

Data collection to support evolutionary models

ABM Symposium, Munich
11-13 December 2024

Aruna Sivakumar, Imperial College London

Summary: Data collection to support evolutionary models

- Brian Lee (Household Travel Survey Data Collection: Meeting the needs of planning practice and research)
 - **Data collection needs to be able to support the requirements of the planning practice such as analysis of equitability, climate change impacts etc**
 - For e.g., ensuring sufficient representation of minority populations and their travel patterns, in order that equitability questions may be addressed
 - All too often important variables are dropped out of models because there isn't enough data to estimate them reliably or lack of suitable 'weights' to ensure representativeness...
- Greg Erhardt (The Potential for Linked Longitudinal Data in Transportation Research)
 - **Evolutionary models need longitudinal panel data**
 - Case study of extracting panel data from the American Community Survey (ACS)
 - Are elasticities estimated from cross-sectional data an over-estimate?

Innovations in/Practice of... data collection

- Mobile network data
- Public transport fare card data
- Open data of PT and shared bike operations
- Enhanced surveys e.g. life-course surveys
- Mobile app-based surveys
- ... data for understanding specific behaviours vs capturing all relevant behaviours
(**CREDIBILITY** of data? Depends on purpose?)

In parallel, national statistics agencies and transport planning organisations continue to collect large (quite rich) datasets from repeated cross-sectional surveys (e.g. UK NTS continuous/monthly since 1988 with approx. 16-20k households per year since 2002; LTDS continuous since 2005)

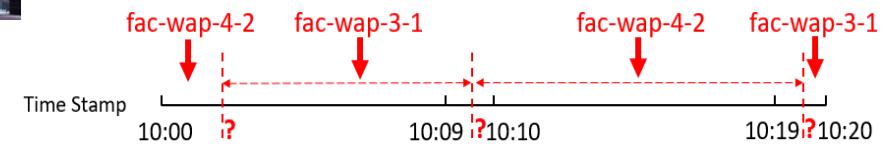
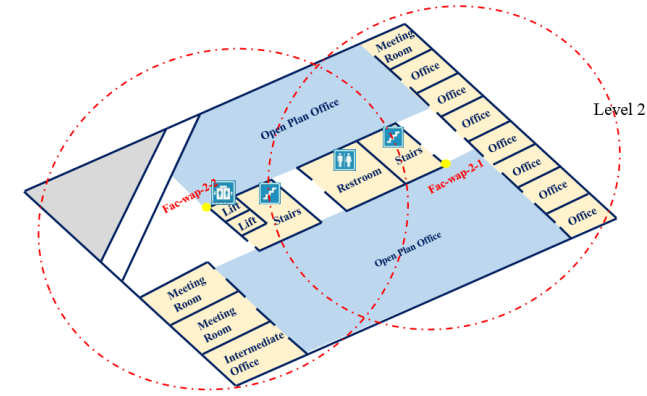
Challenges with emerging data sources and implications for credibility

- Privacy and security concerns
 - Regulatory environment
- Pre-processed data (e.g. mobile phone data)
 - Implications for bias can be complex
- Wide range of data standards and formats need to be reconciled
 - GIS-T, GTFS, Open data initiatives... development of data exchange standards
- Data gaps
 - not all private services make data available
 - missing data and poor quality of data
- Degree of semantic content: 'thick' and 'thin' data; Lack of qualitative insight
- Data driven vs theory driven analysis – reliability, validation

Fusing passive data: cross-sectoral contexts

Integrated modelling of building occupancy & urban systems

- Challenge:
 - Buildings have major footprint on urban systems: consumer of energy, driver of transport demand, economic opportunities
 - Building occupancy modelling has so far remained largely done in isolation from dynamics of the surrounding systems
- The project looks at opportunistic data from a variety of systems within the building (Wi-Fi, entry, BMS, HVAC) and beyond, e.g. weather, transport API



Time	AP
10:00	fac-wap-4-2
10:10	fac-wap-3-1
10:10	fac-wap-4-2
10:20	fac-wap-4-2
10:20	fac-wap-3-1

- Case study: Imperial College Faculty Building
 - **Use of Wi-Fi logs data, translated into occupancy data**
 - Hazard-based approach
 - Impact of facilities, time of day and weather on how people move within and depart from the building

Privacy-preserving big data enrichment

- Challenge:
 - Ever-growing volume of (big) data
 - Low-semantic content ('thin')
- How can we 'enrich' big datasets while not infringing user privacy?
 - Use of 'small but thick' to enrich 'large but thin' data
- Focus on using fundamental principles derived from information *and* microeconomic behavioural theories

Transportation Research Part B 155 (2022) 101–134



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb



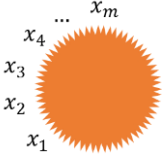



Theory for socio-demographic enrichment performance using the inverse discrete choice modelling approach

Yuanying Zhao^{a,b,*}, Jacek Pawlak^{a,b}, Aruna Sivakumar^{a,b}

^a Urban Systems Lab, Imperial College London, SW7 2AZ, UK

^b Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London, SW7 2AZ, UK

	Thick Dataset: High semantic content (e.g. travel survey)	Thin Dataset: Low semantic content (e.g. GPS data)
Overall Volume		
Each Record	x_1 x_2 x_3 x_4 ... x_m 	y_1 y_2 y_3 y_4 ... y_n 
Remarks	m, n : number of variables contained in each data record ($m \gg n$); $\{x_1, x_2, \dots, x_m\}, \{y_1, y_2, \dots, y_n\}$: variables contained in each data record.	

A typical research project: Project ITINIERANT

- An interdisciplinary framework for disaggregate assessment of **productivity** and **well-being** impacts of **digital technologies** on **knowledge workers in non-traditional settings**: Project ITINIERANT
- Challenge:
 - Lots of anecdotal and qualitative evidence concerning the role of technologies in impacting productivity and well-being
 - Investment appraisal and policy-making requires a suitable modelling framework, supported by empirical evidence from larger-scale data



- **An approach that combines use of a variety of data:**
 - **Secondary large-scale survey data:**
 - A task-based approach to propensity of undertake work in non-traditional settings (a combination of tasks as a 'genome' of particular occupations)
 - Which tasks associated with particular occupations make them more likely to work when travelling or from cafes, public spaces?
 - What role does the technology play?
 - **Interview data:**
 - What is meant by 'being productive'?
 - Do conventional metrics of productivity align with people's perception?
 - **Primary survey data:**
 - Dedicated modelling effort to quantify the interview insights

What do our evolutionary (yet stable) models need?

- Life course information, including major life events
- Multiple weeks of activity-travel patterns (1-week at least, but more to understand intrapersonal heterogeneity)
- That include details of time use (physical and digital activities) and spatial travel patterns
- Lifestyle factors and preferences
- Data from a wide range of population segments, including minorities
- ... ?? discuss

We rarely manage to combine all these and do justice, all the more so in cross-sectoral contexts. More proactively adopt and improve data fusion techniques in order to combine several data sets? Pseudo-panel methods with choice models? Bayesian techniques?

Ensure that when multiple datasets are used (almost always) they are credibly combined