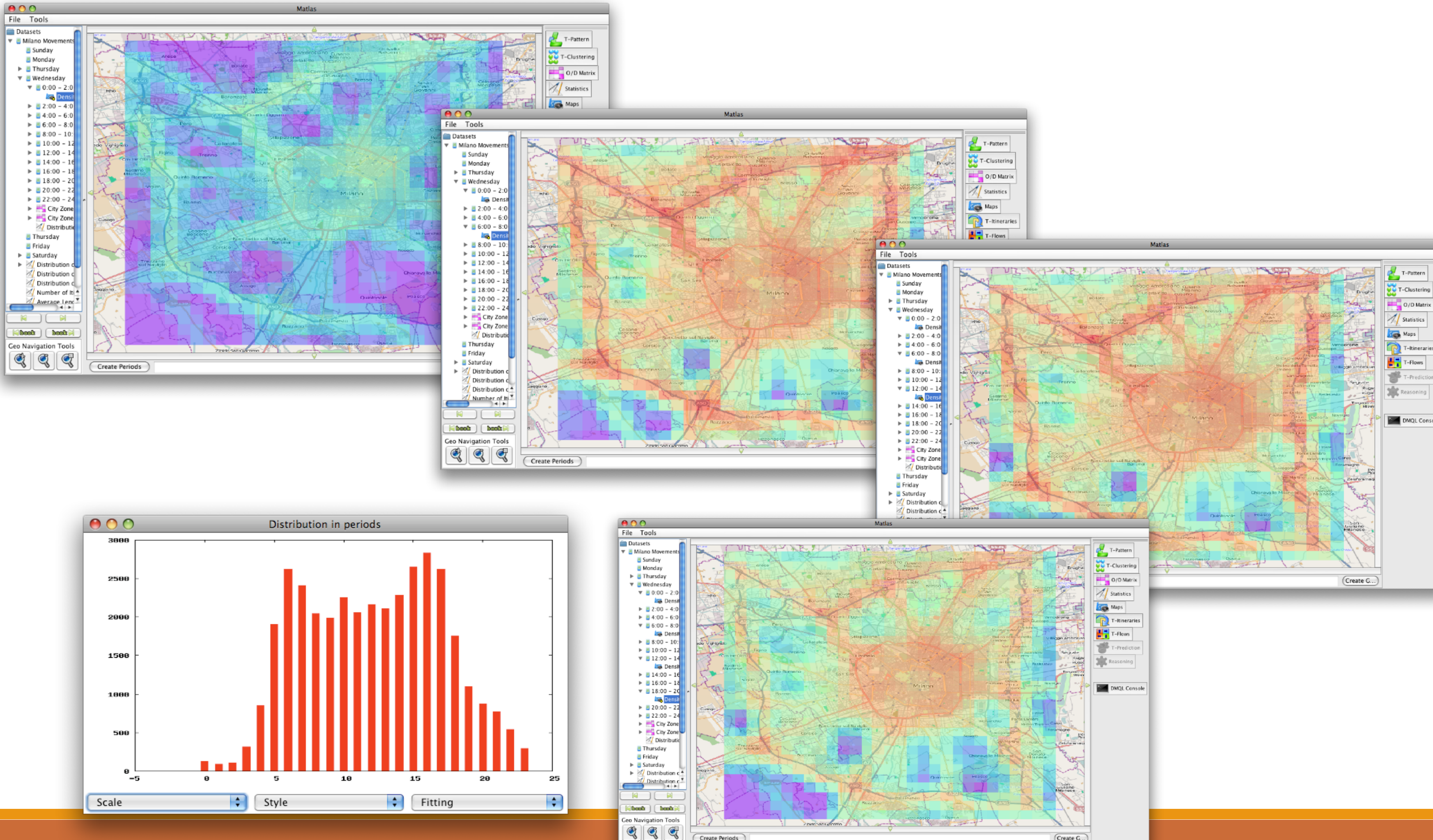
Flow From Tirrenia



Flow To Pisa

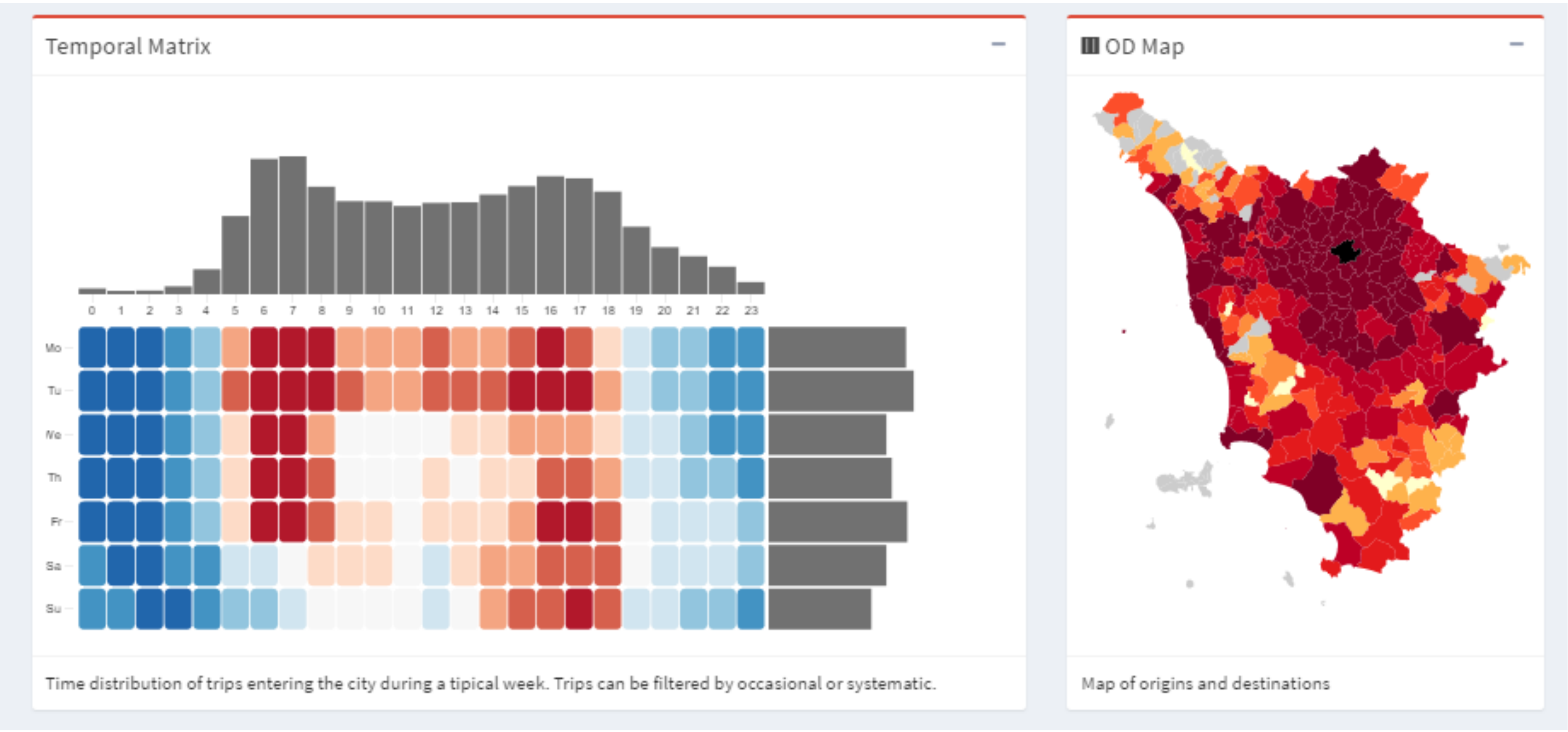

Matlas
File
Tools
Distribution in periods
Distribution of speeds
Distribution of Lengths
Distribution of Durations
Average Length on Speed
Small cells grid\Small cells grid
Flow (Pisa)(Empoli,Fucecchio,Livorno,Luc
Flow (Pontedera)(Empoli,Fucecchio,Livor
Flow (Pistoia)(Empoli,Fucecchio,Livorno,L
Flow (Montecatini)(Empoli,Fucecchio,Livo
Flow (Lucca)(Empoli,Fucecchio,Livorno,Lu
Flow (Versilia)(Empoli,Fucecchio,Livorno,
Flow (Livorno)(Empoli,Fucecchio,Livorno,
Flow (Tirrenia)(Empoli,Fucecchio,Livorno,
Flow (Tirrenia)(Empoli,Fucecchio,Livorno,
Flow (Empoli)(Empoli,Fucecchio,Livorno,L
Flow (Rosignano)(Empoli,Fucecchio,Livor
Flow (Empoli,Fucecchio,Livorno,Lucca,Mc
book
book
Geo Navigation Tools
T-Pattern
T-Clustering
O/D Matrix
Statistics
Maps
T-Itineraries
T-Flows
Flocks
T-Prediction
Reasoning
DMQL Console
Create Periods
Create Grid
Start
M-Atlas
C:\WINDOWS\system3...
Matlas
STATISTIC parameters ...
C:\WINDOWS\system32...
Matlas
C:\WINDOWS\system32...
Matlas
MAP parameters setup
13.43

How do people move during the day?



How do people move during the day?

Urban Mobilty Atlas: http://kdd.isti.cnr.it/uma2/

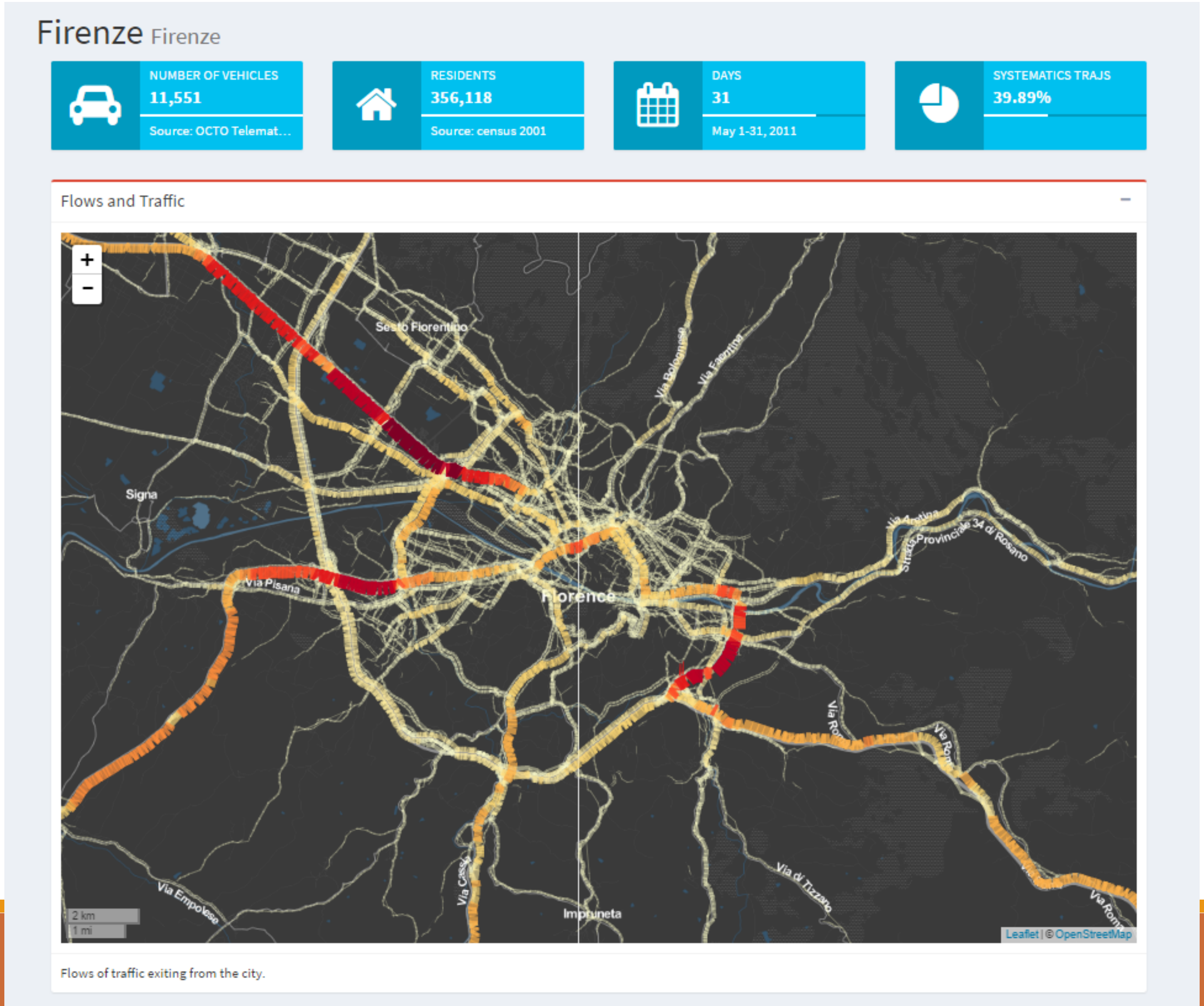


Temporal distribution

Origin-Destination matrix (regional scale)

How do people move during the day?

Urban Mobilty Atlas: http://kdd.isti.cnr.it/uma2/



Traffic on the road

Giannotti · Pedreschi (Eds.)

Mobility, Data Mining and Privacy

The technologies of mobile communications and ubiquitous computing pervade our society, and wireless networks sense the movement of people and vehicles, generating large volumes of mobility data. This is a scenario of great opportunities and risks: on one side, mining this data can produce useful knowledge, supporting sustainable mobility and intelligent transportation systems; on the other side, individual privacy is at risk, as the mobility data contain sensitive personal information. A new multidisciplinary research area is emerging at this crossroads of mobility, data mining, and privacy.

This book assesses this research frontier from a computer science perspective, investigating the various scientific and technological issues, open problems, and roadmap. The editors manage a research project called GeoPKDD, Geographic Privacy-Aware Knowledge Discovery and Delivery, funded by the EU Commission and involving 40 researchers from 7 countries, and this book tightly integrates and relates their findings in 13 chapters covering all related subjects, including the concepts of movement data and knowledge discovery from movement data; privacy-aware geographic knowledge discovery; wireless network and next-generation mobile technologies; trajectory data models, systems and warehouses; privacy and security aspects of technologies and related regulations; querying, mining and reasoning on spatiotemporal data; and visual analytics methods for movement data.

This book will benefit researchers and practitioners in the related areas of computer science, geography, social science, statistics, law, telecommunications and transportation engineering.

ISBN 978-3-540-75176-2

9 783540 751762

springer.com

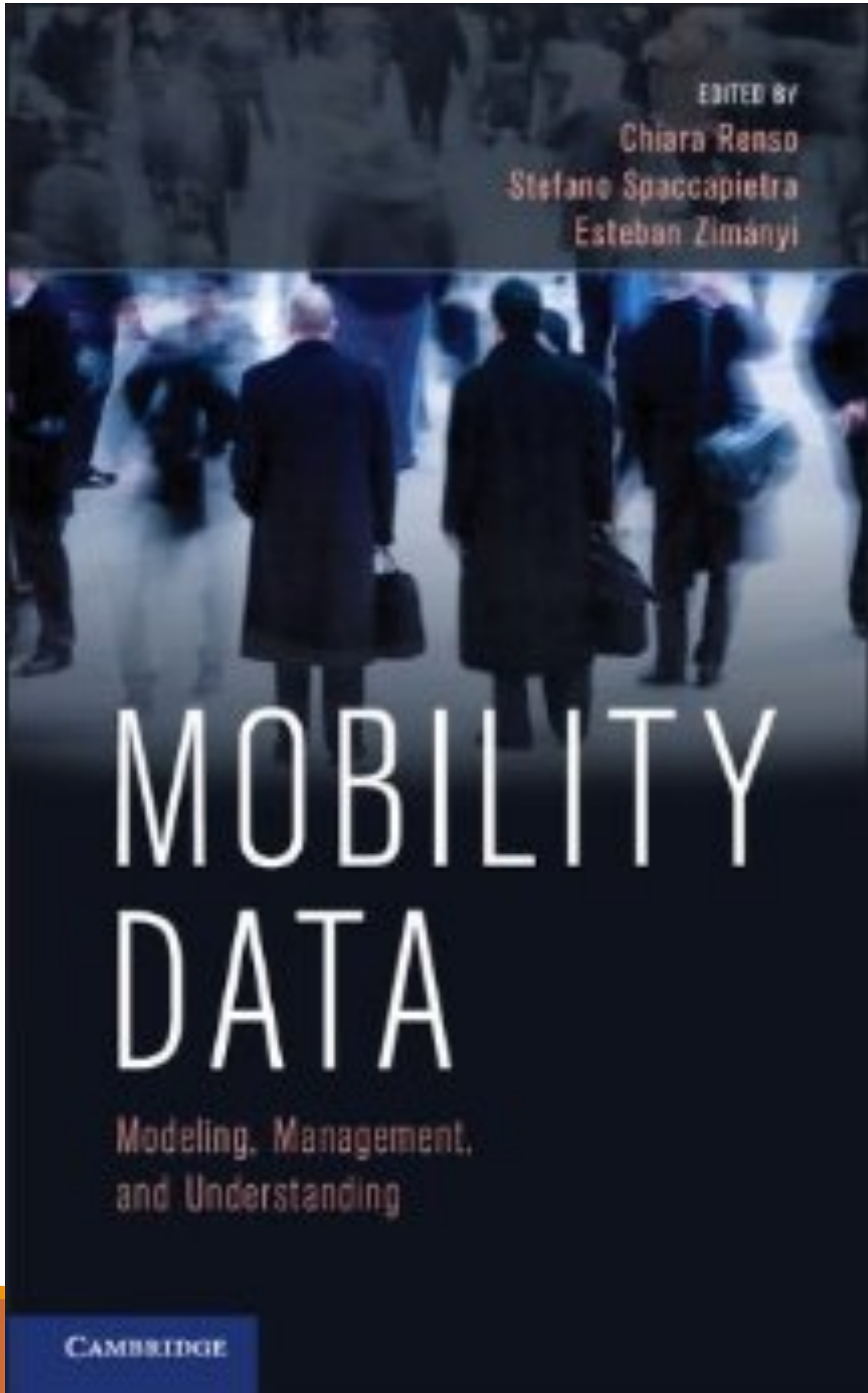
Giannotti
Pedreschi (Eds.)

Mobility, Data Mining
and Privacy

Mobility, Data Mining and Privacy

Geographic Knowledge Discovery

Springer


EDITED BY
Chiara Renso
Stefano Spaccapietra
Esteban Zimányi
MOBILITY
DATA
Modeling, Management,
and Understanding
CAMBRIDGE