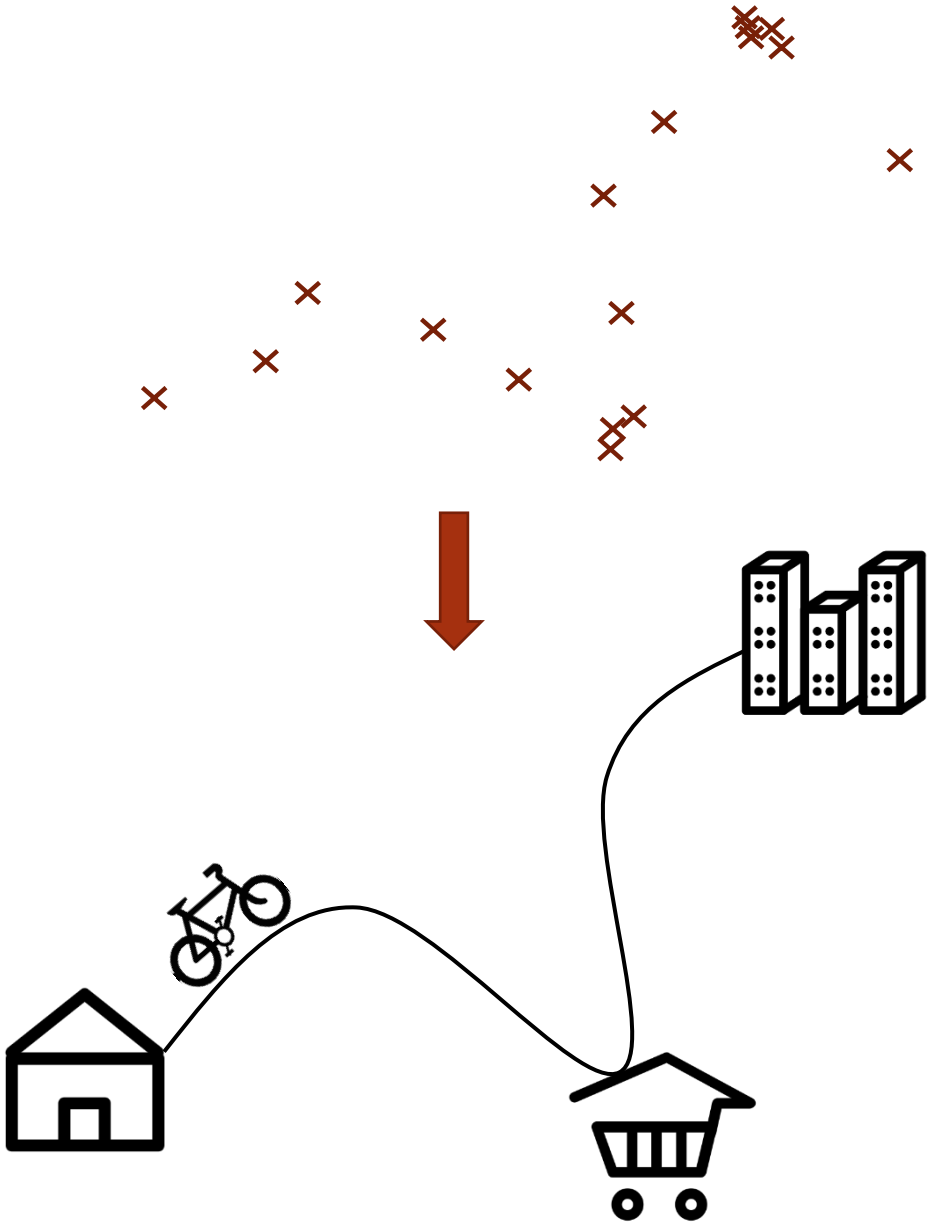
Algorithm-induced biases in data representativeness

The case of activity inference from mobile phone traces

Julie Chrétien
Postdoctoral fellow

Laboratoire Ville Mobilité Transport

Context



Logos created by
Daniele Catalanotto,
from Noun Project

Focus

Home + work inferences

Rule-based algorithms:

If (condition) then the activity is…

Focus

Temporal conditions

Conditions		
Only time or frequency	Time or frequency and other factors	Without time or frequency
Schneider et al. (2013) Ahas et al. (2010) Çolak et al. (2015) Li et al. (2015) Picornell et al. (2015) Luo et al. (2016) Malleson and Birkin (2012) Järv, Ahas, and Witlox (2014) Kung et al. (2014) Jiang (2015) Jurdak et al. (2015) Toole et al. (2015) Janzen et al. (2016) Phithakkitnukoon et al. (2017) Vanhoof et al. (2017)	Wolf, Guensler, and Bachman (2001) Huang, Li, and Yue (2010) Cho, Myers, and Leskovec (2011) Wargelin et al.(2012) Alexander et al. (2015) Gong et al. (2015) Toader et al. (2017) Yin et al. (2017)	Bohte and Maat (2009)

Literature review of rule-based activity-inference techniques from mobile-phone and GPS data

PROBLEM

Problem

1/ Conditions built for the masses

- Presence at certain times
- Frequency of presence
- Duration of presence

Problem

2/ Accuracy evaluated in a aggregated manner

Accuracy estimation	
Direct (ad hoc survey)	Schneider et al. (2013)
Aggregated (other survey)	Ahas et al. (2010) Çolak et al. (2015) Li et al. (2015) Picornell et al. (2015) Luo et al. (2016)
None	Malleson and Birkin (2012) Järv, Ahas, and Witlox (2014) Kung et al. (2014) Jiang (2015) Jurdak et al. (2015) Toole et al. (2015) Janzen et al. (2016) Phithakkitnukoon et al. (2017) Vanhoof et al. (2017)

Literature review of rule-based activity-inference techniques from mobile-phone and GPS data

Questions

1/ Do parameters induce bias?

2/ Does probability of being correctly inferred depend on social factors?

METHOD

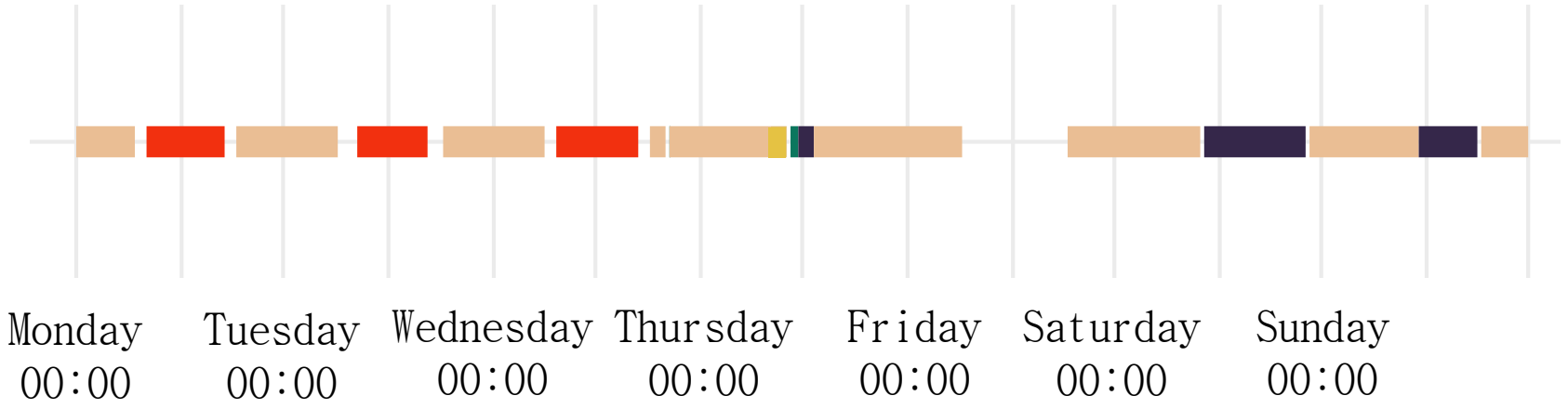
Dataset

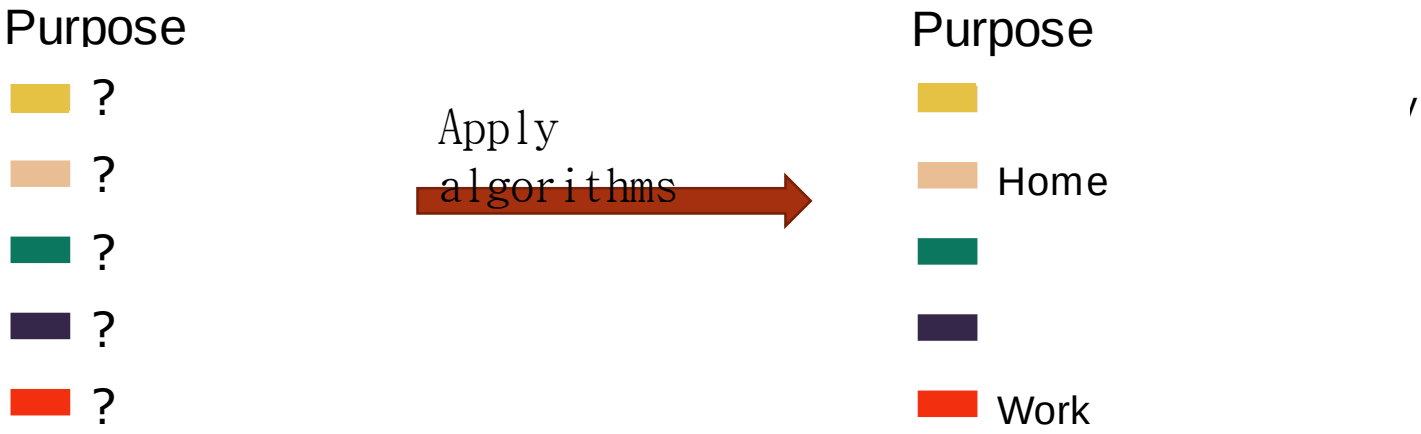
➔UK National Travel Survey (UK NTS)

- Social characterization of individuals:
  - 15,112 individuals 15 years or older
- Activity schedule over a week
- 2016

Method

Treatment





Test if accuracy depends on:

➔Temporal parameters chosen

➔ Social class

➔ Employment type

RESULTS

Test on UK NTS (1 week)

Home inference

Inferring algorithm		Correct (%)	Error (%)	No place ID'ed (%)	Total (%)
First place present at start of day (start of day = 3 a.m.)		94,4	4,2	1,4	100
Place where present most often		94,9	5,1	0	100
Most often at night during week (Monday -Thursday)	00AM-6AM	93,9	4,7	1,4	100
	00AM-9AM	94,4	4,9	0,6	100
	7PM-6AM	94,5	5,1	0,4	100
	7PM-9AM	94,7	5,1	0,2	100
Most often at night any day of the week	00AM-6AM	94,5	4,3	1,2	100
	7PM-9AM	95,3	4,6	0	100
Most often at night during week (Monday -Thursday) and all of the weekend	00AM-6AM	93,8	6,1	0,1	100
	7PM-9AM	95,1	4,9	0	100

Test on UK NTS (1 week)

Workplace inference

Inferring algorithm		People with work place inferred by algorithm		% of real active pop with workplace correctly inferred
Home is:	Place where spends most time:	% workplace correctly inferred	% who actually work	
included in potential workplaces	1PM - 5PM, weekdays	28	59	40
included in potential workplaces but exclude cases where home = work	1PM - 5PM, weekdays	78	89	39
	1PM - 5PM, weekdays + >50% duration threshold	85	92	38
	1PM - 5PM, weekdays, min 3 days of presence	83	92	38
excluded from potential work places	1PM - 5PM, weekdays	39	60	60
	1PM - 5PM, weekdays, minimum 3 days of presence	68	81	45

Workplace inference: distribution of accuracy

Inferring algorithm		% of occupation class correctly inferred		% correctly inferred: job type	
Home is:	Place where spends most time:	Partly skilled	Managerial and technical	Part time	Full time
included in potential workplaces but exclude cases where home = work	1PM - 5PM, weekdays	32	53	21	55
	1PM - 5PM, weekdays + >50% duration threshold	26	44	11	47
	1PM - 5PM, weekdays, min 3 days of presence	31	50	18	53
excluded from potential work places	1PM - 5PM, weekdays	51	64	41	67
	1PM - 5PM, weekdays, minimum 3 days of presence	45	55	30	60

CONCLUSION

Conclusion

Home inference ≠ work inference

Need tackle problem of excluding incomplete data

Thank you for your attention

Algorithms

Home inference

- First place present at start of day
- Place where spends most time
  - Anytime
  - On certain time slots, on any day
  - On certain time slots, only weekdays
  - On certain time slots on weekdays, all of weekend

Work inference

- Can be identical to home:
  - Yes
  - No
- Place where spent most time on certain days at certain times
- Where is a minimum number of days
- Where is a minimum number of time