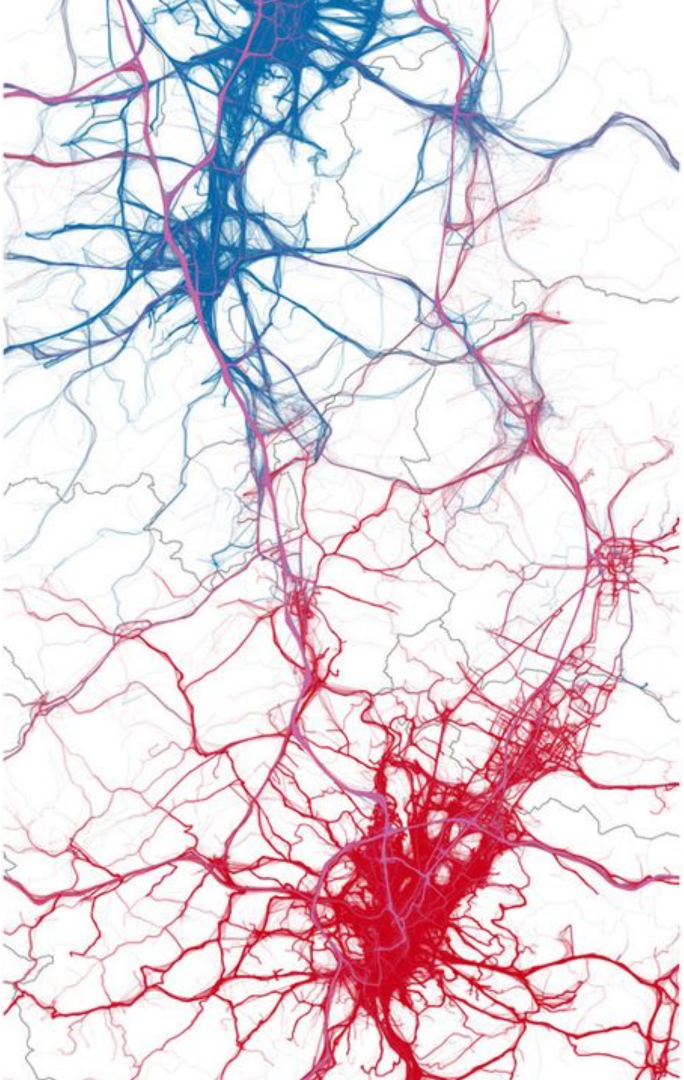


Preprocessing Mobility Data

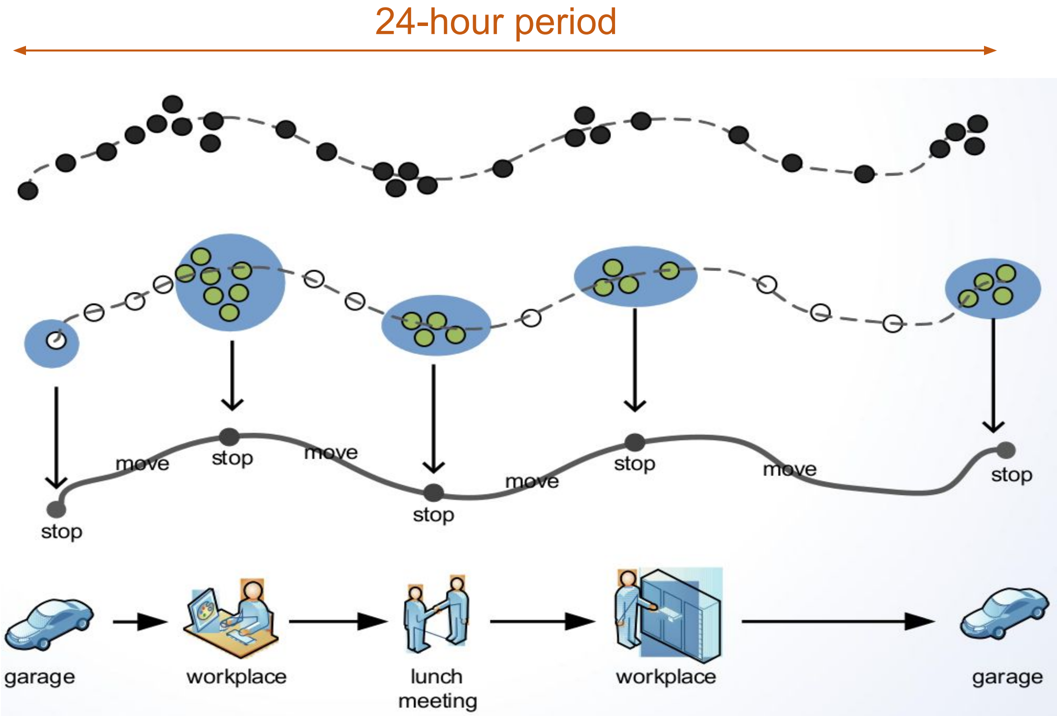
Consiglio Nazionale delle Ricerche

Content of this lesson

- Preprocessing trajectories – Part II
    - Semantic enrichment
        - stop detection / trajectory segmentation
        - home location detection (GPS & MobPhones)
        - activity recognition (POI-based)

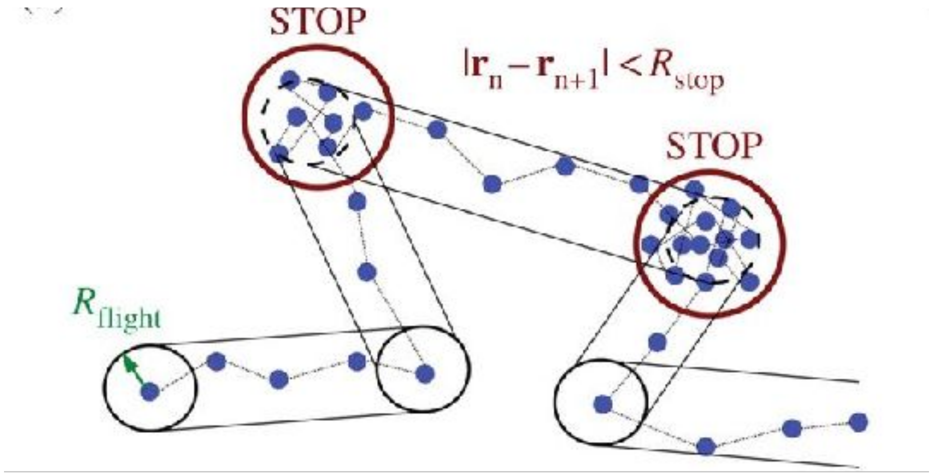
Stop detection & Trajectory segmentation

- Raw data forms a continuous stream of points
- Typical unit of analysis: the trip
- How to segment?
  - Basic idea: identify stops



Stop detection & Trajectory segmentation

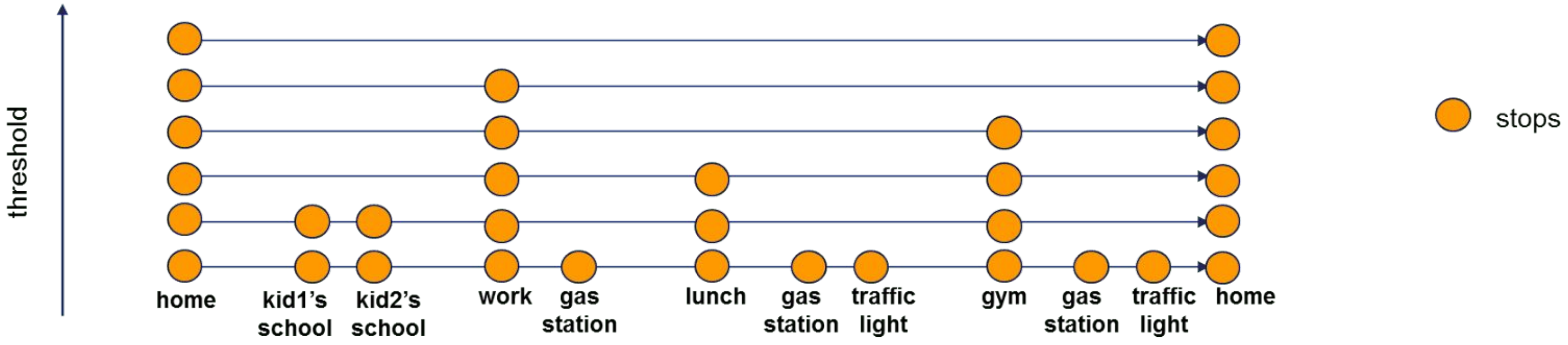
- General criteria based on speed
  - If it **moves very little** (threshold Th_S) over a significant **time interval** (threshold Th_T)
    => it is practically a stop
  - Trajectory (trip) = contiguous sequence of points between two stops



- Typical values:
  - Th_S within [50, 250] meters
  - Th_T within [1, 20] minutes

Stop detection & Trajectory segmentation

- Different time thresholds yield different semantics

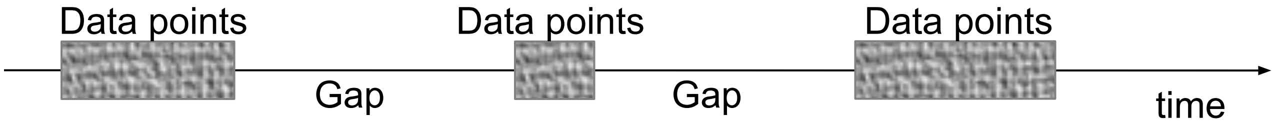


- Which one is the best for you?
  - Application dependent

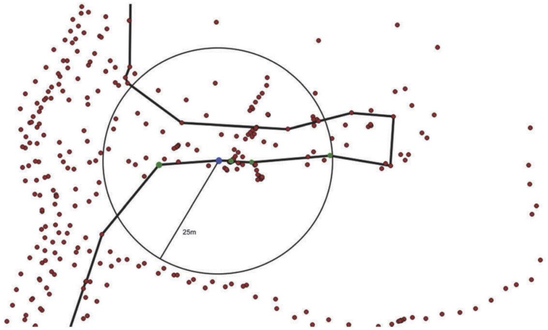
Stop detection & Trajectory segmentation

- 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

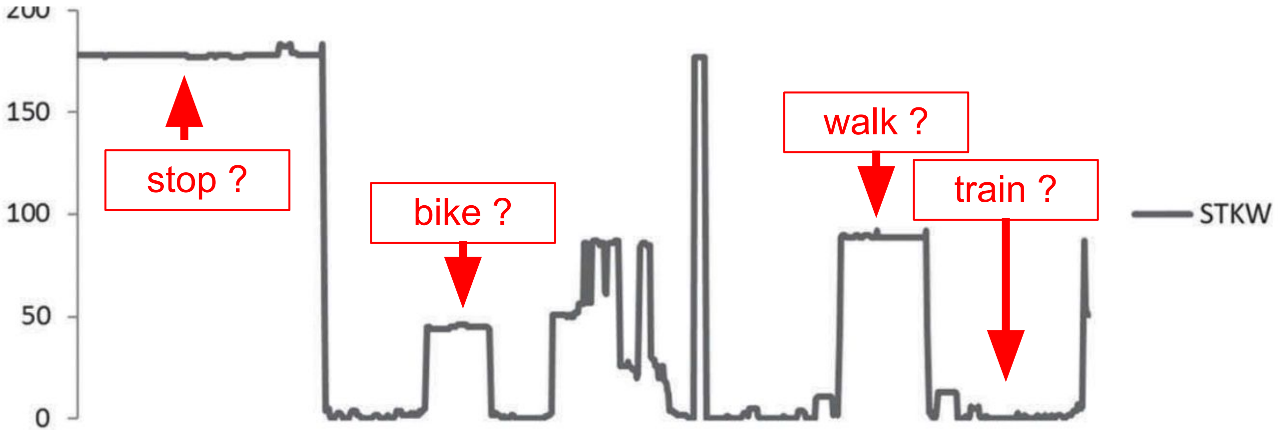


Generalization: transportation means segmentation



- Speed / density-based approach
- Idea: faster means less of my points around me

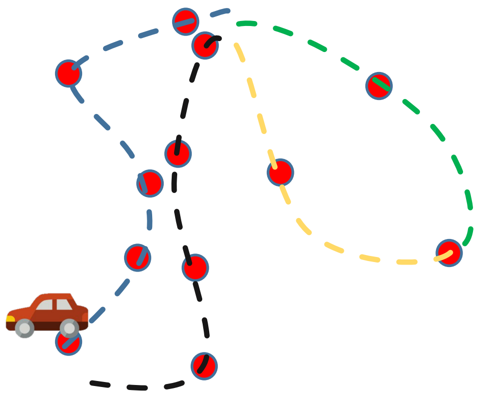
Number of points within radius R



Katarzyna Siła-Nowicka, Jan Vandrol, Taylor Oshan, Jed A. Long, Urška Demšar & A. Stewart Fotheringham (2016) Analysis of human mobility patterns from GPS trajectories and contextual information, International Journal of Geographical Information Science, 30:5, 881-906

User's Mobility History

- What do we get after segmentation?
- Several trajectories associated to the same subject
- Enables individual-level analyses
  - E.g. explore user's habits, find deviations from usual, etc.



Inferring Home / Work locations

- Take all trips of a vehicle / user
- Build a “Individual Mobility Network”
    - Graph abstraction of the overall mobility based on locations (nodes) and movements (edges).

Individual Mobility Network

- Focus on start and stop points
  - Dense areas represent important places

Individual Mobility Network

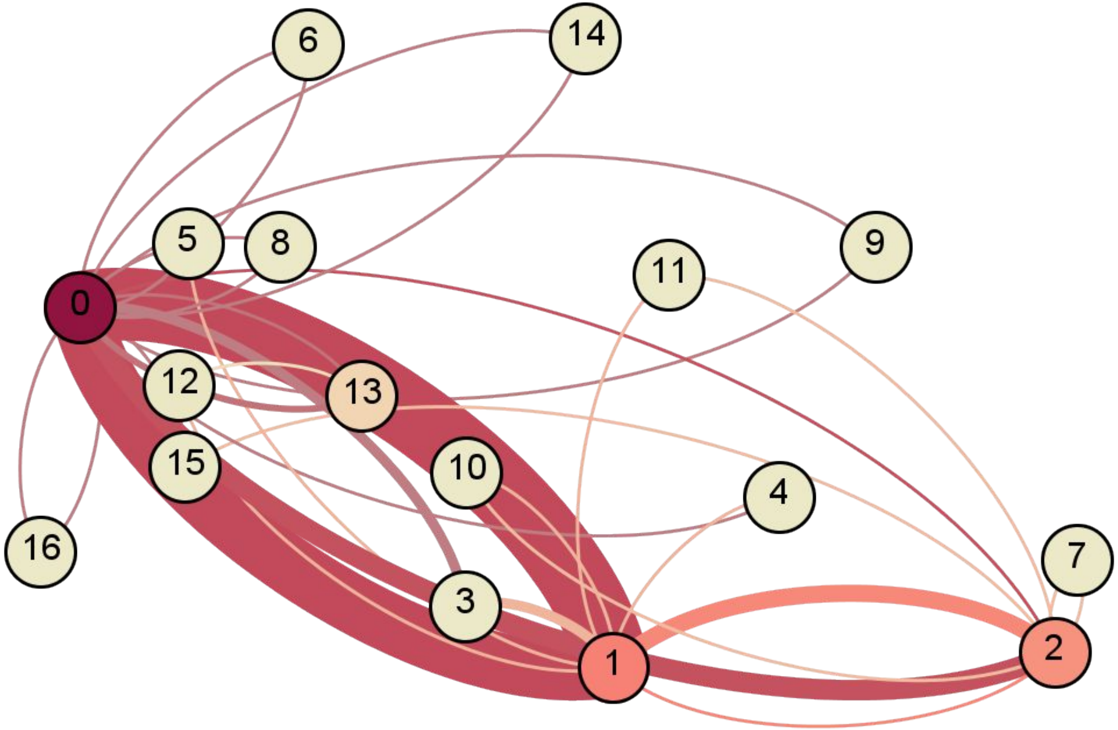
- Cluster points to identify locations

Individual Mobility Network

- Each location is characterized by its frequency

Individual Mobility Network

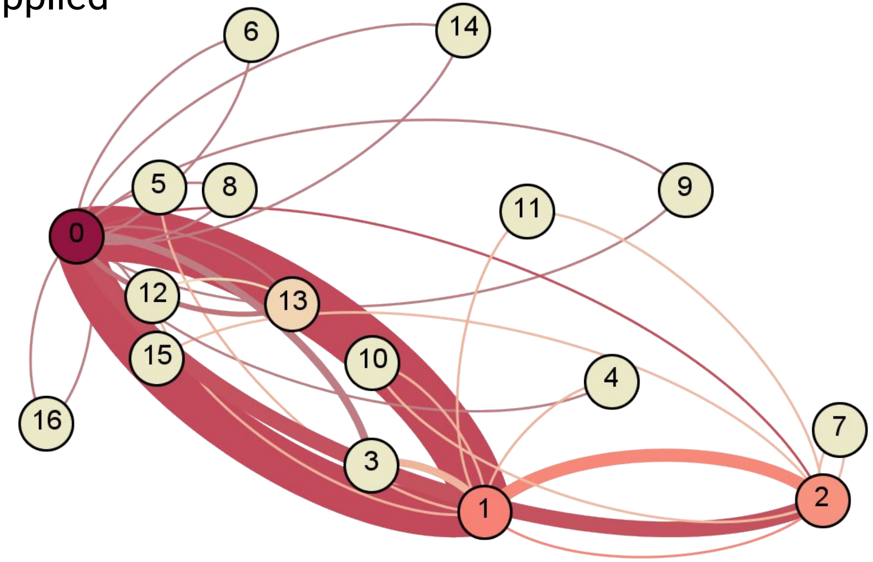
- Trips between points area aggregated as edges between nodes/locations



Inferring Home / Work locations

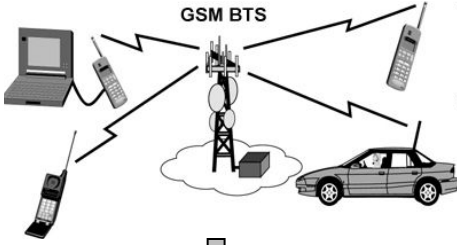
- Basic approach is based on frequency only
  - Most frequent location (L0) := Home
  - Second most frequent location (L1) := Work
    - A minimum frequency threshold is applied

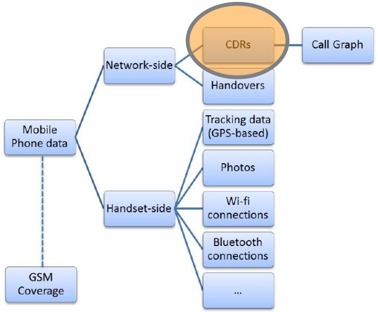
- Various alternatives & refinement are possible
  - Check time of stop & stay duration
    - Home: stop at 20-22, stay 8-10 hrs
    - Work: stop at 7-10, stay 6-9 hrs



Inferring Home / Work locations with Phone Data
Tha case of GSM traces

Data gathered from mobile phone operator for billing purpose





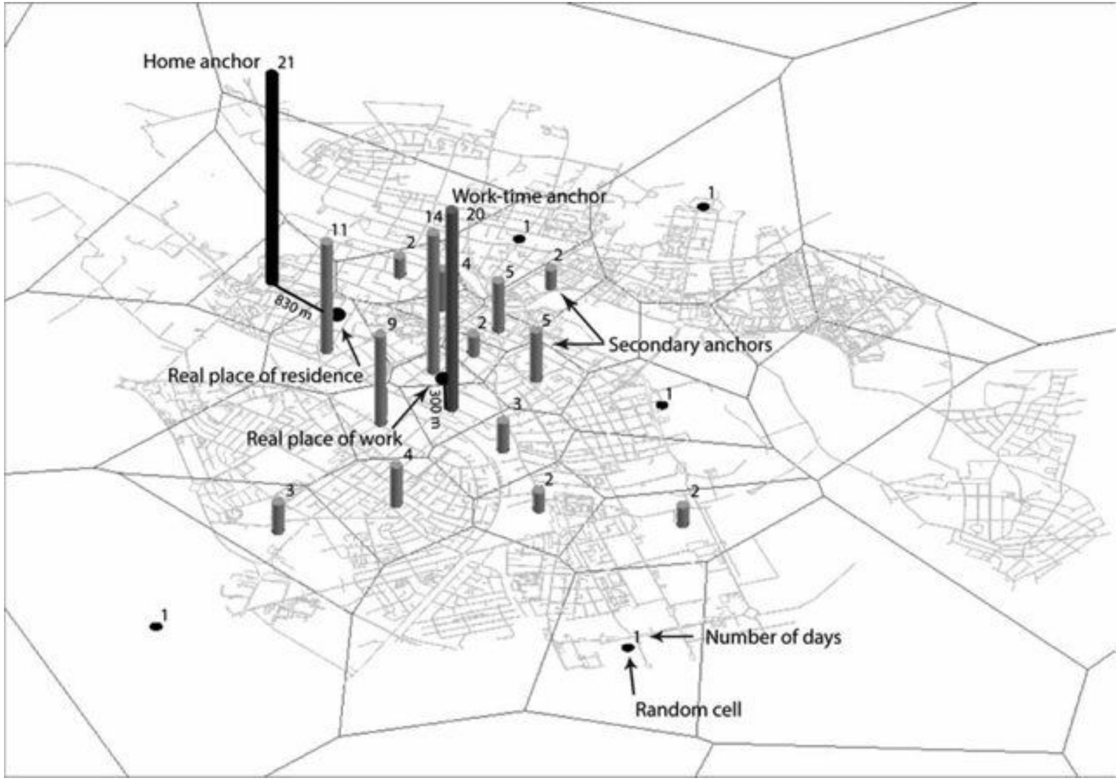
User id	Time start	Cell start	Cell end	Duration
10294595	"2014-02-20 14:24:58"	"PI010U2"	"PI010U1"	48
10294595	"2014-02-20 18:50:22"	"PI002G1"	"PI010U2"	78
10294595	"2014-02-21 09:19:51"	"PI080G1"	"PI016G1"	357

Inferring Home / Work locations with Phone Data
Tha case of GSM traces - 1

- "**Personal Anchor Points**": high-frequency visited places of a user
  - Select top 2 cells with max number of days with calls
  - Determine home and work through time constraints:
    - Based on average start time (AST) of calls and its deviation (std)
    - IF AST<17:00 & std<0.175 ⇒ WORK
    - ELSE HOME

Inferring Home / Work locations with Phone Data
Tha case of GSM traces - 1

- “Personal Anchor Points”



AHAS, R., SILM, S., JARV, O., SALUVEER, E., AND TIRU, M. 2010. *Using mobile positioning data to model locations meaningful to users of mobile phones*. Journal of Urban Technology 17, 1, 3–27.

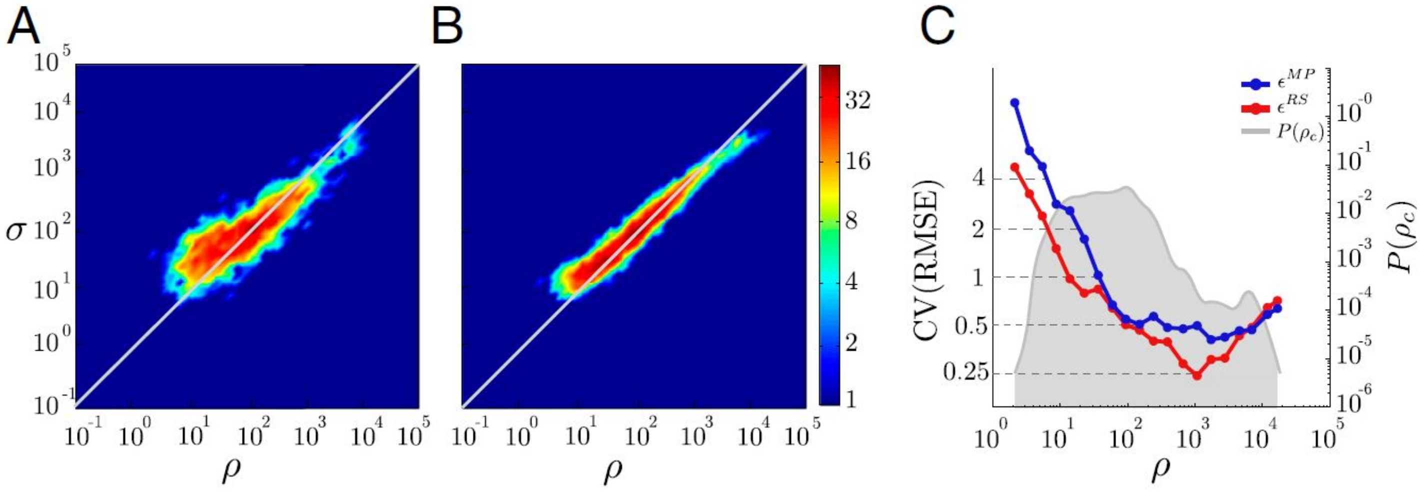
Inferring Home / Work locations with Phone Data
Tha case of GSM traces - 2

- Estimating users' **residence through night activity**
  - Home = region with highest frequency of calls during nighttime
  - More suitable for larger scales
    - E.g. region = municipality

Pierre Deville et al.
Dynamic population mapping using mobile phone data.
PNAS vol. 111 no. 45, pp. 15888–15893, doi: 10.1073/pnas.1408439111

Inferring Home / Work locations with Phone Data
Tha case of GSM traces - 2

- Sample results on national level (France)

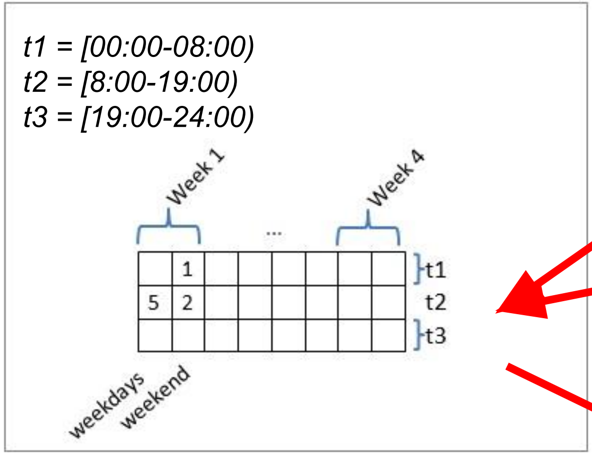


A = GSM data   B = Environment/Infrastructures-based

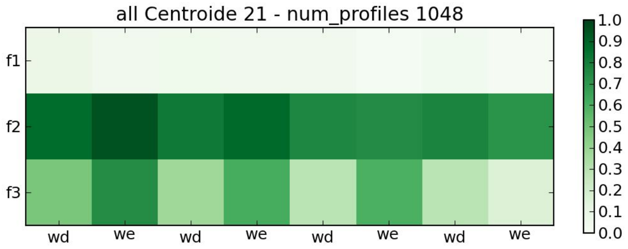
Inferring Home / Work locations with Phone Data
Tha case of GSM traces - Sociometer

Step 1: build individual profiles

- Derive presence distribution for each < user, municipality >



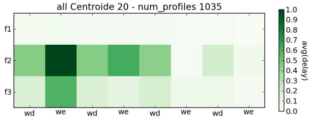
123643 Cell12 24/06/2012 14:05
123643 Cell12 24/06/2012 18:13
123643 Cell15 25/06/2012 11:05
123643 Cell15 25/06/2012 20:42
123643 Cell11 25/06/2012 21:05
123643 Cell12 26/06/2012 10:01
....

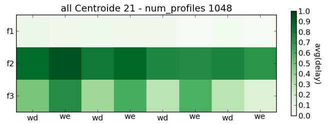


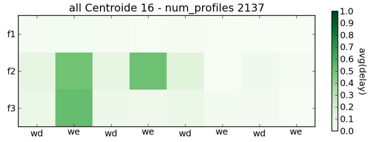
Inferring Home / Work locations with Phone Data
Tha case of GSM traces - Sociometer

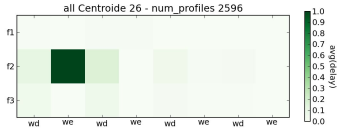
Step 2: find representative profiles across all dataset

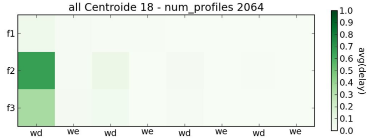
- Based on clustering
  - simple k-means: start with K random representatives, and iteratively refine them
  - in our experiments, k=100
- Output: set of reference (unlabelled) profiles







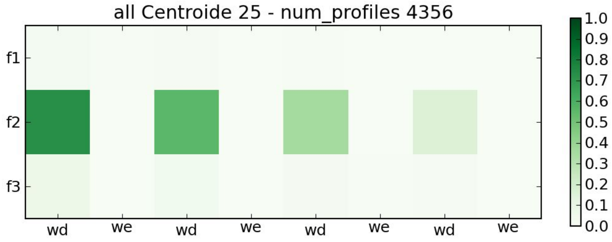




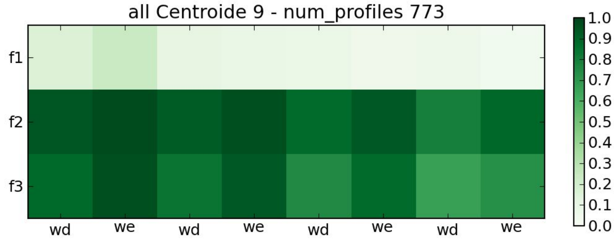
Inferring Home / Work locations with Phone Data
Tha case of GSM traces - Sociometer

Step 3: associate representative profiles to categories

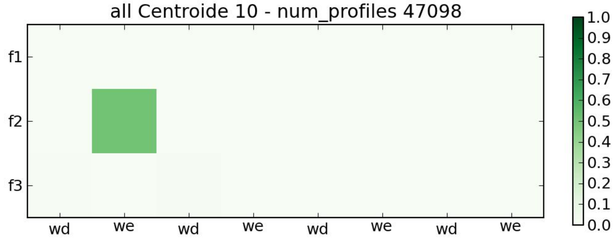
- Manual labelling
  - Use fuzzy rules, difficult to formalize
  - Crisp classification, no weights (reliability of labels)



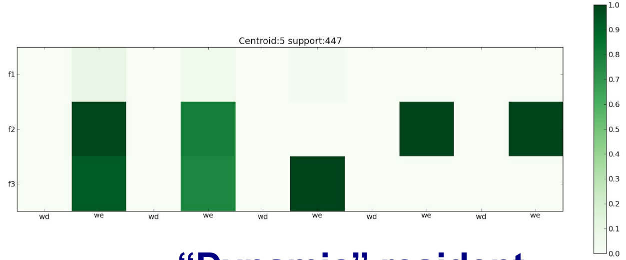
Commuter



"Static" resident



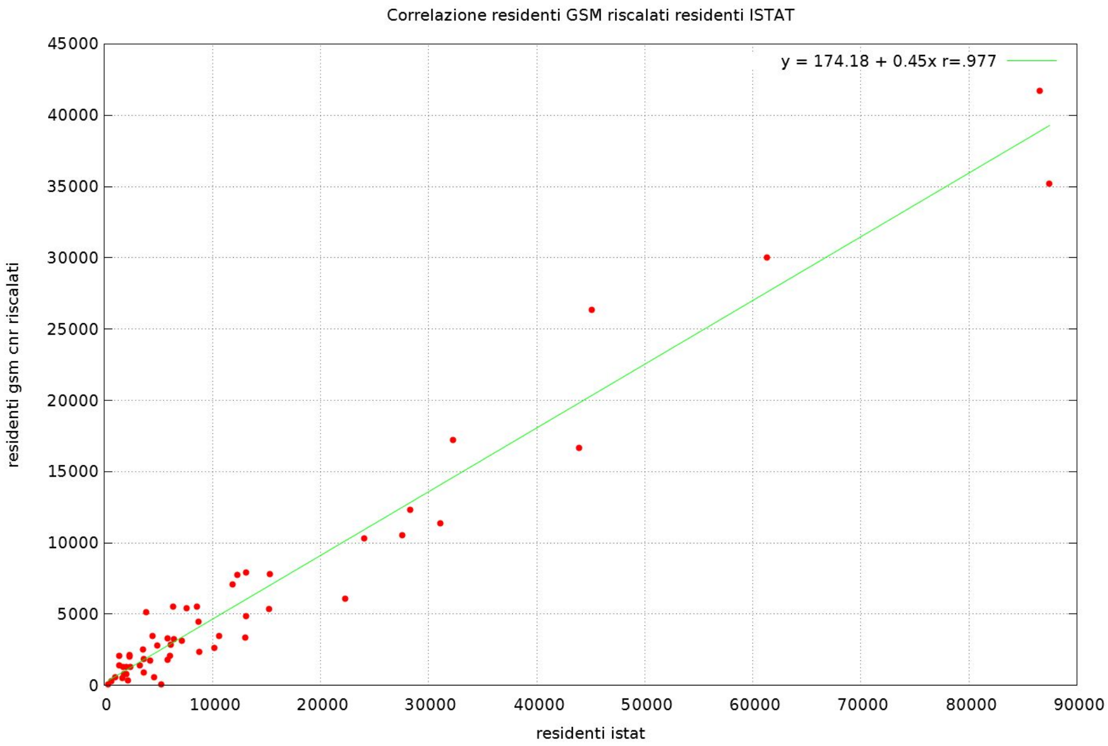
Occasional



"Dynamic" resident

Inferring Home / Work locations with Phone Data
Tha case of GSM traces - Sociometer

Comparing Static residents



Activity labelling / recognition

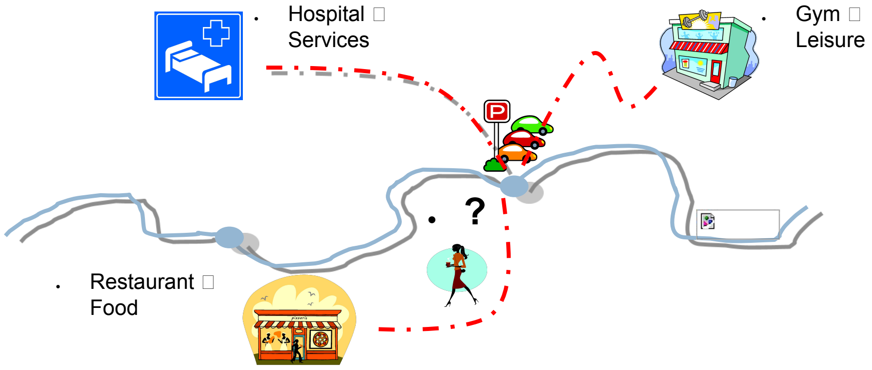
Objective: adding information to points / locations

Two main ways:

- Assign a single activity
- Assign a distribution of POIs / activity types

Activity labelling / recognition

Given a dataset of GPS tracks of private vehicles, annotate trajectories with the most probable activities performed by the user.

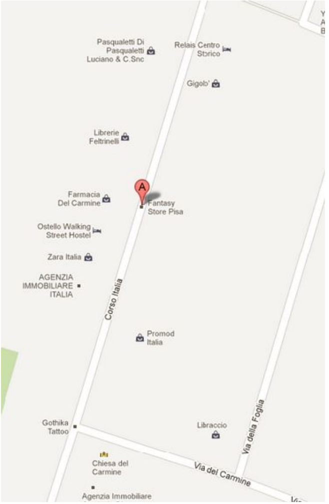


Associates the list of possible POIs (with corresponding probabilities) visited by a user moving by car when he stops.

A mapping between POIs categories and Transportation Engineering activities is necessary.

Activity labelling / recognition

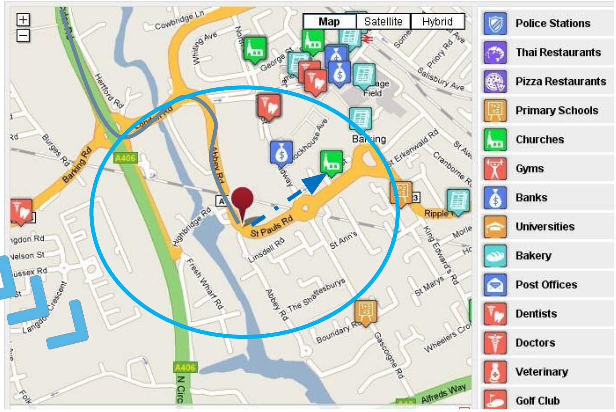
- **POI collection**: Collected in an automatic way, e.g. from Google Places.
- **Association POI – Activity**: Each POI is associated to a ``activity". For example Restaurant → Eating/Food, Library → Education, etc.
- **Basic elements/characteristics:**
  - C(POI) = {category, opening hour, location}
  - C(Trajectory) = {stop duration, stop location, time of the day}
  - C(User) = {max walking distance}
- **Computation of the probability to visit a POI/ to make an activity:** For each POI, the probability of ``being visited" is a function of the POI, the trajectory and the user features.
- **Annotated trajectory**: The list of possible activities is then associated to a Stop based on the corresponding probability of visiting POIs



Activity labelling / recognition

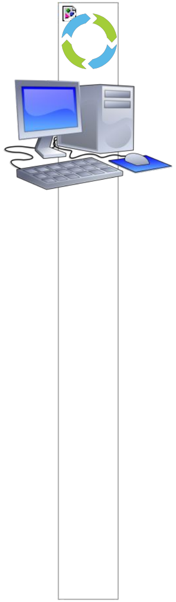
Filter POIs based on time constraints

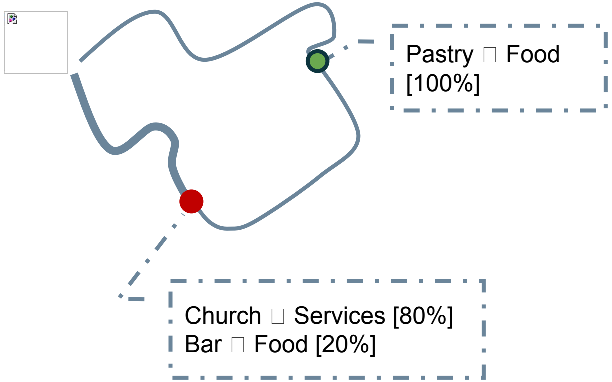
- Lat; Lon
- TimeStamp: Sun 10:55 – 12:05



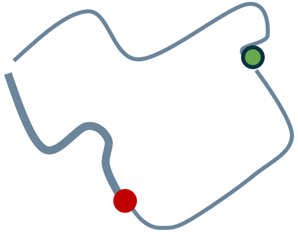
- Bank: Mon – Fri [8:00 – 15:30]
- Dentist: Mon – Sat [9:00 – 13:00] [15:30 – 18:00]
- Church: Mon – Sat [18:00 – 19:00]
- Sun [11:00 – 12:00]
- Primary School: Mon – Sat [8:00 – 13:00]

Activity labelling / recognition

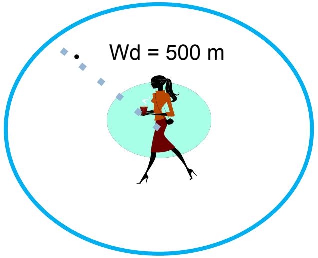




- Stop: Lat; Lon
- Time: Mon 7:00 – 7:15



- Stop: Lat; Lon
- Time: Sun 10:55 – 12:05

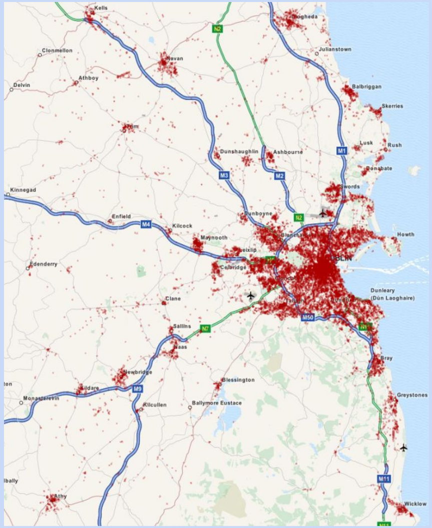


INTERVALLO

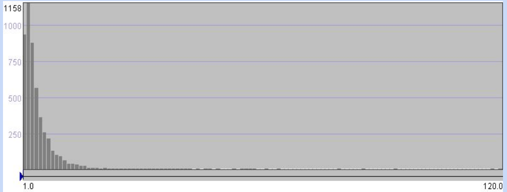
Reading social media to find POIs

An Irish experiment on Twitter

The points of each trajectory taken separately were grouped into spatial clusters of maximal radius 150m. For groups with at least 5 points, convex hulls have been built and spatial buffers of small width (5m) around them.
1,461,582 points belong to the clusters (89% of 1,637,346); 24,935 personal places have been extracted.



Examples of extracted places



Statistical distribution of the number of places per person

INTERVALLO

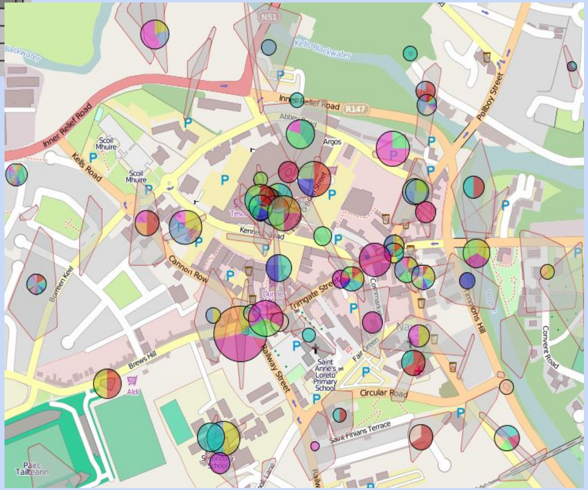
Reading social media to find POIs

Topics have been assigned to 208,391 messages (14.3% of the 1,461,582 points belonging to the personal places)

Message	Features	topic=family: Occurrences of topic	topic=home: Occurrences of topic	topic=education: Occurrences of topic	topic=work: Occurrences of topic
@joe_lennon I usually	education	0	0	1	0
@joe_lennon together	education	0	0	1	0
@jas_103 deadly; don	work	0	0	0	1
Just got home and see	home	0	1	0	0
So excited about my ne	sweets	0	0	0	0
@lamtcdizzy I haven't b	shopping	0	0	0	0
Get in from my night ou	family;home;work	1	1	0	1
Home again at 6pm! N	home	0	1	0	
Bussing it home for th			1	0	
Ah shite. It's been a p			0	0	
@ronanhutchinson be	education	0	0	1	

Get in from my night out; my dad gets home from work two minutes later. Great timing :)

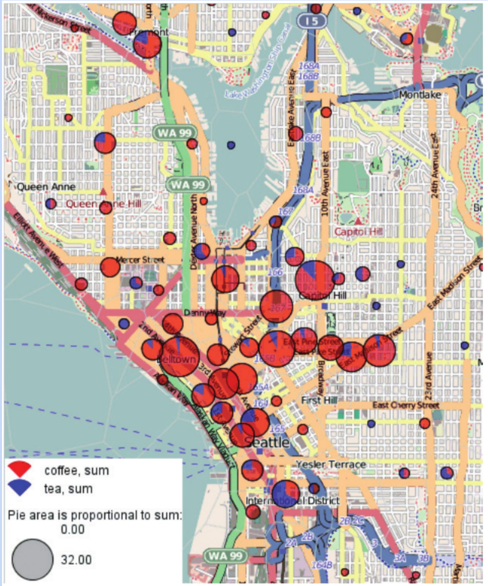
...

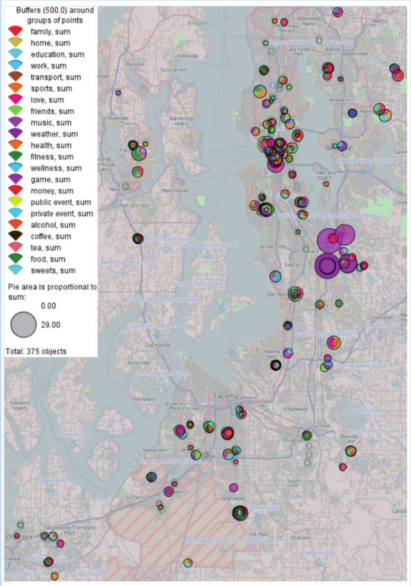


1) Some places did not get topic summaries (about 20% of the places)
2) In many places the topics are very much mixed
3) The topics are not necessarily representative of the place type (e.g., topics near a supermarket: family, education, work, cafe, shopping, services, health care, friends, game, private event, food, sweets, coffee)

INTERVALLO

In the meanwhile, in Seattle…





G. Andrienko et al. Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. Computing in Science & Engineering, 2013.

Homeworks

to be delivered by Friday, October 14th 2022

Homework 4.1

How fast are users?

Choose one of the datasets seen at lesson (taxis, Geolife, etc.), select at least 10 users/vehicles and compute distributions of lengths. Remove 10% of points in each trajectory and repeat the distribution. Do the same for 20%, 30%, … 90%. How does length distribution change?

- Submit a (well commented) python notebook

Homework 4.2

Implement your own time-aware trajectory compression method, and test it on a dataset of your choice, e.g. a subset of taxis or Geolife users.

- Show the effects of simplification on some sample trajectories
- Study how the lengths of trajectories are affected
- Submit a (well commented) python notebook

Homework 4.3

Inferring Home locations is often used to estimate the resident population of geographical areas. What are the existing approaches to face the problem?

- Make a research on Internet on the methods, including big data-based ones (GPS, GSM data, maybe satellite data or others) but also any other approach – e.g. coming from statistics/demography, sociology, etc.
- Prepare a blog (basically a survey) summarizing your discoveries.

Homework 4.4

Estimating GPS errors. Choose a bounding rectangle covering SF city. Download the road network/graph of that area. Select the GPS points of taxis in the same area. Assign each point P to its closest road segment R. Define pseudo-error(P) as the distance dist(P,R).

- Analyze the overall distribution of the pseudo-errors. Is it coherent with GPS.gov estimates of errors?
- Are pseudo-errors the same downtown vs. out of city?
- Submit a (well commented) python notebook