

Cleansing GPS-data from person based travel surveys in urban environments

Peter Bro

Agenda

- The survey
- Challenges in GPS surveys
- Trip identification

THE SURVEY

Overall goal

- To evaluate GPS surveys as a mean of collecting travel information as a supplement to traditional trip diaries

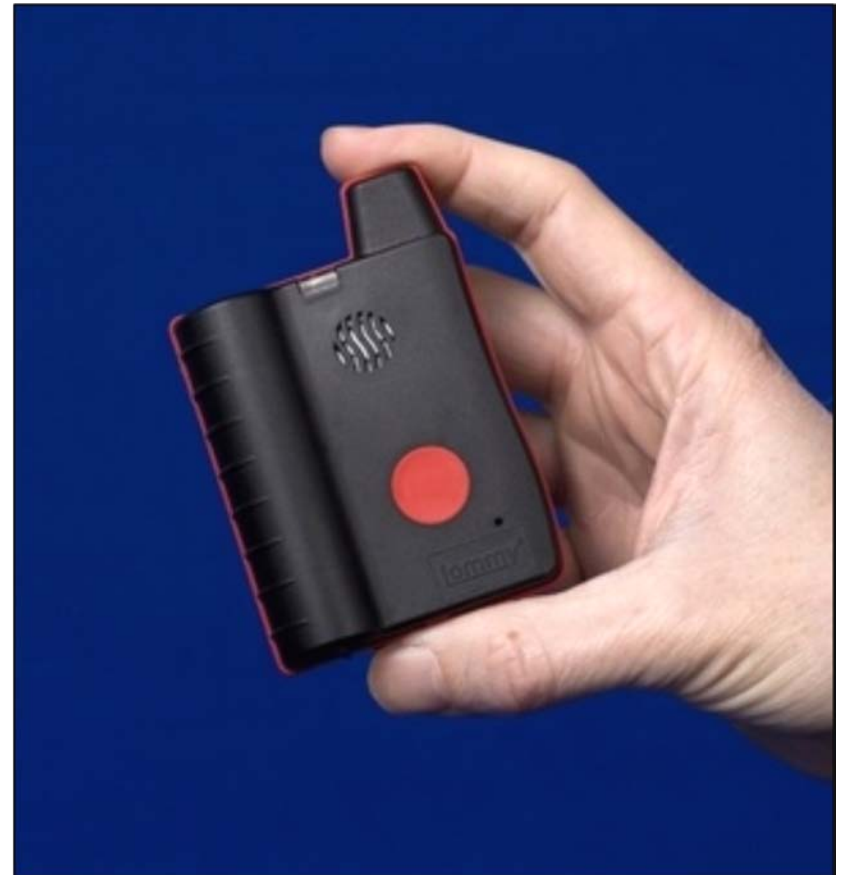
Hardware

Diverse Urban Spaces

På baggrund af flere tests besluttede vi at anvende Flextrack Lommy©, der har både GPS, GSM og GPRS enheder

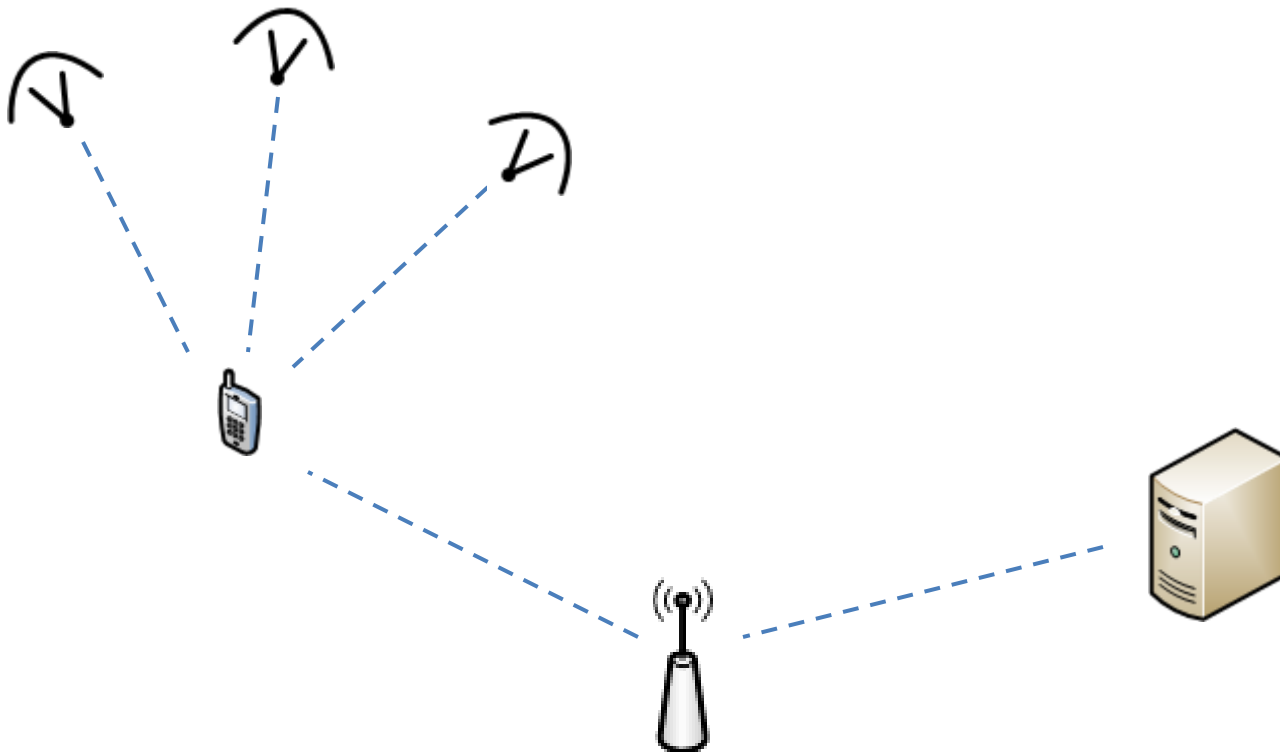
Designet af enheden er simpelt og den er ganske lille (74x61x23 mm og 99 gram) og den har kun én tænd/sluk knap

Lommyen giver desuden mulighed for at følge enheden online og i real-time, sådan undersøgelserne kan monitoreres løbende, og bortkomne enheder kan trackes og indhentes



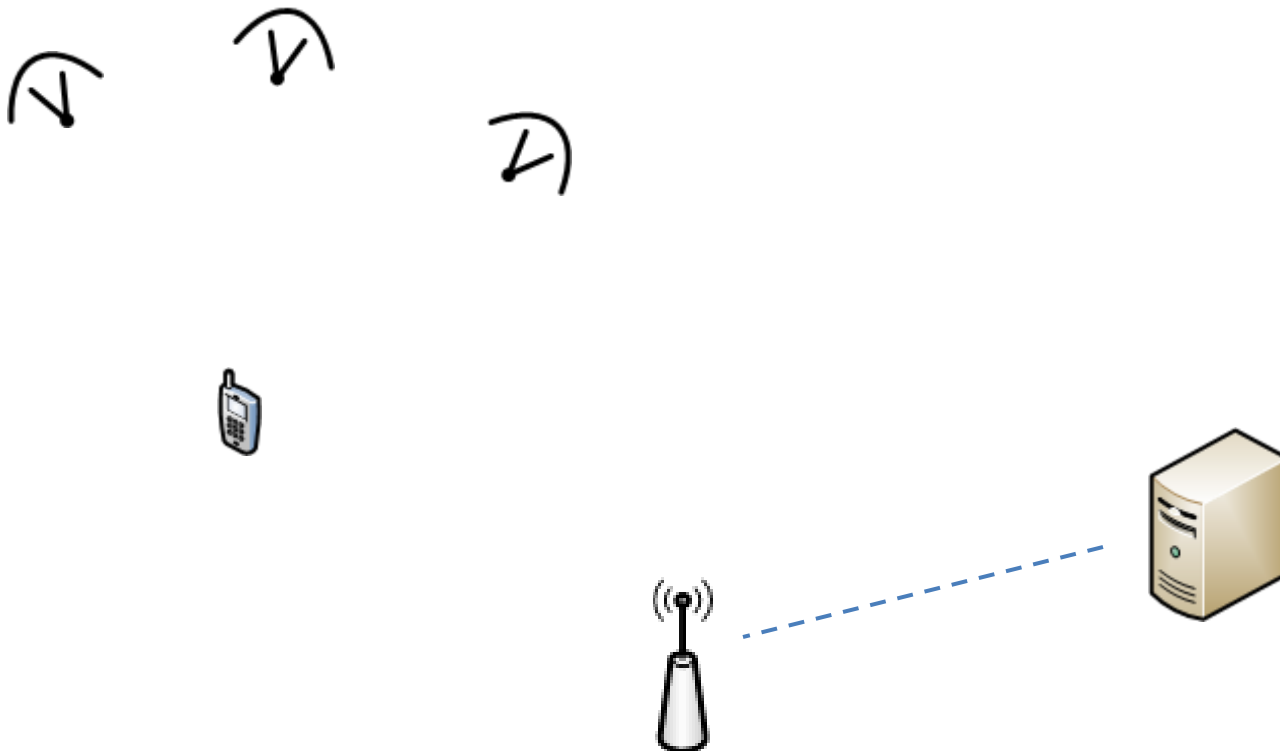
Data flow

Diverse Urban Spaces



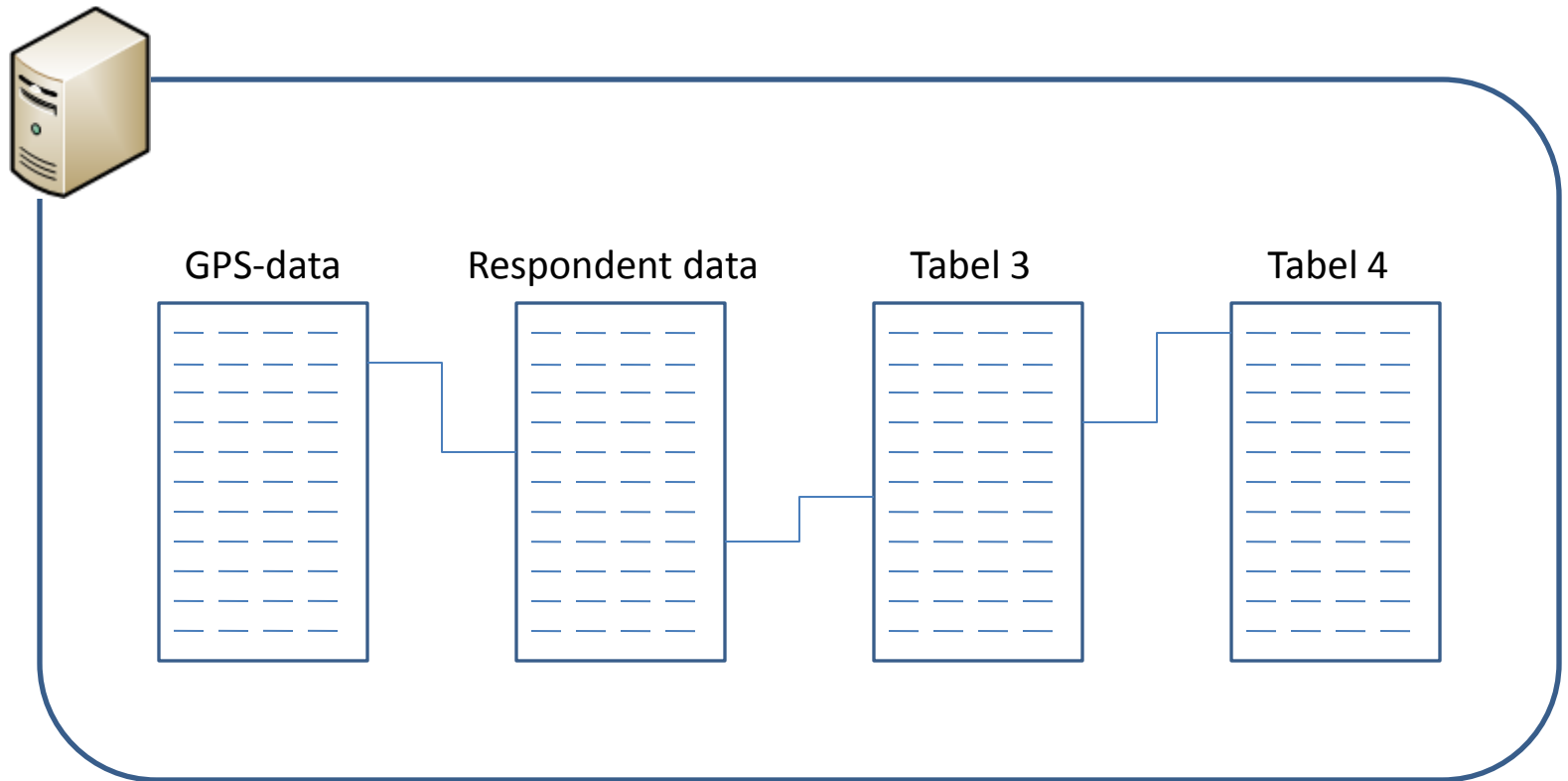
Data flow

Diverse Urban Spaces



Data flow

Diverse Urban Spaces



Methodological setup

- 250 young people were selected to participate and carry a GPS and answer a trip-diary each evening throughout a period of seven days
- All participants are students at high school level in Aalborg Municipality
 - 50 respondents from each school at the time
- Data collection were done outside the holidays
 - 4 surveys before the summer holiday
 - 4 surveys after the summer holiday

Methodological setup



Trip diary

Unges mobilitet - en undersøgelse af unges brug af byen - Windows Internet Explorer

http://www.detmangfoldigebymum.dk/aalborg/byrum/1b3.php

Google C- Go | Bookmarks | 1 blocked | Check | AutoLink | AutoFill | Send to | Settings

Det mangfoldige byrum

Tur nr. 3, den 20/5 [Send en e-mail til du@2.aod.aau.dk](mailto:du@2.aod.aau.dk) hvis der er problemer med GPS'en eller spørgeskemaet i dag (bemærk at GPS punkterne kan godt springe lidt)

1. Hvornår begyndte turen? Skriv time: Skriv minut: Hvis du i løbet af dagen har glemt din gps eller skudt din gps, skal du udfylde spørgeskemaet alligevel.

2. Hvornår sluttede turen? Næste dag Vælg Vælg Husk at afpasse tiden for turstart og turafslutning med kortet.

3. Hvilket transportmiddel benyttede du til størstedelen af turen? Vælg

4. Hvem foretog du turen sammen med? Vælg hvem

5. Hvad kostede denne tur for dig (Udregn hvis du har abonnement)? Vælg

6. Hvilken aktivitet foretog du på dit bestemmelsessted? Vælg

7. Brugte du internet på dette bestemmelsessted (angiv min. du aktivt brugte siden)? Vælg

8. Hvornår besluttede du at foretage aktiviteten? Vælg

9. Hvilken udgift var forbundet med aktiviteten (Udregn hvis du har abonnement)? Vælg

Klik her hvis turen/aktiviteten falder udenfor ovenstående svarmuligheder (f.eks. var hjemme hele dagen)

<< Tilbage | Næste tur >> | Gem tur og afslut turdagbog

Tur nr.	Transportmiddel til aktivitet	Ragestid til aktivitet	Aktivitet	Vanghed af aktivitet	Beslutning	Sammen med:		Udgift
				Fra kl.	Til kl.	Antal	Hvem	
Slet	Rat 2	30 min	Fritidsjob	14:0	14:30	Mere end én dag	Mere end fire personer	Familie 101 - 200 kroner
Slet	Rat 1	840 min	Fritids- og sociale aktiviteter - Frivilligt arbejde, kurser og foreningsmøder	0:0	14:0	Fast tilbagevendende aktivitet	Ingen, jeg var alene	Familie 51 - 100 kroner

Done

Microsoft PowerPoint... | Indbakke - Microsoft... | Report | problemformulering... | Unges mobilitet - e...

Internet | Protected Mode: On | 100% | 71.16 | tirsdag

Trip diary

Young people's urban mobility

Unges mobilitet - en undersøgelse af unges brug af byen - Windows Internet Explorer

http://www.detmangfoldigebym.dk/aalborg/byrum/1b3.php

Google C- Go | Bookmarks | 1 blocked | Check | AutoLink | AutoFill | Send to | Settings

Det mangfoldige byrum | Unges mobilitet - en u...

1. Hvornår begyndte turen? Skriv time: 12 Skriv minut: 0 Hvis du i løbet af dagen har glemt din gps eller skullet din gps, skal du udfylde spørgeskemaet alligevel.

2. Hvornår sluttede turen? Næste dag 13 55 Husk at afpasse tiden for turstart og turafslutning med kortet.

3. Hvilket transportmiddel benyttede du til størstedelen af turen?
Gang

4. Hvem foretog du turen sammen med?
Familie
Hvor mange personer foretog du turen sammen med?
To personer

5. Hvad kostede denne tur for dig (Udregn hvis du har abonnement)?
51 - 100 kroner

6. Hvilken aktivitet foretog du på dit bestemmelsessted?
Fritidsjob

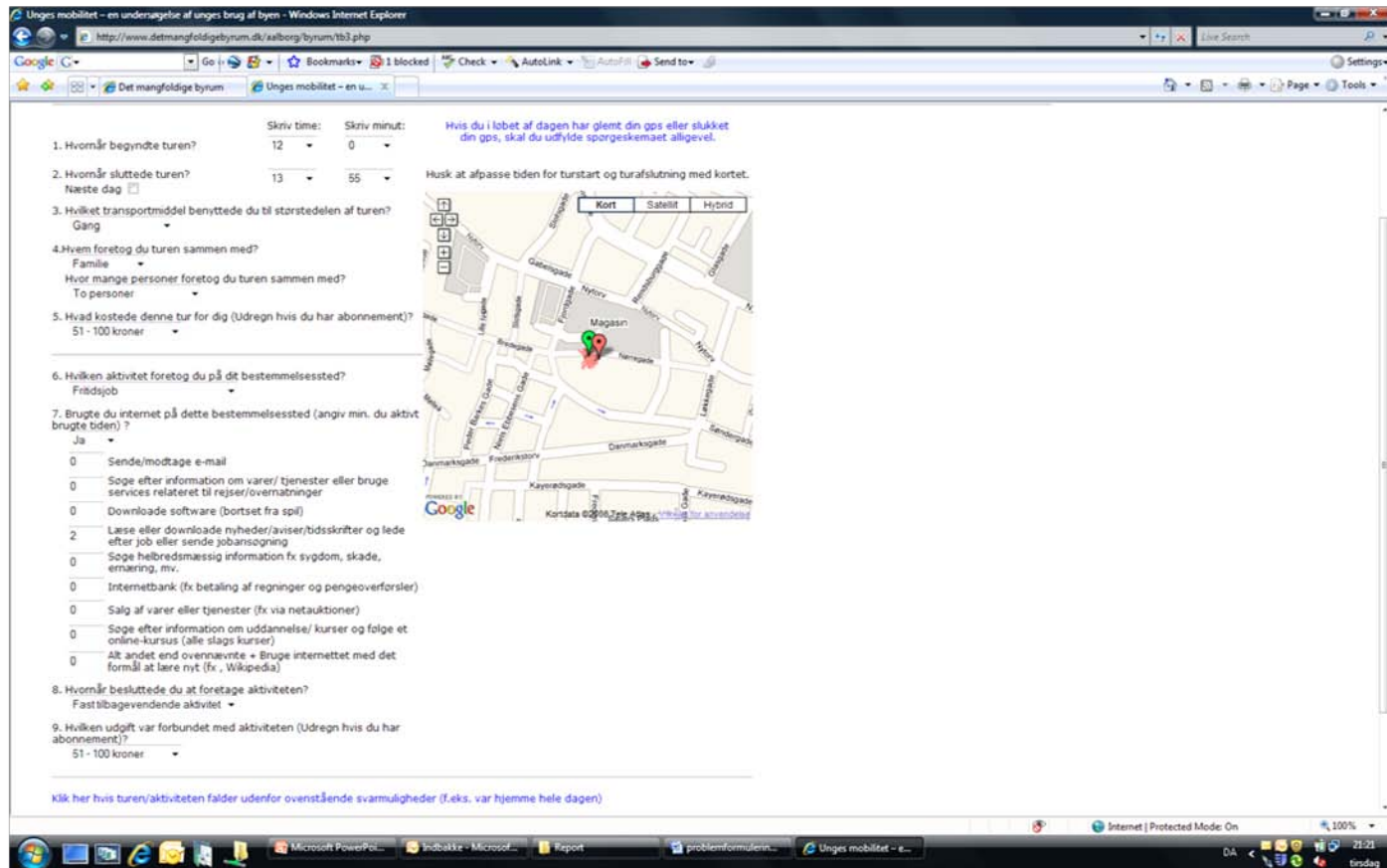
7. Brugte du internet på dette bestemmelsessted (angiv min. du aktivt brugte tiden)?
Ja

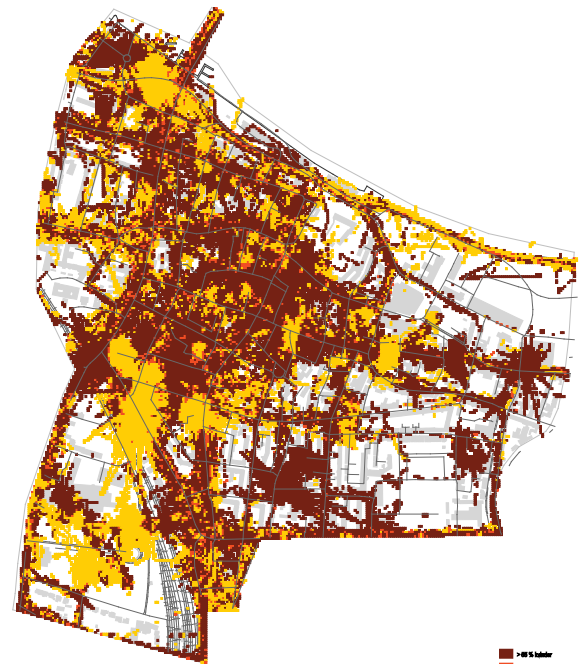
0 Sende/modtage e-mail
0 Soge efter information om varer/ tjenester eller bruge services relateret til rejser/overnatninger
0 Downloade software (bortset fra spil)
2 Læse eller downloade nyheder/avisser/bidsskrifter og lede efter job eller sende jobansøgning
0 Soge helbredsmaessig information fx sygdom, skade, ernæring, mv.
0 Internetbank (fx betaling af regninger og pengeoverforsler)
0 Salg af varer eller tjenester (fx via netauktioner)
0 Soge efter information om uddannelse/ kurser og følge et online-kursus (alle slags kurser)
0 Alk andet end ovennaevnte + Brug internetet med det formål at lære nyt (fx., Wikipedia)

8. Hvornår besluttede du at foretage aktiviteten?
Fasttilbagevendende aktivitet

9. Hvilken udgift var forbundet med aktiviteten (Udregn hvis du har abonnement)?
51 - 100 kroner

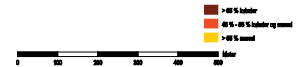
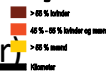
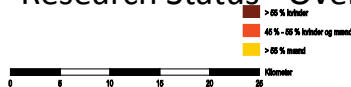
Klik her hvis turen/aktiviteten falder udenfor ovenstående svæmuligheder (Læks. var hjemme hele dagen)





<http://www.detmangfoldigebyrum.dk/>

Research Status - Overview - (Unge i alderen 16 -23 år)

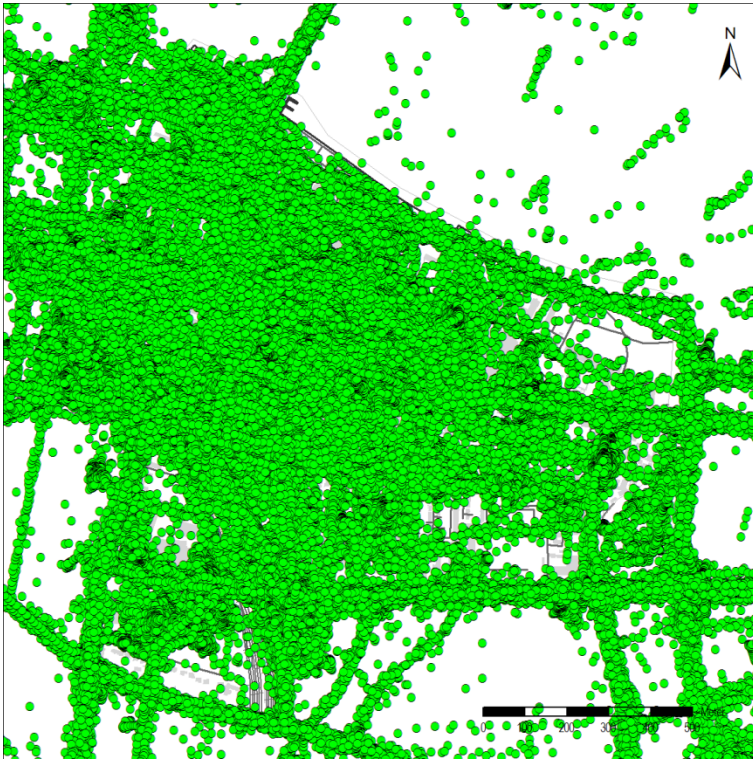


CHALLENGES IN GPS SURVEYS

The pros and cons of GPS surveys

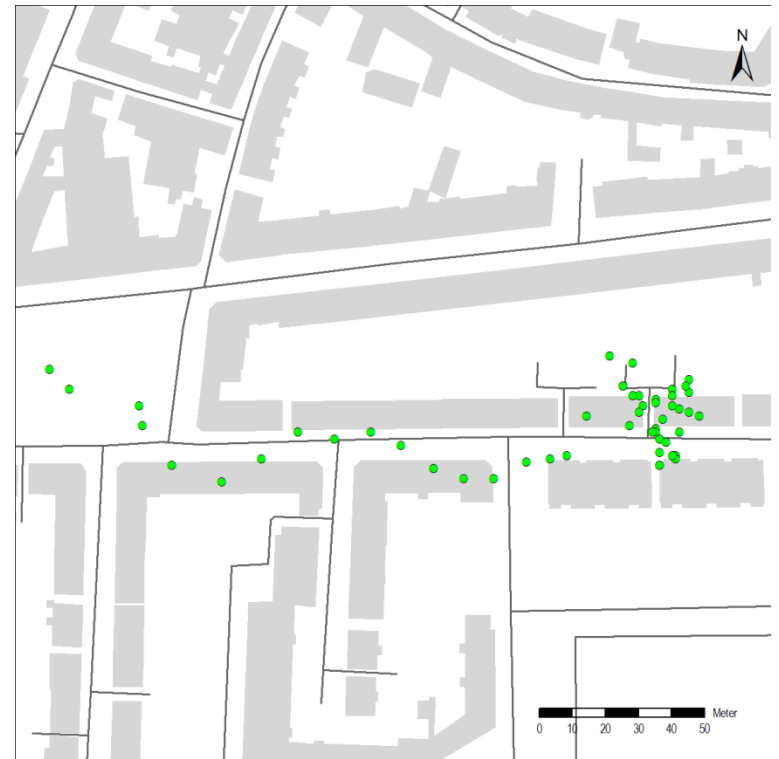
- Easy to collect a lot of data
- A possibility to eliminate errors due to limited memory
- Too much data is collected
- Data is hard to interpret and process
- No guarantee that respondents carry the GPS all the time

All the data is one big bunch



A closer look

- During trips the loggings are in a nice line
- During stays they scatter

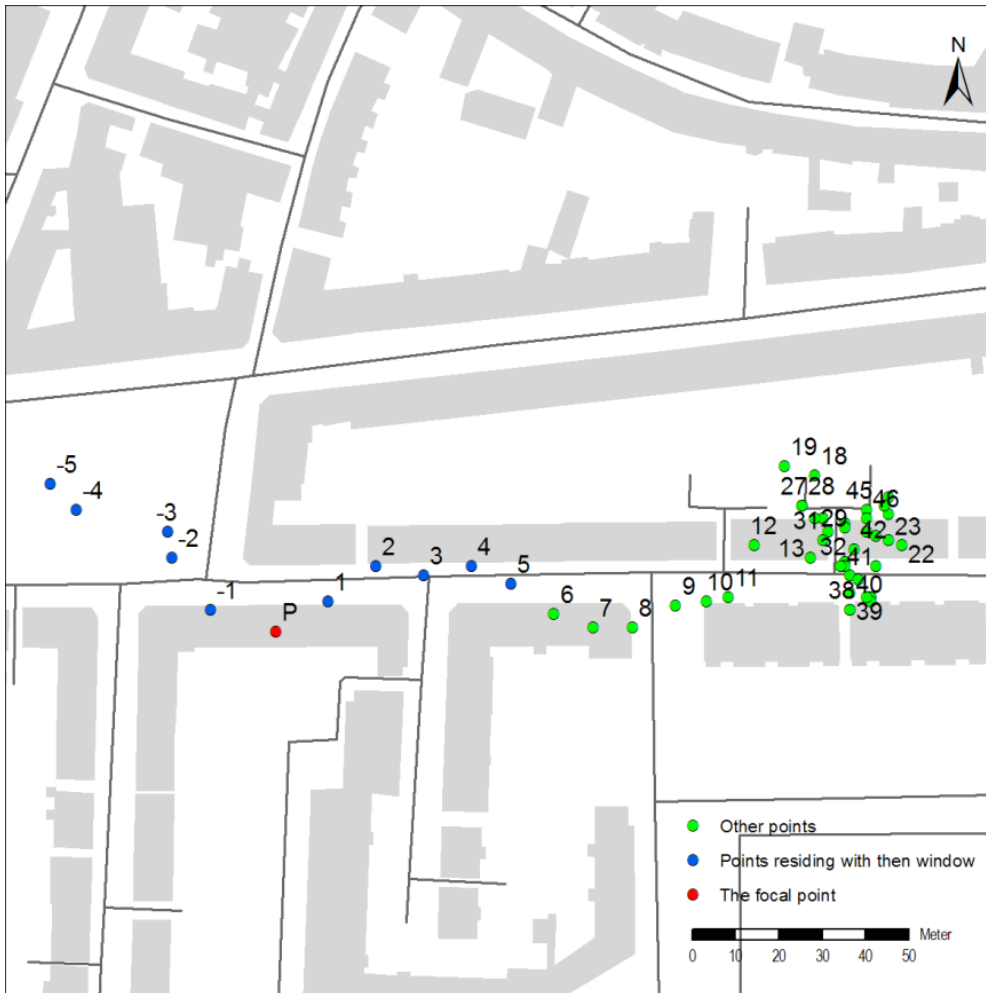


TRIP IDENTIFICATION

Data cleansing

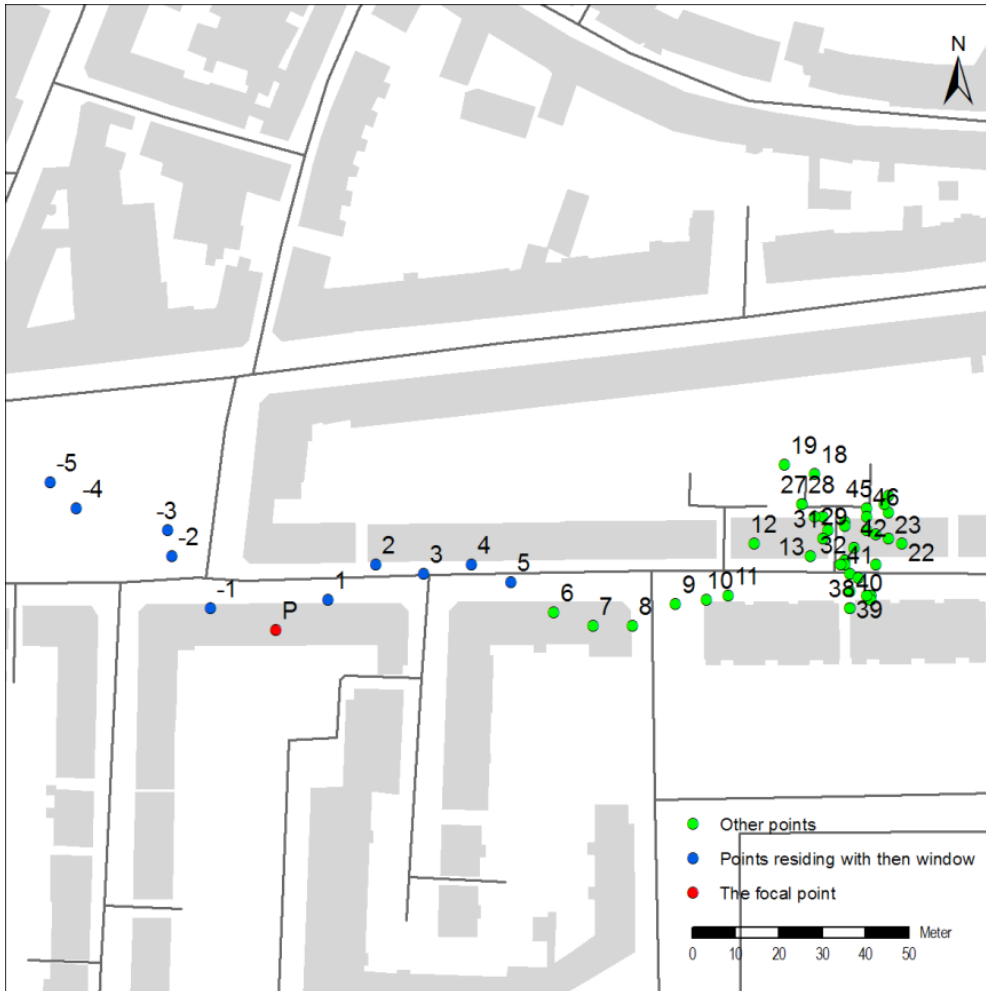
- A sample data set of roughly 120.640 loggings were manually sorted
 - 413 activities
 - 451 trips
 - 24 different respondents
 - 8 different days

The window approach



Point ID	Speed	Direction	Focal point relation
7	4	139	-8
8	3	144	-7
9	4	103	-6
10	7	142	-5
11	3	90	-4
12	0	177	-3
13	0	157	-2
14	5	140	-1
15	3	135	P
16	4	78	1
17	5	79	2
18	4	97	3
19	4	85	4
20	3	98	5
21	5	113	6
22	5	105	7
23	4	98	8
24	4	85	9
25	3	84	10
26	2	110	11

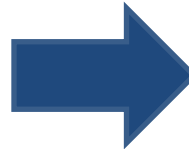
The window approach



Point ID	Speed	Direction	Focal point relation	Direction difference
7	4	139	-8	
8	3	144	-7	
9	4	103	-6	
10	7	142	-5	
11	3	90	-4	52
12	0	177	-3	87
13	0	157	-2	20
14	5	140	-1	17
15	3	135	P	62
16	4	78	1	1
17	5	79	2	18
18	4	97	3	12
19	4	85	4	13
20	3	98	5	
21	5	113	6	
22	5	105	7	
23	4	98	8	
24	4	85	9	
25	3	84	10	
26	2	110	11	

The window approach

Point ID	Speed	Direction	Focal point relation	Direction difference
7	4	139	-8	
8	3	144	-7	
9	4	103	-6	
10	7	142	-5	
11	3	90	-4	52
12	0	177	-3	87
13	0	157	-2	20
14	5	140	-1	17
15	3	135	P	62
16	4	78	1	1
17	5	79	2	18
18	4	97	3	12
19	4	85	4	13
20	3	98	5	
21	5	113	6	
22	5	105	7	
23	4	98	8	
24	4	85	9	
25	3	84	10	
26	2	110	11	



Point ID	Speed	Direction	Focal point relation	Direction change sum
7	4	139	-8	
8	3	144	-7	
9	4	103	-6	
10	7	142	-5	
11	3	90	-4	
12	0	177	-3	
13	0	157	-2	
14	5	140	-1	
15	3	135	P	282
16	4	78	1	
17	5	79	2	
18	4	97	3	
19	4	85	4	
20	3	98	5	
21	5	113	6	
22	5	105	7	
23	4	98	8	
24	4	85	9	
25	3	84	10	
26	2	110	11	

Data cleansing

- Different variables were developed in order to automate data cleansing
 - $\text{avg}(\text{DIRCHN})_{t1,t2}$
 - $\text{avg}(\text{SPEED})_{t1,t2}$
 - $\text{sum}(\text{DIST}_r)_{t1,t2}$
 - $\text{avg}(\text{HDOP})_{t1,t2}$

Data cleansing

$$P(X = \text{trip}) = \frac{1}{(1 + \exp(1,688 + 0,278x))}$$

– Where

- x = the number of loggings within a 20 meter radius from the focal point and registered within the time span of 120 seconds before and after the focal point

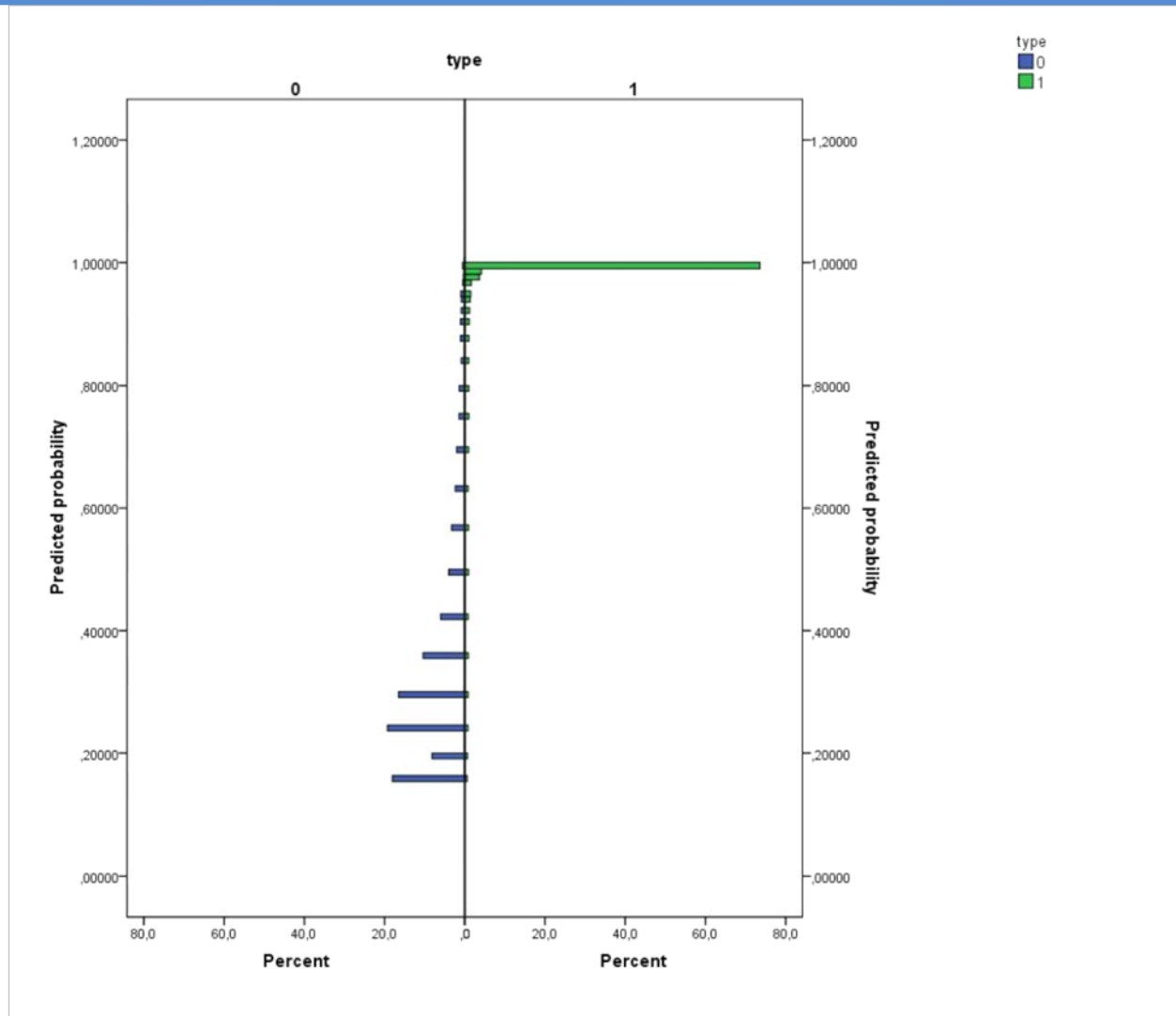
Data cleansing

- This classifies 92,8% of all loggings correct
 - 94,6% of the activity loggings are classified correctly
 - 82,3% of the trips loggings are classified correctly
 - Nagelkerke R²: 0,695
 - Sig: 0,000

Cleansed data



Systematic errors?



Systematic errors?



CONCLUSIONS

Conclusions

- General
 - Collecting travel data with GPS is relatively easy
 - Data processing is time consuming and requires good computational power
 - It is possible to automatically cleanse the data based upon attributes in the loggings
- Specific
 - The developed algorithm classifies 92,8% of the loggings correct
 - It classifies 82,3% of the trip loggings correct
 - The algorithm tends to misclassify the first and the last loggings of trips