A generic data-driven sequential clustering algorithm determining activity skeletons

Wim Ectors, Bruno Kochan, Luk Knapen, Davy Janssens, Tom Bellemans

ANT 2016
Table of contents

- Motivation and introduction
- Methodology
  - Data description
  - Sequential clustering algorithm
  - Sensitivity analysis
- Results
- Discussion and conclusion
Motivation and introduction

- ABM: need for transportation as derived demand from people’s activity patterns
  - Mandatory (inflexible) activities scheduled before more flexible activities
  - Conventional mandatory activities: work & education
- HTS Flanders, Belgium (OVG):
  - Only 45% contains a ‘mandatory’ activity
  - No structure in other 55%?

Data-driven approach to reveal the real basic structure of individuals’ schedules: skeleton schedule
Methodology – Data description

- HTS of Flanders, Belgium
  - Single-day, including weekends
  - Only out-of-home activities
  - 17,300 individuals
    - 13,200 at least one trip
- Weights
  - 14 (of 2600 different) most frequent day-long schedules:
    - 45% of observations (each other pattern <1%)
      - 55% more complex behavior → skeleton schedules??
- Pre-processing
  - Consecutive activities merged
Methodology – Sequential clustering algorithm

- Main idea:
  - Find common activity patterns in otherwise highly heterogeneous activity schedules
  - $\Rightarrow H-S-H-X-H$ ?
  - Optimization of location $X$ ?

<table>
<thead>
<tr>
<th>H</th>
<th>Home</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>Shopping</td>
</tr>
<tr>
<td>R</td>
<td>Recreation</td>
</tr>
<tr>
<td>Se</td>
<td>Services</td>
</tr>
<tr>
<td>X</td>
<td>‘Wildcard’</td>
</tr>
</tbody>
</table>
Methodology – Sequential clustering algorithm

**Pre-processing**
- Input survey data
- Conformity check data
- Extract list of single-day schedules
- Generate all possible wildcard-containing schedules for each given schedule

**Sequential clustering**
- Calculate frequencies
- Select wildcard-containing schedule with highest frequency
- Assign to compliant given schedule and move to output dataset
- Max # iterations reached or all given schedules assigned?
  - Yes
  - No

**Post-processing**
- Remove outliers
- Generate output tables and figures
Methodology – Overview of the research

OVG

Sequential clustering [3 steps]

Skeletons

ID3 DT training

DT

Prediction by DT

Predicted Skeleton

Settings

Obtaining Skeletons

Socio-demo extract

Socio-demo

Regression

Regression coefficients

CMA

Compare
Methodology – Sequential clustering algorithm

- Pre-processing
  - Cleaning
    - Remove schedules with >x activities?
  - ∀ schedules: find all possible wildcard-containing schedules according to settings:
    - Minimum # activities not replaced by X?
    - H cannot become X?
    - W cannot become X?
    - Merge consecutive X?

\[ N = \sum_{r=s}^{n} \frac{n!}{r! (n - r)!} \]

- Sequential clustering
  - determine the largest groups of unique wildcard-containing patterns
Methodology – Sequential clustering algorithm

- Post-processing
  - Exclude *odd patterns* ("outliers")
  - ⑥ Cutoff after cum. freq. of x %

![Diagram](image-url)

*Outliers according to wildcard-containing schedule frequency, based on OVG 3.0-4.5*

*Wildcard-containing schedules (from left to right: most frequently to least frequently occurring). Only a few labels are shown on the axis.*
Methodology – Sensitivity analysis

- Effect of ①, ②, ③, ④, ⑤, ⑥ ...?
- Ultimate goal: predictions
  - Use DTs as in ABMs such as FEATHERS, ALBATROSS
- Two stages
  1. Generate many sets of skeletons with different setting combinations
    - 2520 sets were generated
  2. Use ID3 algorithm to train DT and estimate accuracy of skeleton classification
    - ⑦ minimum number of records in a leaf?
    - ± 44,000 DTs fitted
    - Training (75%) and test set (25%) CMAs
Methodology – Overview of the research

- OVG
  - Settings
  - Sequential clustering [3 steps]
  - Skeletons
  - ID3 DT training

- Socio-demo extract
  - Regression
  - CMA
  - Regression coefficients

- DT
  - Predicted Skeleton
  - Prediction by DT
  - Compare

Obtaining Skeletons
Methodology – Sensitivity analysis

- Influence of ①, ②, ③, ④, ⑤, ⑥, ⑦ on classification accuracy?
- Analyzed in regression model (adj. $R^2$ 0.82)
  - Minimum # activities not replaced by X: inversely correlated
  - Cutoff after cum. freq. of x %: inversely correlated
  - Remove schedules with >x activities from input dataset: Marginal effect on CMA
  - H cannot become X: marginal negative effect
- ‘Practical optimum’ set of settings yields test set CMA of 32% (↔ null model accuracy 13.3%)
Results

- **Two runs**

  1. ① Minimum # activities not replaced by $X = 3$
     - 733 skeletons from 2,600 schedules
  2. ① Minimum # activities not replaced by $X = 2$
     - 341 skeletons from 2,600 schedules
     - 14 skeletons = 70% of all records (↔ 45% in original data)
Discussion and conclusion

- Only single-day data is limitation
- Temporal component not accounted for
- Number of trips affected by merging of consecutive X

Yet:
- Activity-distribution in X quite complex; common travel behavior extracted
- Algorithm universal and simple
- Data driven
- Compatible with current ABM approaches
Thank you

Wim Ectors*
Bruno Kochan
Luk Knapen
Davy Janssens
Tom Bellemans

*wim.ectors@uhasselt.be