Using smartphone data for travel demand analysis: challenges and opportunities

Michel Bierlaire

Transport and Mobility Laboratory

School of Architecture, Civil & Environmental Engineering

Ecole Polytechnique Fédérale de Lausanne, Switzerland

Outline

- Data
  - Challenges
  - Opportunities
- Models: route choice
  - the chosen route
  - the non chosen routes

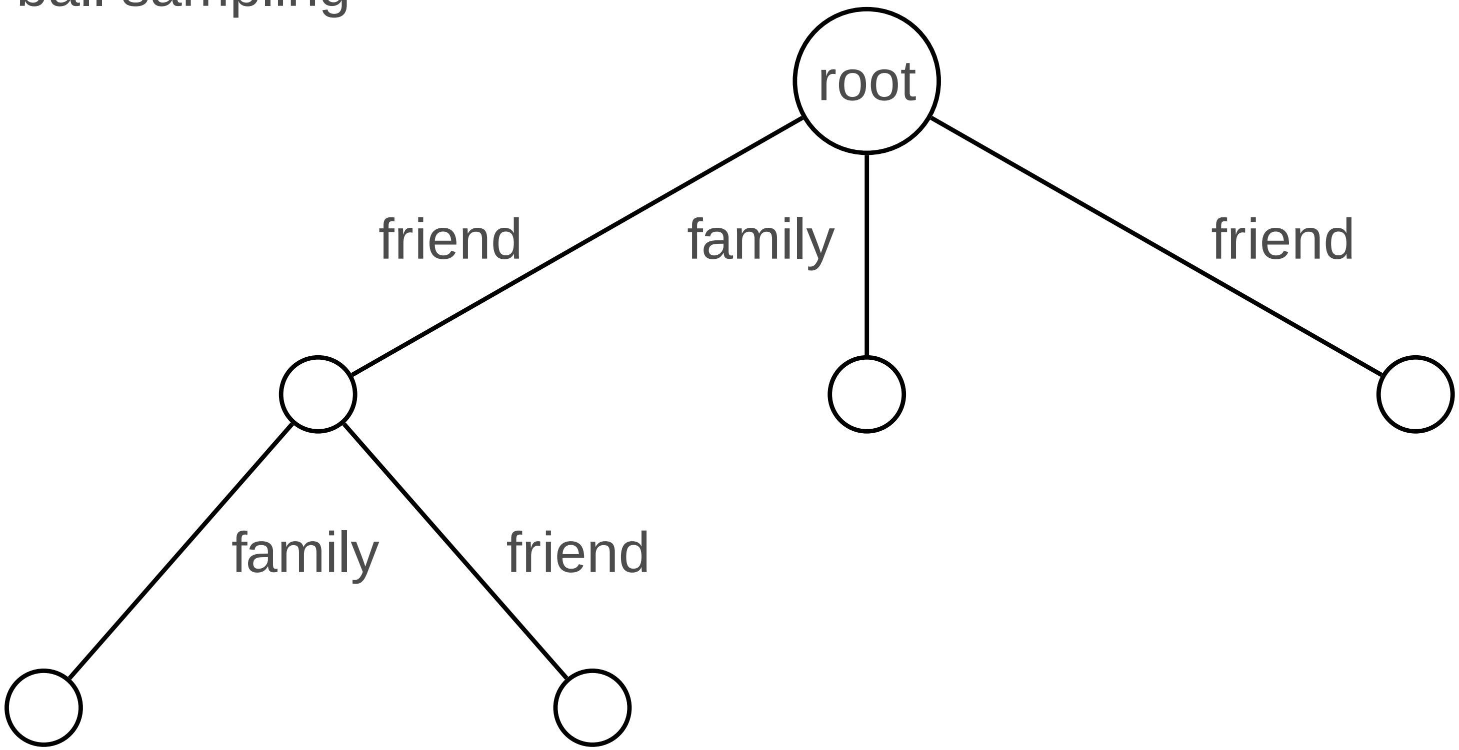
Nokia data collection campaign

Nokia data collection campaign

- Funding source: Nokia Research Center (NRC) at EPFL.
- Participants: About 185.
- Since: September 2009.
- Phone: Nokia N95.
- Collaborators: NRC Lausanne, IDIAP (Switzerland).

Recruitment

snow ball sampling



Participants

- About 185 participants.
- Mostly from Lausanne area.
- $\sim 1/3$ females.
- $< 1/4$ students.

Software design

Phone software (EPFLSCOPE)

- written in python Symbian S60;
- starts with the operating system, runs in backend;
- cannot be turned off by users;
- records data constantly;
- uploads data automatically to DB A via wireless network (WIFI, 3G), every 2 hours.

Databases

- are administrated by Nokia;
- a remote database (DB A) with data access API (httprequest, JSON format);
- another geographical database (DB B) copies data from DB A with $\sim$ 12 hours lag (SQL access).

Energy performance

The original software was developed by Nokia.

- With GPS on, one fully charged battery lasts less than 4 hours.

The energy performance was improved by TRANSP-OR, IDIAP and NRC Lausanne.

- Turn off GPS if stationary.
- Determines stationary/moving: GPS, known WLAN, cell ID, accelerometer.
- One fully charged battery can last ~ 10 hours.

Privacy and security

- Data is owned by participants. They can delete their data from DB A.
- The campaign is permitted and controlled by an ethical committee.
- Nokia and authorized research partners (in CH) get access to the data.

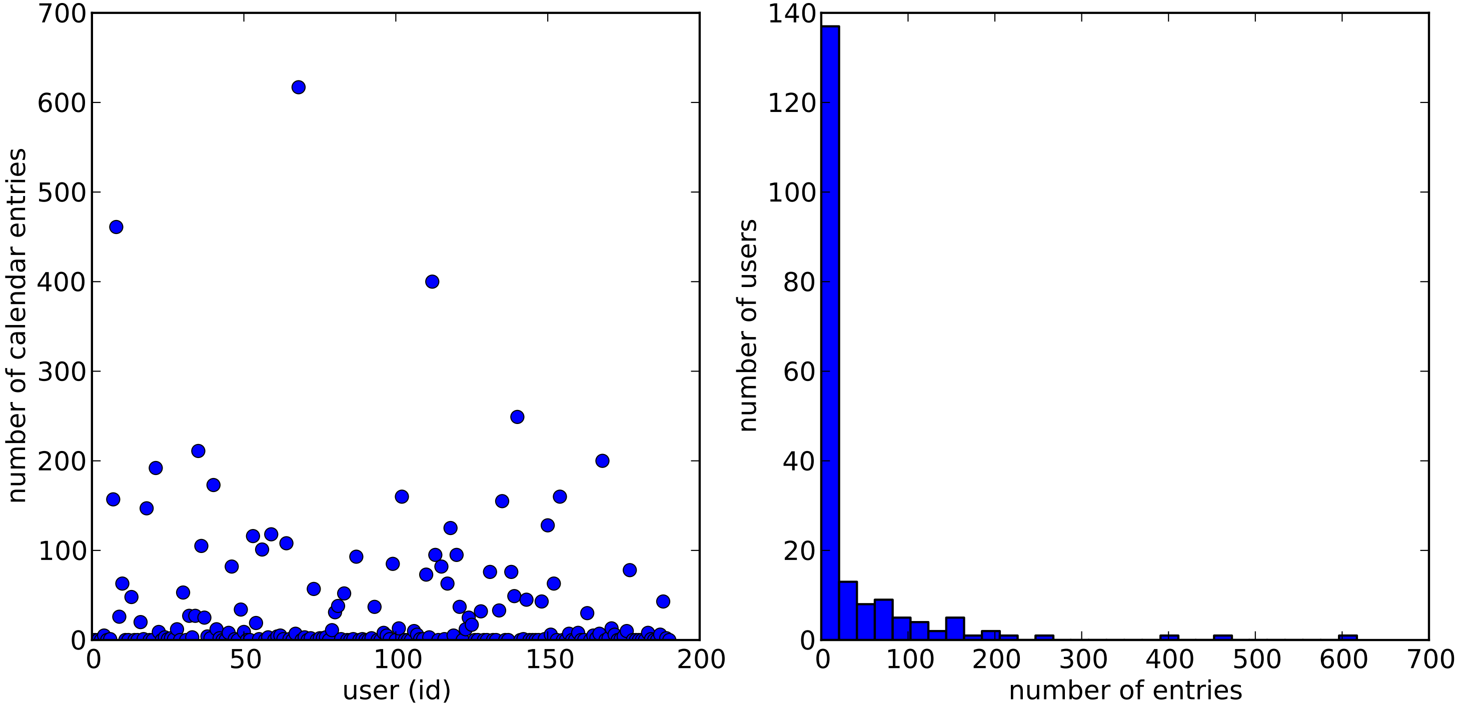
It took ONE YEAR for EPFL to get data access (although data had already been in Nokia's databases).

Data volume

∼ 150k-entries/100MB of data per user per month

Number of GPS points	11,531,652
Number of calls	247,448
Duration of calls	6,903h
Number of sms	179,358
Number of video made	3,890
Number of pictures taken	54,537
Number of unique BT	543,517
Number of unique WIFI	572,910
Number of unique cell towers (63 countries)	100,505
Number of unique cell towers (CH)	28,945
Number of acceleration samples	1,344,198
Number of application events captures	8,280,554
Number of phone book entries	115,134

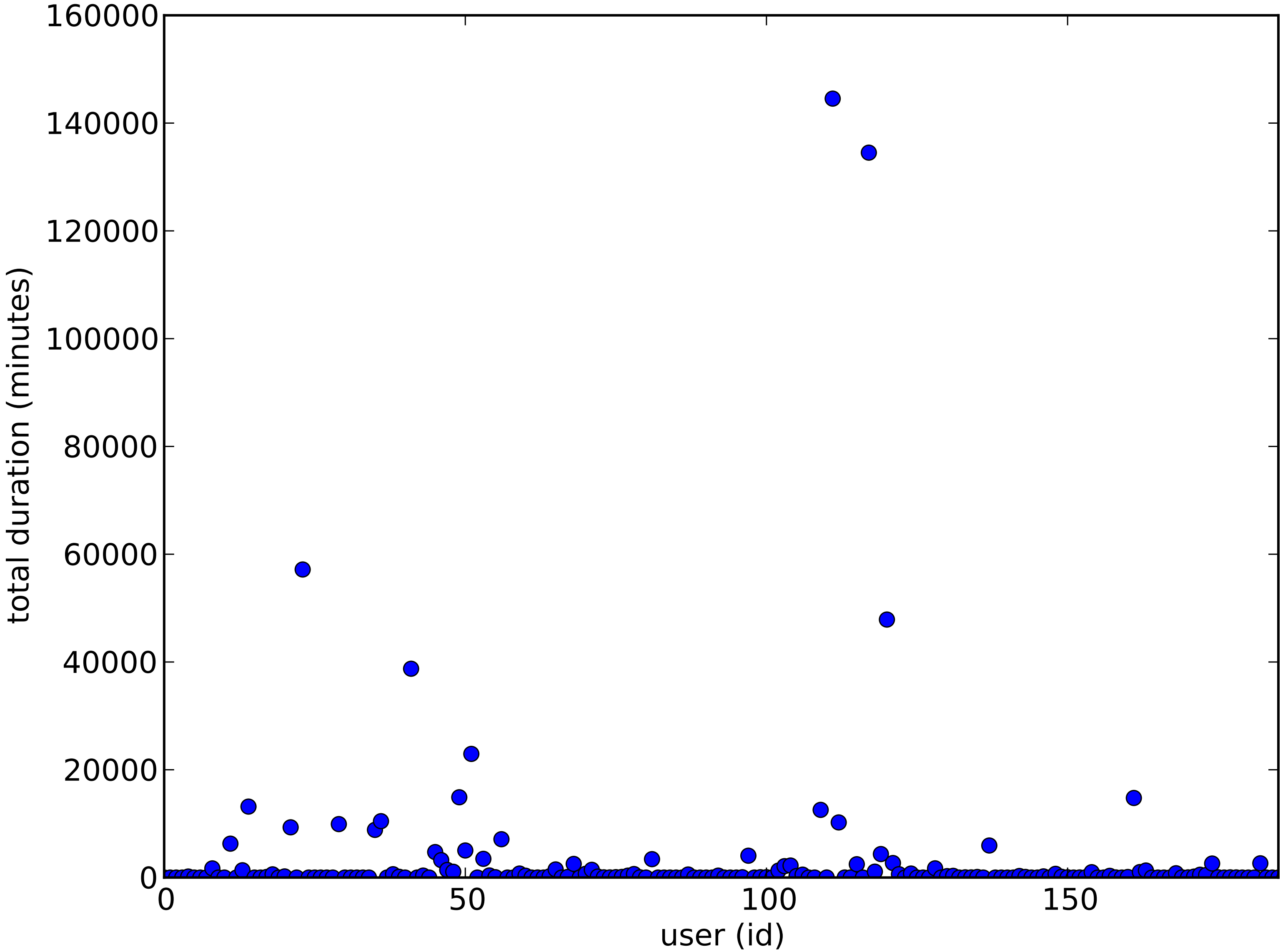
Calendar: number of entries



TRANSP-OR

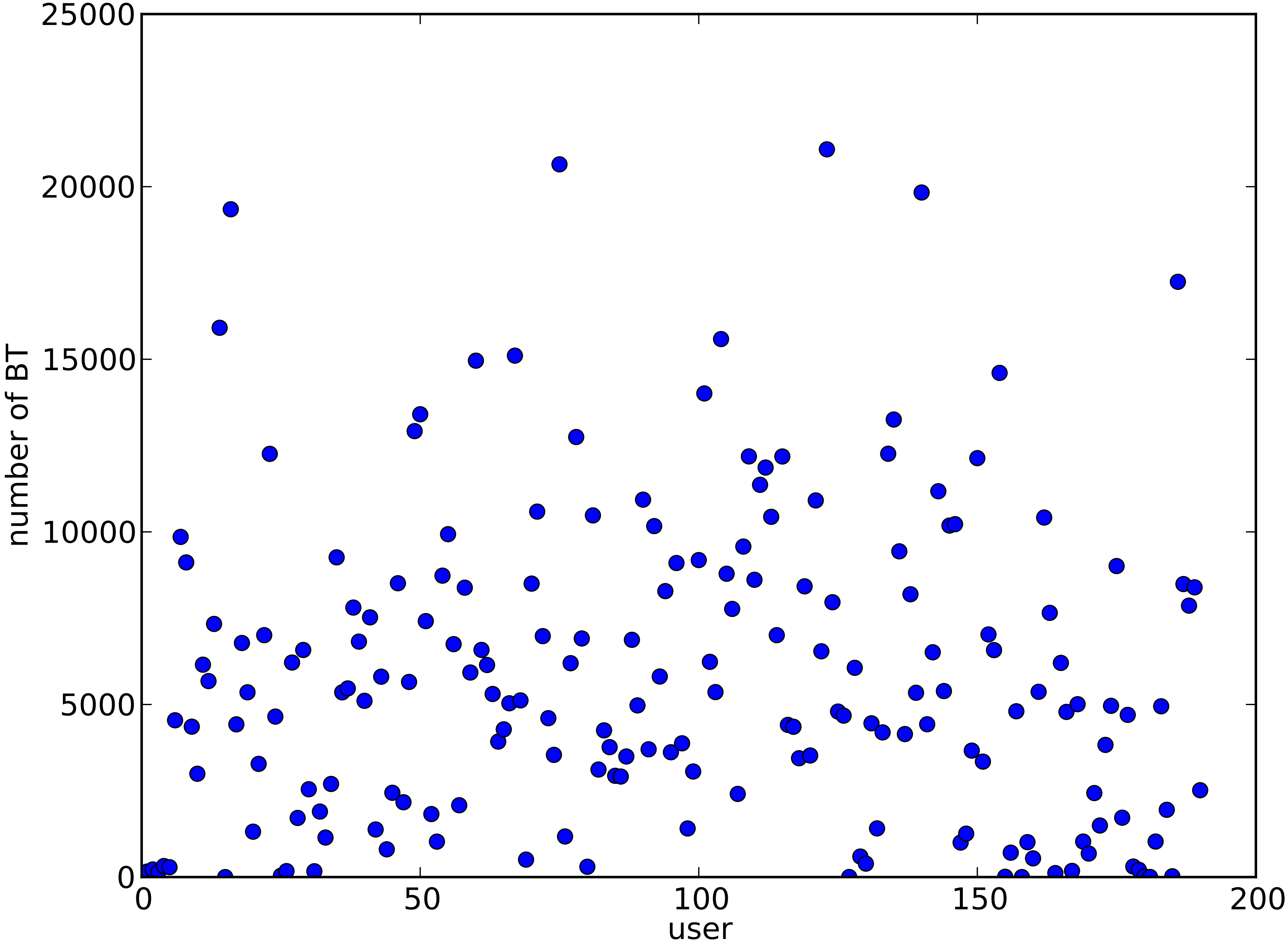
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Media play

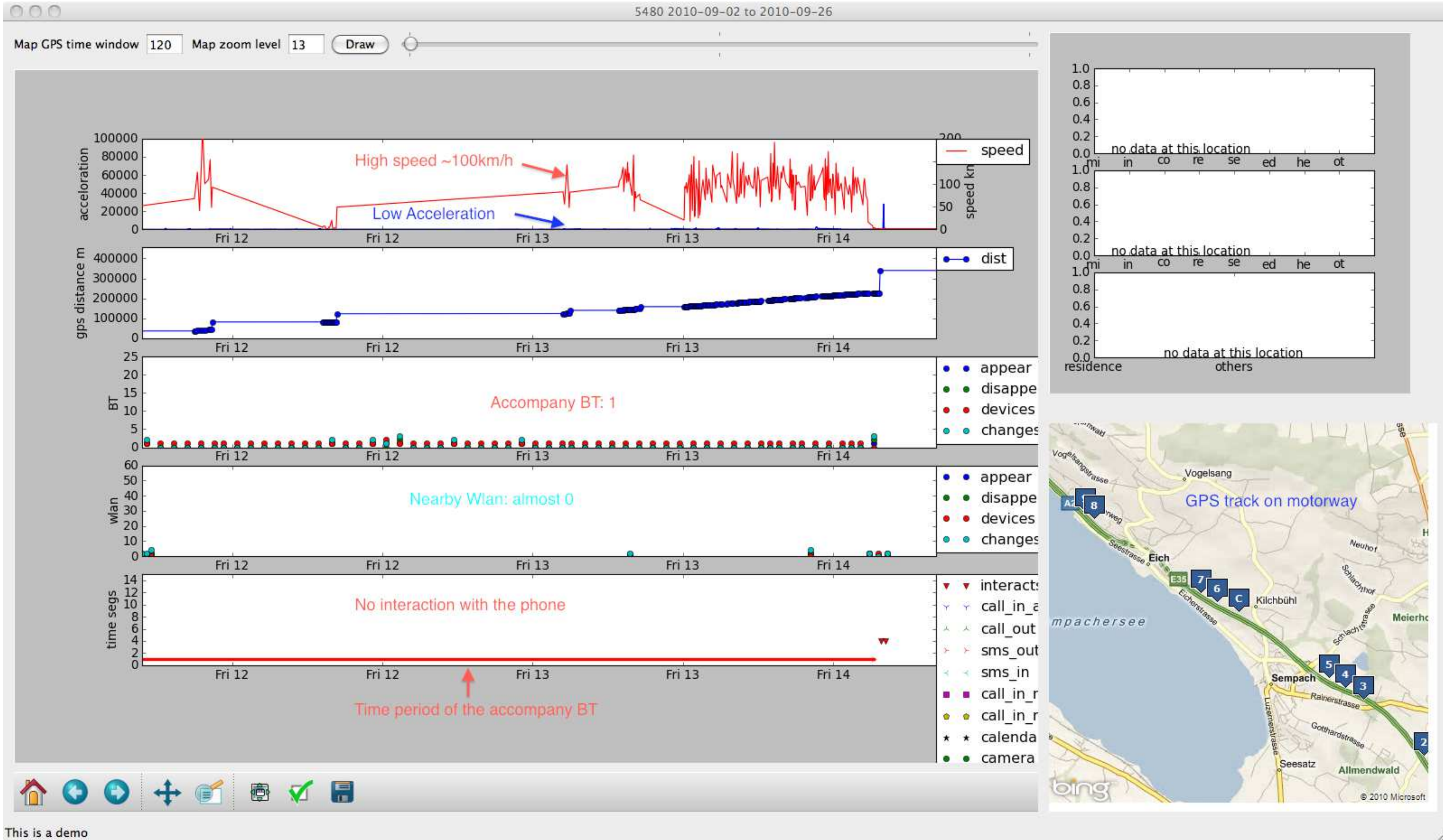


160000
140000
120000
100000
80000
60000
40000
20000
0
total duration (minutes)
0
50
100
150
user (id)

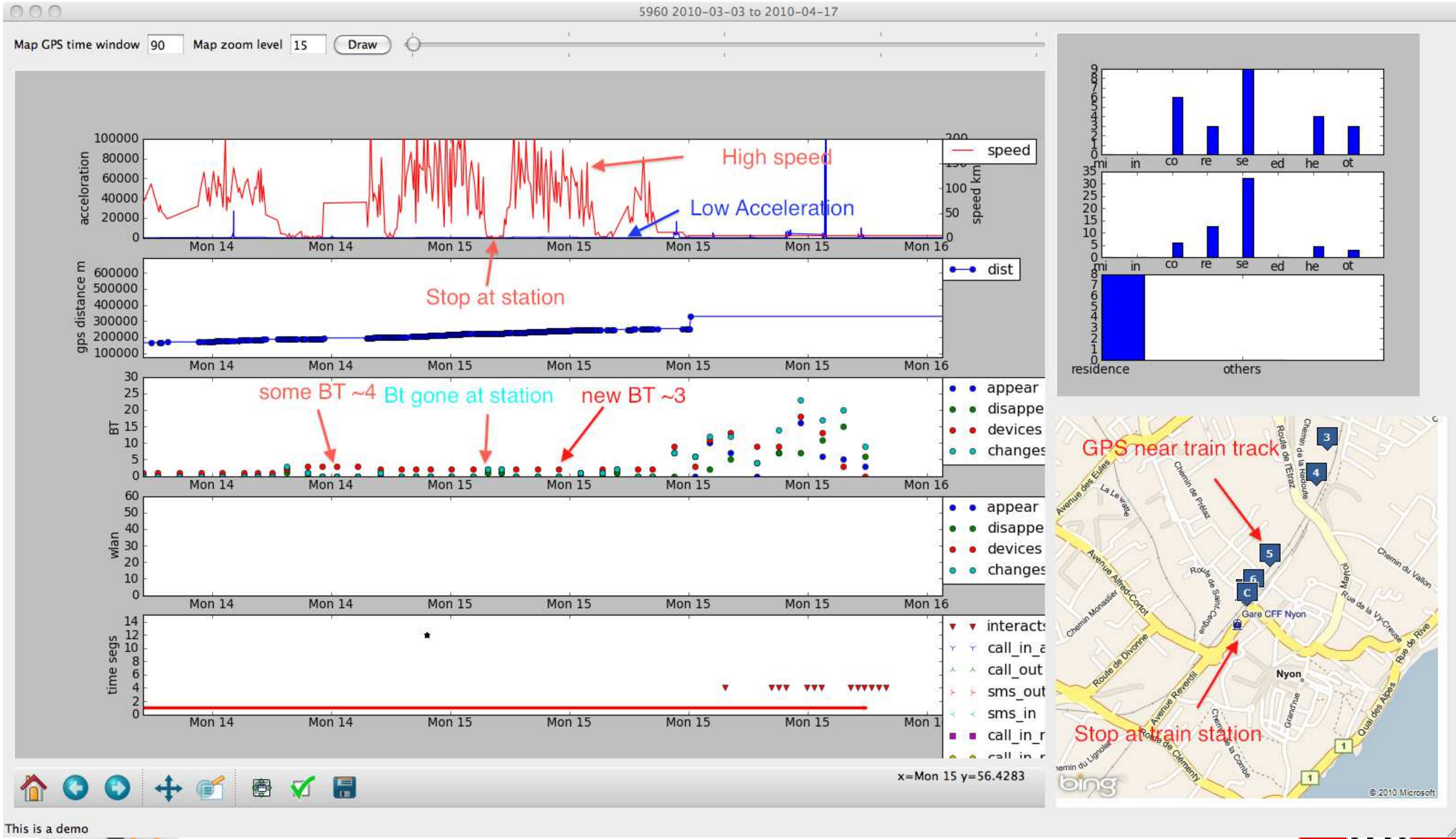
Number of Bluetooth devices



Mobility patterns: car



Mobility patterns: train



Data

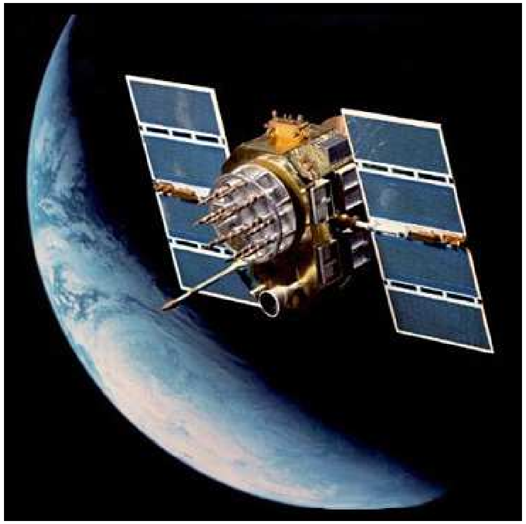
Challenges

- Technological: battery life
- Legal: privacy
- Technical: huge volume of date

Opportunities

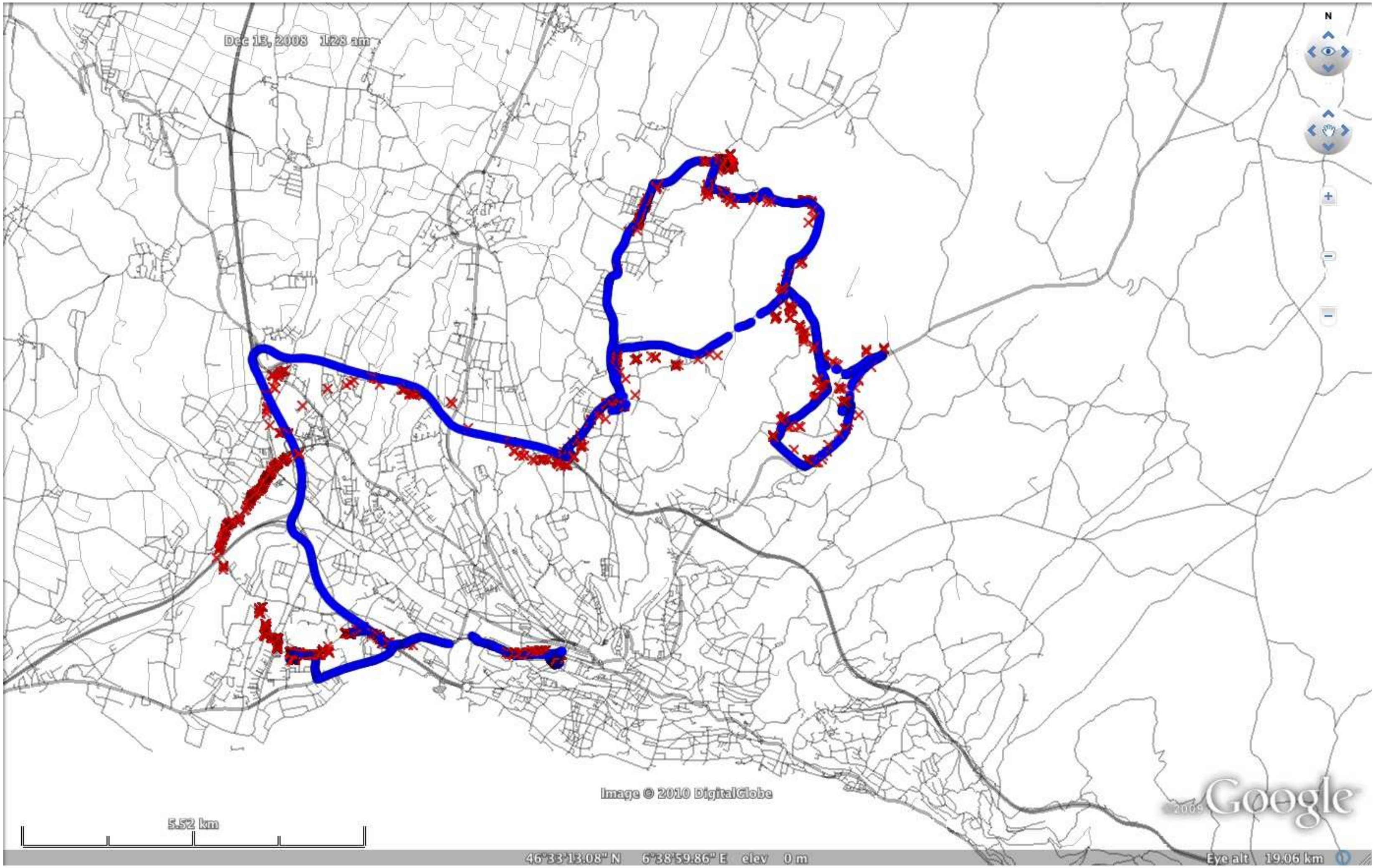
- Complex mobility patterns
- Mode
- Route
- Activities

Route choice: the chosen route

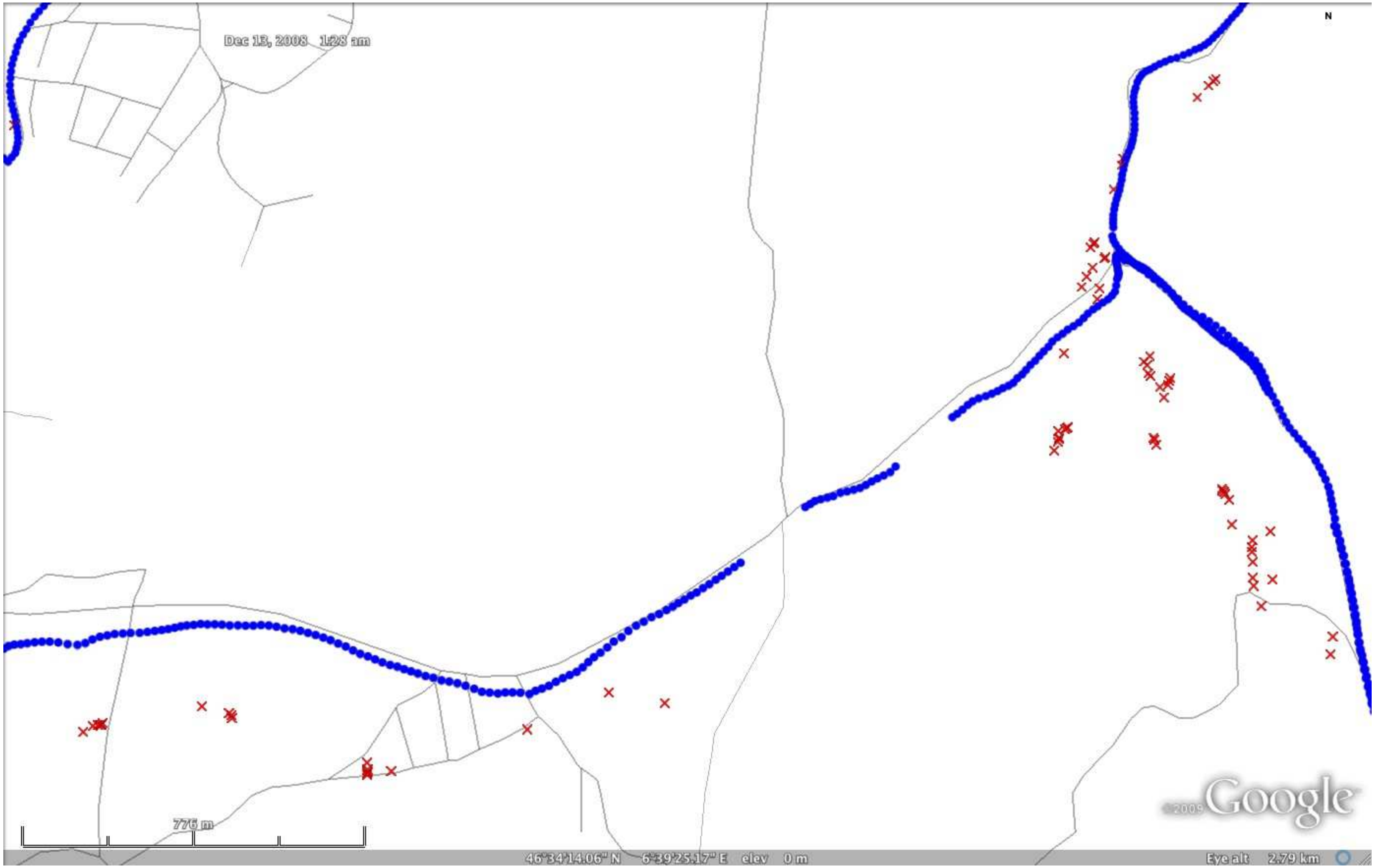


- Focus on GPS data from smartphone
- Objective: reconstruct actual paths

Issues


Dec 13, 2008 1:28 am
Image © 2010 DigitalGlobe
5.52 km
46°33'13.08" N 6°38'59.86" E elev 0 m
Eye alt 19.06 km
Google

Issues



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Issues

- Low data collection rate to save battery (every 10 seconds)
- Inaccuracy due to technological constraints
- Smartphone carried in bags, pockets: weaker signal
- Map matching algorithms do not work with this data

Measurement equations

Objective (derivation in the appendix):

- Given a path p
- Given a sequence of GPS data $(\widehat{x}_1, \ldots, \widehat{x}_T)$
- What is the likelihood that the sequence has been generated by a smartphone moving along path p?
- Note: different approach from map matching, which is essentially a projection procedure.
- We derive

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p),$$

- ... recursively

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p) = \Pr(\widehat{x}_T | \widehat{x}_1, \ldots, \widehat{x}_{T-1}, p) \Pr(\widehat{x}_1, \ldots, \widehat{x}_{T-1} | p).$$

Case study: true path — [-11.3]

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Case study: path with a deviation (1) — [-12.9]

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Case study: path with a deviation (2) — [-13.2]

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Case study: log likelihood from measurement equati

True path	-11.3
Deviation 1	-12.9
Deviation 2	-13.2

- Results are consistent with intuition

Route choice: the non chosen routes

- Choice model: $P_n(i|\mathcal{C}_n)$
- Route choice: what is $\mathcal{C}_n$?
- Many "behaviorally motivated" heuristics proposed in the literature.
- Most of the time, the chosen route is not included.
- `Frejinger, Bierlaire and Ben-Akiva (2009)` propose an econometric approach.
- Idea:
    - Assumption: **all** paths connecting the OD pair are relevant.
    - Issue: enumeration is prohibitive.
    - Solution: sampling of alternatives.

Sampling of alternatives

- Sample $\mathcal{C}_n$ with replacement from $\mathcal{C}$ according to $\{q(i)\}_{i\in\mathcal{C}}$
- Add the chosen alternative
- k_{in} is the number of times alternative i is contained in $\mathcal{C}_n$
- Correct for sampling when estimating logit model

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in} + \ln\left(\frac{k_{in}}{b(i)}\right)}}{\sum_{j\in\mathcal{C}_n} e^{\mu V_{jn} + \ln\left(\frac{k_{jn}}{b(j)}\right)}}$$

where $\{b(i)\}_{i\in\mathcal{C}}$ is such that $q(i) = b(i)/\sum_{j\in\mathcal{C}} b(j)$

Objective: sample paths according to pre-specified $\{b(i)\}_{i\in\mathcal{C}}$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Metropolis-Hastings algorithm

- Given
  - a finite state space,
  - positive weights $\{b(i)\}_i$,
  - and irreducible Markov process
- the Metropolis-Hastings algorithm generates a Markov chain that converges to

$$q(i) = b(i)/\sum_j b(j).$$

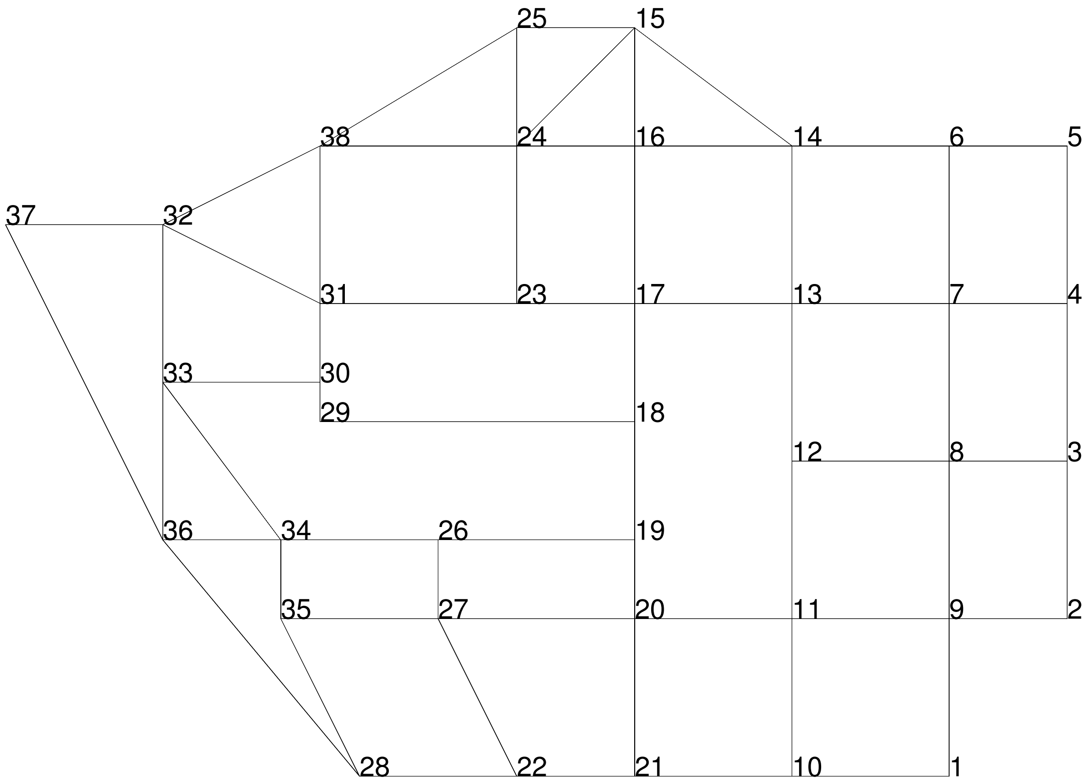
Using MH for path sampling

- State space comprises all possible paths
- Weights $b(i)$ favor plausible paths (importance sampling)
- Typically, paths with length close to the shortest path have high probability to be sampled
- Based on a Markov process creating local path modifications
  - too little variability: slow convergence
  - too much variability: random search
- a great deal of technical details must be addressed to obtain a valid algorithm.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Simple example



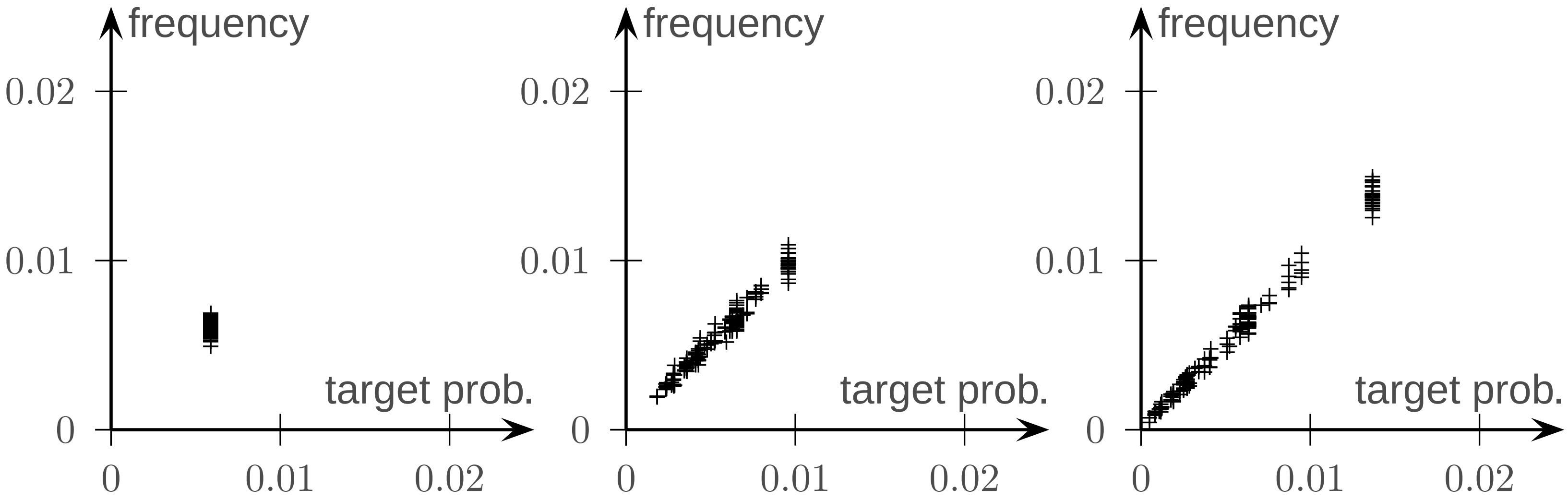
Simple example

- Target weights:

$$b(i) = \exp[-\mu\delta(\Gamma)]$$

  where $\delta(\Gamma)$ is the length of path Γ.
- Note: $\mu = 0$ means equal probability.

Scatter plots



(a) $\mu = 0.0$ (b) $\mu = 2.0$ (c) $\mu = 4.0$

Tel-Aviv example

Conclusion

- Route choice modeling is difficult.
- Data: smartphones
- Identify the chosen route
  - Deal with inaccuracy and low rate
  - Probabilistic map matching
- Identify the non chosen routes
  - Sampling of paths
  - Markov Chain Monte-Carlo method
  - The devil is in the details...
  - but it works!

References

- Bierlaire, M., and Frejinger, E. (2008). Route choice modeling with network-free data, Transportation Research Part C: Emerging Technologies 16(2):187-198.

  http://dx.doi.org/10.1016/j.trc.2007.07.007

- Bierlaire, M., Chen, J., and Newman, J. P (2010). Modeling Route Choice Behavior From Smartphone GPS data. Technical report TRANSP-OR 101016. Transport and Mobility Laboratory, ENAC, EPFL.

  http://transp-or.epfl.ch/php/abstract.php?type=1&id=BierChenNewm10

References

- Flötteröd, G., and Bierlaire, M. (2011). Metropolis-Hastings sampling of paths. Technical report TRANSP-OR 110606. Transport and Mobility Laboratory, ENAC, EPFL.

  http://transp-or.epfl.ch/documents/technicalReports/FloeBier11.pdf

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE