Where did you come from? Where did you go? Robust policy relevant evidence from mobile network big data

Danaja Maldeniya, Amal Kumarage, Sriganesh Lokanathan,
Gabriel Kreindler, Kaushalya Madhawa

CPR*south* 2015
Taipei, Taiwan
26 August 2015

LIRNEasia
Pro-poor. Pro-market.

This work was carried out with the aid of a grant from the International Development Research Centre, Canada and the Department for International Development UK.

IDRC CRDI Canada

UKaid from the British people

Policy implication

- Mobile Network Big Data (MNBD) can support urban transport planning as a continuous exercise
  - Greater spatio-temporal detail than corresponding traditional output
  - Negligible incremental cost of generating forecasts
  - Single source for understanding different aspects of mobility
- Inherent limitations mean MNBD cannot replace traditional process entirely

Transport forecasting in developing countries are based on infrequent surveys

- These are done on an as needed basis
  - The Colombo transport master plan (COMTRANS) survey in 2013 in Sri Lanka, funded by JAICA
- Expensive
  - COMTRANS 2013 survey cost approximately $400,000[1]
- Time consuming
  - By the time the results are ready, they are often already outdated
- Cannot support continuous monitoring of transport patterns
- Not very useful to to help evaluate impact of policies

[1] Estimated based on interviews

Mobility insights from MNBD need to be aligned with the traditional process

- MNBD based insights provide greater temporal and spatial resolution
- MNBD cannot replace the diverse data collected by the traditional survey
- MNBD insights need to be aligned with different stages of traditional process familiar to planners

The data: historical and anonymized Call Detail Records (CDRs) from Sri Lanka

- Call Detail Record (CDR):
  - Records of all calls made and received by a person created mainly for the purposes of billing
  - Similar records exist for all SMS-es sent and received as well as for all Internet sessions

Calling Party Number	Called Party Number	Caller Cell ID	Call Time	Call Duration
A24BC1571X	B321SG141X	3134	13-04-2013 17:42:14	00:03:35

  - The Cell ID in turn has a lat-long position associated with it
- CDR data for 13 contiguous months in 2012-2013
  - Nearly million 10 SIMs
  - Over 25 billion records

Origin – Destination Matrices

- Key intermediate output of the traditional forecasting process
- Estimated people/vehicle flows between regions
  - E.g.: DSD (3rd level administrative division) level O-D matrix for the Western Province of Sri Lanka
- Used in traffic studies, identifying key transport corridors, etc.

	Agalawatta	Attanagalla	Bandaragama	Beruwala	Biyagama	Bulathsinhala	Colombo	Dehiwala	Divulapitiya	Dodangoda
Agalawatta	11013	4	48	338	4	7750	167	56	0	1722
Attanagalla	2	177073	9	31	757	7	3085	64	936	4
Bandaragama	33	8	77516	212	91	255	1392	509	8	280
Beruwala	242	34	229	206834	16	113	1178	392	11	12246
Biyagama	2	836	80	12	199621	19	9866	337	96	12
Bulathsinhala	5477	8	280	114	27	34435	128	27	11	2350
Colombo	90	2769	1040	891	6150	80	605077	9451	646	354
Dehiwala	14	35	235	135	175	12	7684	73180	11	42
Divulapitiya	0	977	4	9	84	9	646	24	119516	2
Dodangoda	1159	5	294	11801	13	2300	407	102	3	32690

Multiple methods exist for extracting O-D matrices from MNBD

- We extracted O-D matrices for the Western Province of Sri Lanka using 3 methods:
  - Stay based method
  - Transient Trip method
  - Frequent Trip method
- Methods have two stages:
  - Identify individual movement as a sequence of trips (varies across methods)
  - Aggregate individual trips across the origin and destination locations of the trips (same for all methods)

Stay based approach

- Identify instances when a user has been stationary - **Stays**
  - Geographical location with the associated time period
- With CDR a Stay is contiguous series of records such that,
  - Any two records in the series are less than a distance D apart, where $D = 1km$
  - The entire series of records should span a period of more than 10 minutes
  - Two contiguous records are separated by a time interval T such that $T \leq 1hour$
- Each pair of consecutive stays for a person is taken as origin and destination of a trip
- Built on prior work
  - Calabrese, F., Di Lorenzo, G, Liu, L., Ratti, C. (2011). Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area.
  - Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E., & González, M. C. (2013). A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In *Proceedings of 2nd ACM SIGKDD International Workshop on Urban Computing*. Chicago, IL.

Transient trip approach

- A trip is identified from the CDRs by a consecutive pair of records such that,
  - The records indicate a displacement, i.e. the BTS-es utilized for each record is different
  - The records are separated by a time interval $T_{Interval}$ where, 10 minutes $\leq T_{Interval} \leq$ 1 hour
- Maximizes amount of extracted mobility information by capturing intermediate points in trips
- Extracted trips likely to correspond to segments of real trips
- Built on prior work
  - Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012). Understanding road usage patterns in urban areas. *Scientific Reports*, *2*, 1001. doi:10.1038/srep01001
  - Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. doi:10.1016/j.trc.2014.01.002

Frequent trip approach

- Frequent trip approach attempts to capture regular travel
  - Identify frequent sequences of two locations in the daily trajectories of a person (Frequent Sequence Mining)
  - A sequence can be non-contiguous.
  - A sequence is frequent if it occurs at least on 10% of the days a person is observed
  - A frequent sequence defines the endpoints of a frequent trip
- Estimate the likelihood of making a frequent trip during a period of a given day

$$P(Trip_i|D_j Period_k) \approx \frac{Frequency\ of\ Trip_i\ during\ D_j\ and\ Period_k}{\#\ of\ times\ i\ had\ at\ least\ 1\ record\ during\ D_j\ and\ Period_k}$$

$$k\ = \{Morning, Afternoon, Evening, Night\}$$
$$D_j\ = day\ of\ the\ week, j = \{0{,}6\}$$

- Built on prior work
  - Bayir, M. A., Demirbas, M., & Eagle, N. (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing*, *6*(4), 435–454. doi:10.1016/j.pmcj.2010.01.003

Each method has strengths/weaknesses

	Stay based	Transient trips	Frequent trips
Amount of mobility	Low	High	Regular mobility
Sensitivity to noise	Low	High	Low
Bias towards active users	High	Moderate	Low
Use	Identifies congregations of people	Suitable for short term mobility analysis	Aligns best with outputs from traditional process

Validation with traditional output

- Compared with best available traditional forecast
  - Number of trips generated by region at the DSD level, from COMTRANS 2013

- Constructed weighted linear models for all three methods

Method	Intercept	MNBD estimate	R^2
Stay based	35,516***	76.41***	0.819
Transient trip	25,460**	2.66***	0.903
Frequent Trips	14,770·	1.16***	0.909

MNBD insights within a traditional transport forecasting process

- MNBD insights have inherent limitations
  - Sampling biases: high activity users, mobile phone penetration in different regions
  - Sparsity of data: less than 25 records per day for 90% of the users
  - Lack of socioeconomic, demographic, travel motivation

- Solutions exist to mitigate these issues to some extent
  - Adjust for penetration and operator market share by scaling flows
  - Associating demographic parameters from travel surveys with MNBD insights using machine learning techniques to match mobility variables in both
  - Probabilistic models (E.g : Hidden Markov Models) to estimate locations for people where no records exist, can improve mobility estimates

Policy implication

- Mobile Network Big Data (MNBD) can support urban transport planning as a continuous exercise.
  - Greater spatio-temporal detail than corresponding traditional output
  - Negligible incremental cost of generating forecasts
  - Single source for understanding different aspects of mobility
  - Inherent limitations mean cannot replace traditional process entirely

Thank you.