Research on Data Mining and Complex Networks Meets Transportation Research: New Approach to Existing Challenges in Travel Behavior Analysis

13.30-14.10 Existing challenges in travel behavior analysis and modeling solved from the perspective of large datasets: a take-off in the DATASIM project, **Prof. dr. Davy Janssens**

14.10-14.50 Unveiling the complexity of human mobility by querying and mining massive trajectory data, **Dr. Fosca Giannotti**

14.50-15.10 BREAK

15.10-15.50 Spatial Data Mining in Practice: Principles and Case Studies, **Dr. Michael May**

15.50-16.30 *A universal model for mobility and migration patterns*, **Dr. Filippo Simini**
Existing challenges in travel behavior and modelling, seen from the perspective of large datasets: a take-off in the DATASIM FP 7 project

Prof. dr. Davy Janssens, davy.janssens@uhasselt.be
Context and background of this workshop/presentation
Why transportation modelling?

- Transportation problem is multi-dimensional:
  - Traffic jams, CO₂-emissions, Impact on economy, Traffic accidents,…

- Transportation models
  - Support ex ante management decision making (What if?)
  - Make predictions in uncertain circumstances:
    - Changing infrastructure, environment
    - Changing behaviour of people
    - Changing socio-demographic circumstances
    - …

- Assess impact of policies/ other domains
  - Mobility
  - Traffic safety
  - Environmental impact (emissions)
  - Exposure (human health)
  - Electrical vehicles
Transportation models

Some scenario’s almost cannot be calculated by means of trip and tour-based models.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Decision units</th>
<th>Trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip-based</td>
<td>1 HBW-trip</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>3 NHBW-trips</td>
<td>B,C,D</td>
</tr>
<tr>
<td></td>
<td>3 HBO-trips</td>
<td>E,F,G</td>
</tr>
<tr>
<td>Tour-based</td>
<td>1 HBW-tour</td>
<td>A,D,E</td>
</tr>
<tr>
<td></td>
<td>1 WB-tour</td>
<td>B,C</td>
</tr>
<tr>
<td></td>
<td>1 HBO-tour</td>
<td>F,G</td>
</tr>
<tr>
<td>Activity-based</td>
<td>1 Activity agenda</td>
<td>A,B,C,D,E,F,G</td>
</tr>
</tbody>
</table>
Activity-based models

“Travel demand derived from activities”

- Sequences of episodes
  - Activities
  - Travel
- Account for constraints
  - Household interactions
  - Time-space constraint
Activity-based models

- Simulate
  - Which activities?
  - Where?
  - When?
  - How long?
  - With whom?
  - Transport mode?

"Travel demand derived from activities"
Research challenges in travel behaviour analysis and modelling
Challenges in travel behaviour analysis and modelling

- **Challenge 1: A problem of data**
  - Travel surveys and activity diaries:
    - burden for respondent
    - very expensive
    - time lag between data collection and data entry
    - spatial component (granularity, level of detail)
    - temporal component ("rounding" issues)
    - Response rates often low
Challenges in travel behaviour analysis and modelling

- **Challenge 2: Efficient model validation**
  - Currently model validation only done at the final step in the modelling phase (comparing with traffic counts)
    - E.g. no intermediary validation at the level of origin-destination matrices
    - Difficult to see how errors propagate in the model
  - Furthermore: in traditional modelling, OD matrices are estimated from the perspective of pre-defined zones (corresponding to administrative districts, communities etc), while fuzzy (non-administrative but data driven) OD matrices may reveal important and more valueable information
Challenges in travel behaviour analysis and modelling

- **Challenge 3: A problem of currently existing AB models**
  - Operationalisation of AB models!
  - Real-world applications of AB models (increasing though) while in theory there are plenty of opportunities
  - Furthermore:
    - Scalability of models is an issue!
    - In traditional straightforward (non agent-based) applications, behavioural sensitivity in the AB model which was present during demand estimation is lost in the assignment step
Big datasets as a possible solution?: The DATA SIM Approach
DATA SIM

- Data Science for Simulating the Era of Electric Vehicles (DATA SIM): an open-FET FP7 EU project
  - FET-Open supports the exploration of new and alternative ideas that, because of their risky or non-conventional nature, would not be supported elsewhere.
  - Only highly ambitious projects are accepted, low acceptance rate (~5%)
## Participants in DATA SIM

<table>
<thead>
<tr>
<th>Participants</th>
<th>Expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>UH-IMOB (Belgium)</td>
<td>Transportation modelling, traffic safety, and logistics</td>
</tr>
<tr>
<td>CNR-ISTI (Italy)</td>
<td>Knowledge discovery and data mining</td>
</tr>
<tr>
<td>BME-TTK (Hungary)</td>
<td>Network science applied in economics, computer science, sociology and biology</td>
</tr>
<tr>
<td>FRAUNHOFER (Germany)</td>
<td>Data mining, statistical relational learning, geographic information systems, and distributed computing</td>
</tr>
<tr>
<td>UPM (Spain)</td>
<td>Electric power systems, control engineering, chemistry and topography, geographical information systems</td>
</tr>
<tr>
<td>VITO (Belgium)</td>
<td>Industrial innovation, energy, and quality of the environment;</td>
</tr>
<tr>
<td>TECHNION HAIFA University (Israel)</td>
<td>Distributed knowledge discovery, large-scale distributed data mining and data streams</td>
</tr>
<tr>
<td>UPRC (Greece)</td>
<td>Database management, knowledge discovery and data mining</td>
</tr>
</tbody>
</table>
Challenge 1: a problem of data (diaries)

- Solution:
  - GPS data are in theory a possible solution to this problem (!)
  - However: there is a long way to go from raw data of individual trajectories up to high level collective mobility knowledge:
    - Only detailed spatio and temporal resolutions are covered
    - Annotation (activity, transport mode) is necessary
      → A lot has been done in data mining and informatics in the past!
    - Behavioural reflection of annotation process often not taken into account
      → Annotation of vehicle trajectories: translate to individuals’ trajectories: account for e.g. Hägerstrands’ constraints
Challenge 1: a problem of data (diaries)

- Mobile phone data:
  - Spatio-temporal information also available but less detailed (call data records, handover data)
  - Are Massive datasets
  - Covered in the area of Network Science
State-of-the-art

- The latest developments in the research of human mobility include:
  - **Mobility data mining**: extending traditional data mining techniques to location sequences of individuals’ movement for pattern mining, clustering and location prediction (*M-Atlas*);
  - **Statistical physics of human mobility**: uncovering statistical laws that govern the key dimensions of human travels, e.g. travel distance and activity duration;
  - **Semantic-enrichment of mobility data**: inferring semantic and context aspects of travel behavior; annotation process
  - **Social network analysis**: investigating the dynamics of social network to characterize mobility behaviours of subpopulations based on their social relations.

- Combine data mining and statistical physics into a uniform analytical framework, able to develop macro-micro models of human mobility with an unprecedented explanatory and predictive power

- Combine mobility patterns with social networks to explore how mobility patterns depend on demographic factors and social network characteristics.

- Extend mobility patterns with semantics to explain the purpose of people’s whereabouts.
Goal: try to combine micro- and macro-laws of human mobility and use them in AB-modelling!

- E.g. location choice model can be developed from GPS/GSM traces (see also presentation by dr. Simini)
- When fully annotated, travel diaries can (in theory) be reconstructed (from GPS traces), and can be fed into AB-models
Implementation of the solution

- Merge the **raw and behaviorally poor big** data with the **smaller but behaviorally richer** travel survey data, building a novel agent-based reality mining modeling standard of mobility behavior.

  - Modelling anchor points from big data (home, work, school)
  - Modelling stop frequencies
  - A detailed modeling level for departure time of activities, trips and tours;
  - Destination choice models built from individual accessibility point-of-view, under potential path areas that can be reached within certain spatio-temporal constraints of the agent, derived from the annotated trajectories developed in WP2;
  - Merging the person-based accessibility measures with socio-demographic information.

- Sensitive towards a broad range of behavioral changes, accounting for the impact of policy measures and trends, especially the impact of different scenarios in the era of electric vehicles.
Simulation process demonstration

Home (Departure time; Travel mode)

Starting time
Activity duration
Activity type
Activity Location
Travel mode
Travel companion

Work

Travel mode
Travel distance
Activity sequence
More

Home

Number of trips a day

Stop

Stop

Stop
The final outcome

Big data integrated with activity-travel diaries and social networking data, providing information on:
- Activity location, starting time and duration
- Activity type, travel mode and travel companion (inferred)
- Socio-demographic information (inferred)
Challenge 2: efficient model validation

- Massive mobility data trace people transfer phenomenons, providing direct and objective measures for the validation process.
  - **GPS**: providing movement trajectories in precise spatial locations and a high time rate; but covering small subsets of a population, and related to vehicles rather than individuals;
  - **GSM**: tracking individuals’ movement and covering a significant segment of population; but requiring additional efforts from telecom operators and lacking details in spatial and temporal resolution.

- Overcome the limitation of single data source, combine heterogeneous data types, and yield reliable model validation methodologies applicable to large-scale domains.
- Start from hand-over data to estimate OD Matrix

![Figure 9: GSM handover data issues. (a) resulting origin-destination matrix lays around diagonal, and diagonal is empty; (b) long trips translate to several short ones; (c) aggregation to zones; (d) uncertainty in the geometry of cells](image)
Challenge 3: A problem of currently existing AB models

- Look for applications which shows the real value of an operational model
- Previous research in the past (Beckx et al.)
- In this project: application domain of electric vehicles:
  - Nationwide power demand estimation (using detailed charging schemes)
  - Dynamic electricity pricing scenario’s (prices ↑ when electricity demand ↑, prices ↓ when electricity demand ↓)
Some preliminary research results of DATA SIM

Challenge 1: Activity-diaries
Some preliminary research results of DATA SIM (challenge 1)

- **Semantic annotation of GPS traces** (topic related to challenge 1)
  
  = automatic inference of activity types *only* based on activity start times and activity durations

- Developing an expert system that links GPS trajectories to corresponding diary data based on two models:
  - Predicted probability distribution
  - Point predictions
Challenge 1: Data

- Development of a GPS-enabled Personal Digital Assistant (PDA) (PARROTS: PDA system for Activity Registration and Recording of Travel Scheduling)
- Data collected for 2400 Households, about 900 based on PDA, 1500 paper-and-pencil diaries
- 1 week activity-based survey
- Capturing route information
- Planning activities for the future
- Enabler for future research related activities (see later)
Challenge 1: Data

- Activity attributes:
  - Which type of activity?
  - When?
  - For how long?
  - Where?
  - With whom?

- Trip attributes:
  - When?
  - With whom
  - Route?
  - Transport mode

+ consistency checks
Challenge 1: Data

- Fraction GPS logs with GPS information as a function of time-of-day
  - Day: fraction increases → logging of trips
  - Night: fraction decreases → charging
  - Evening: fraction decreases → charging & imputation
Challenge 1: Data

- Training, test and validation set
  - Diary data after cleaning: N=11906
  - GPS data with matching diary data: N=342
    - Validation data
  - Training data: 75%
    - N=8930
  - Test data: 25%
    - N=2976
Challenge 1: Processing/cleaning of data

- Deviation in time registration:
  - Strong mismatch due to
    - Incomplete schedules!
    - Trip end identification
    - GPS burn in, battery instability, incorrect diary reporting/use of GPS device
  - Small deviation for most trips
    - Due to rounding errors (min), burn in problems
  - 5% large deviations (> 1 h, > 2 h)
    - Due to inefficiency during process of trip end identification?
    - Rounding errors at hour level?

- Perfect match
  - Trip start times: 245 trips
  - Trip end times: 311 trips
  - Trip durations: 421 trips
Challenge 1: Research methodology

- Classification of activity start times and activity durations
  - Decision Tree (C4.5)
  - Optimizing classification by considering minimum class frequencies and by creating balance between number of leaf nodes and impurity
- Start times: 13 categories
- Durations: 8 categories
Challenge 1: Research methodology

- Predicting probability matrix
  - Probability distribution model
  - Distribution of activity probabilities for each class of activity start time and activity duration

- Extracting majority matrix
  - Point prediction model
  - Highest probability for each class
Challenge 1: Results

<table>
<thead>
<tr>
<th>Data set</th>
<th>N</th>
<th>Misclassification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>8930</td>
<td>29.50%</td>
</tr>
<tr>
<td>Test data</td>
<td>2976</td>
<td>27.76%</td>
</tr>
<tr>
<td>Validation data</td>
<td>342</td>
<td>25.44%</td>
</tr>
</tbody>
</table>
## Challenge 1: Results

### Probability Distribution

<table>
<thead>
<tr>
<th>Activity duration &lt;= 18min AND Activity Start time &lt;= 8:56 am</th>
<th>Activity duration &lt;= 18min AND Start Time between 8:56 am and 11:55 am</th>
</tr>
</thead>
<tbody>
<tr>
<td>P (Bring-get) = 66.2%</td>
<td>P (Bring-get) = 22%</td>
</tr>
<tr>
<td>P (Home) = 12.4%</td>
<td>P (Home) = 7.9%</td>
</tr>
<tr>
<td>P (Leisure) = 0.6%</td>
<td>P (Leisure) = 2.9%</td>
</tr>
<tr>
<td>P (Services) = 1.9%</td>
<td>P (Services) = 15.9%</td>
</tr>
<tr>
<td>P (Shopping) = 14.6%</td>
<td>P (Shopping) = 43.3%</td>
</tr>
<tr>
<td>P (Social) = 1.3%</td>
<td>P (Social) = 4.7%</td>
</tr>
<tr>
<td>P (Touring) = 0%</td>
<td>P (Touring) = 0.4%</td>
</tr>
<tr>
<td>P (Working) = 2.9%</td>
<td>P (Working) = 2.9%</td>
</tr>
</tbody>
</table>
**Challenge 1: Results**

*point prediction*

<table>
<thead>
<tr>
<th>MAJORITIES</th>
<th>Activity duration</th>
<th>Activity starting time</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;= 18 min</td>
<td>&lt;= 2u10</td>
<td>2:10 - 8:30</td>
</tr>
<tr>
<td>18 min - 55 min</td>
<td>8:30 - 8:56</td>
<td>8:56 - 9:30</td>
</tr>
<tr>
<td>55 min - 157 min (2u37)</td>
<td>9:30 - 11:55</td>
<td>11:55 - 12:50</td>
</tr>
<tr>
<td>157 min - 159 min (2u39)</td>
<td>12:50 - 14:30</td>
<td>12:50 - 15:40</td>
</tr>
<tr>
<td>159 min - 180 min (3u)</td>
<td>14:30 - 16:40</td>
<td>14:30 - 16:40</td>
</tr>
<tr>
<td>180 min - 550 min (9u10)</td>
<td>16:40 - 18:40</td>
<td>16:40 - 18:40</td>
</tr>
<tr>
<td>&gt; 550 min (9u10)</td>
<td>18:40 - 20:55</td>
<td>18:40 - 20:55</td>
</tr>
<tr>
<td>&gt; 23 min</td>
<td>&gt; 23:00</td>
<td>&gt; 23:00</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
<tr>
<td>home</td>
<td>home</td>
<td>home</td>
</tr>
</tbody>
</table>

*Note: The table shows the distribution of activities based on duration and starting times.*
Challenge 1: Conclusion

- Using these models (and further improvements)
  - To enrich GPS logs with diary variables
  - To be able to use GPS data collection without accompanying diaries (survey questionnaires)

- Contributions to scientific state-of-the-art
  - Pure time annotation
  - Deviation in time registration between both data sources
  - Indication of impact of time on activities
    - Important role in the semantic enrichment process
Some preliminary research results of DATA SIM

Challenge 3: Operationalisation of nationwide activity-based models and application scenario in EV
(by L. Knapen et al. TRB 2012)
Some preliminary research results of DATA SIM (challenge 3)

- The Feathers framework
- Forecasting Evolutionary Activity-Travel of Households and their Environmental RepercussionS (FEATHERS)
- AB simulation research laboratory
- Platform for model development
  - Prototyping
  - Experimenting
  - Simulating
- Configurable for different study areas
- Recuperation of research / implementation efforts
The FEATHERS output

- Activity-travel schedules
  - Most detailed form
  - Aggregation to extract targeted information

- Analysis of personal transportation behavior
  - Usage of means of transportation
    *E.g. Usage of public transportation, carpooling, …*
  - Detailed segmentation
    *E.g. motive: Home-work/school transportation*
The FEATHERS output

- Analysis of tours and trips
  - Frequencies
  - Modal choice
  - Distance and/or time distribution
  - Detailed segmentation based on socio-demographics
    *E.g. Modal split for 65+

Relative amount of trips leaving each municipality in Flanders (Belgium)
The FEATHERS output

- Traffic assignment
  - Vehicle flows
  - Volume to network capacity

Trips for all activities in Flanders (Belgium)

Trips for work activity in Brussel
The FEATHERS output

- Analysis of Activities
  - Type
  - Location
  - Duration
  - Time of day

Quotient work status in Flanders (Belgium)

Person-hours spend from all origin zones in Hasselt

57 - 60
61 - 63
64 - 65
66 - 67
68 - 74

HASSELT
The FEATHERS output

Relative amount of Work Activities per municipality in Flanders
Application domain

- Yearly electric energy consumption
  - For Belgian Household estimated at 3500..3900kWh
  - For EV (15000Km, 200Wh/km): 3000 kWh
- Potential problem if everyone would charge its car at the same time
- Electric Vehicle Scenario
  - Calculate
    - EV energy requirement (based on distance driven)
    - Available charging period (deadline)
    - Location
Application domain

- Charging scenarios
  - Everyone uses timer to start when lowTariffPeriod starts: peak
  - Uniform distribution within lowTariffPeriod: Feasible
  - Immediate charging after last home arrival: Lower peak

- Parameters for scenarios
  - Market share EV: BEV, PHEV
  - Car class: small, medium, large: similar to ICEV cars
  - Charger types
  - Range anxiety coefficient
  - Privately owned vs. company cars
  - Work location charging (cheaper than fuel card)
Application domain

POC - CC / Work Location Charging

- Car Users
  - NoWorkTrip
  - WorkTrip
    - POC
      - can charge at work
    - CC
      - cannot charge at work
Application domain

\[ \forall i, j \in [1, \#L] : C_b - d_{O,i} \cdot c_s + \sum_{j=1}^{j<i} t_j \cdot p_j \geq C_b \cdot k_{dcd} \]  

(1)

\[ \forall i, j \in [1, \#L] : C_b - d_{O,i} \cdot c_s + \sum_{j=1}^{j\leq i} t_j \cdot p_j \leq C_b \]  

(2)

- \( i, j \): location indices
- \( C_b \): battery capacity
- \( L \): the set of all locations used in the schedule
- \( t_j \): charge-period duration at the j-th location
- \( p_j \): power at location j
- \( d_{O,i} \): total distance from the first origin to the i-th destination
- \( c_s \): distance specific energy consumption
- \( k_{dcd} \): deep charge depletion coefficient
Results

Trip distance distribution

- About 78% of trips predicted for Flanders are BEV-feasible.
- Prediction takes small amount (5% of employees) of work location charging into account.
Results
Electric Vehicle Scenario

- Power requirement: All EV are PHEV
  - Scenarios:
    - UniformLowCost
    - LastHomeArrival
    - EachHomeArrival

Electric power demand during low tariff period for a working day
Questions?

Thanks for your attention!

Transportation Research Institute (IMOB)
Hasselt University
Wetenschapspark 5 bus 6
3590 Diepenbeek – Belgium

Tel.: +32 (0)11 26 91 28
E-mail: davy.janssens@uhasselt.be
URL: www.imob.uhasselt.be

Also visit the DATA SIM website: www.datasim-fp7.eu