Optimizing copious activity type classes based on classification accuracy and entropy retention

Wim Ectors[1], Sofie Reumers[1], Won Do Lee[2], Bruno Kochan[1], Davy Janssens[1], Tom Bellemans[1], Geert Wets[1]

[1]Hasselt University, Transportation Research Institute (IMOB) - [2]Manchester Metropolitan University, Crime and Well-being Big Data Centre

Abbreviations

ATC: Activity Type Class
HTS: Household Travel Survey

Intro

(Big) travel data: ✓ large amounts, real time, temporally and spatially referenced data
× personal and activity-travel info are lacking!
→ Behavioral data mining techniques are used to *infer* the activity type (=trip purpose)

Classification accuracy strongly depends on the number of ATCs; different classification approaches exist:

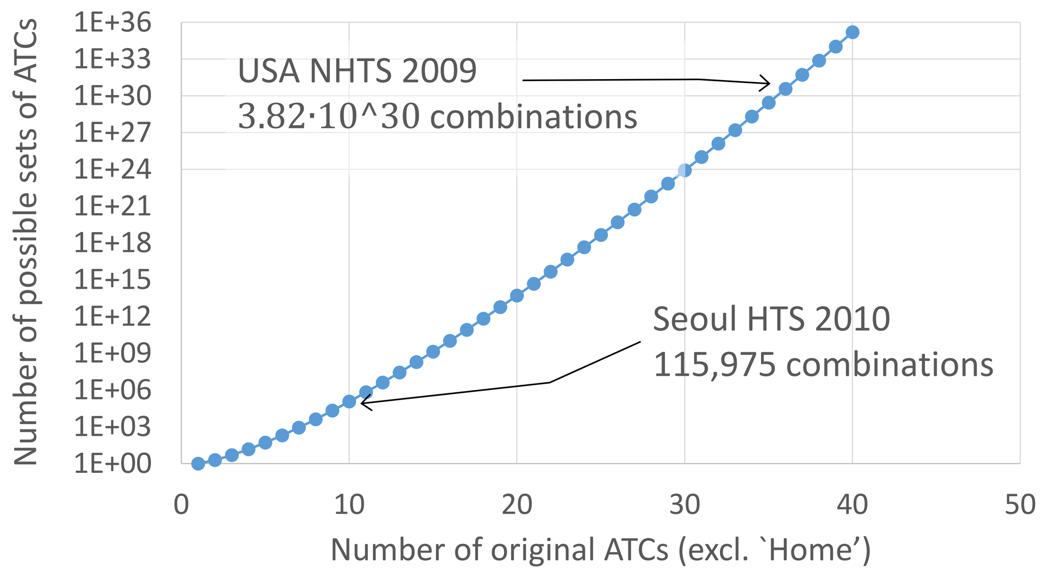
- Some predict many distinct ATCs → rich prediction, but bad classification accuracy
- Others predict few distinct ATCs → unsatisfactory prediction information, but good prediction accuracy

Ectors et al. (2017): previous studies do not provide a strong justification for the choice of ATCs.
Often, ATCs are aggregated to enhance activity inference, but this destroys activity information.

What is the optimal set of ATCs? Optimal balance between:

- Improving inference accuracy by aggregating (grouping) ATCs
- Preserving activity information from the original data (keeping as disaggregated set of ATCs as possible)

Ectors et al. (2017): find optimal set of ATCs *by creating all possible sets first* and then finding the optimum
However, the number of possible sets of ATCs increases exponentially with the number of original ATCs:



→ **This approach is impossible** for i.a. USA NHTS 2009 with 36 original ATCs (*'home'* excluded) because $3.82 \cdot 10^{30}$ aggregation combinations exist (an estimated $\sim 1.13 \cdot 10^{23}$ years of computation time on a high-end server)

→ This research proposes a *local search* algorithm to determine the optimal set of ATCs.

Ectors, W., S. Reumers, W. Do Lee, K. Choi, B. Kochan, D. Janssens, T. Bellemans, and G. Wets. Developing an Optimised Activity Type Annotation Method Based on Classification Accuracy and Entropy Indices. Transportmetrica A: Transport Science, Vol. 9935, No. June, 2017, pp. 1–50.

Methodology

Data

1. Seoul HTS (2010): 11 ATCs; ~76,000 individuals; temporal variables
   - To confirm correct convergence (cfr. Ectors et al. (2017)) and benchmark performance gains
2. USA NHTS (2009): 37 ATCs; ~308,901 individuals; temporal variables
   - To the authors' knowledge the richest activity encoding in a HTS
   - The `ultimate' challenge to optimize ATCs (using the local search algorithm) because of the combinatorial challenge ($3.82 \cdot 10^{30}$ different sets of ATCs exist)
   - Popular data set: optimal set of ATCs useful info

Only temporal variables (activity start time, duration…) are used to infer activity types because

- Research as compatible as possible with other study areas
- Many applications start from e.g. GPS recorded or smart card data (temporal info readily available)

The `Home' activity is excluded from all analyses because

- `Home' is typically easy to classify with a very high accuracy
- The large share of easy-to-classify `Home' activities obscures suboptimal or bad classifications of out-of-home activities

Data was split in train set (75%) and test set (25%) to train and evaluate the ATC classifier

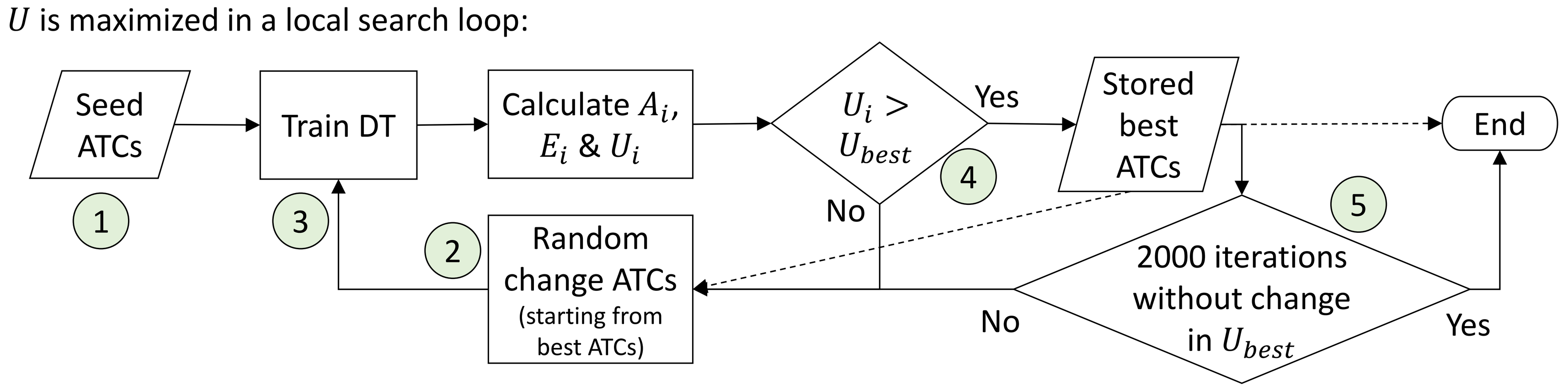
Optimizing through local search

Maximize parameter U

$$U = \frac{A_i - A_0}{A_{max} - A_{min}} - a\frac{E_0 - E_i}{E_{max} - E_{min}} = \frac{A_i - A_0}{1 - A_0} - a\frac{E_0 - E_i}{E_0}$$

The classification accuracy A is determined with the C4.5 (J48 in Weka) decision tree classifier (ATCs are predicted using temporal variables as explanatory variables). The entropy E is calculated with $E = -\sum_i p_i \log_2(p_i)$ where p_i the probability of ATC i. The factor a is a weight parameter ($a = 1$ for this study).

U is maximized in a local search loop:



① Start with reference set of ATCs (e.g. no ATCs grouped [[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]])

② Generate a new grouping scheme based on the current 'best' grouping scheme, but with some random changes, e.g. [[1], **[2, 10]**, [3], [4], [5], [6], [7], [8], [9]]. The number of random changes depend on an exponential distribution, up to a maximum of ten random changes. This approach decreases the probability that the algorithm gets stuck in a local optimum and increases the probability that it will reach the global optimum. Note that this step is not completely random, as previously generated random grouping schemes are never used again (for obvious performance reasons). The random change generator is insensitive to the size of an existing group. This prevents a bias of large groups getting only larger, or vice versa.

③ For this new set of ATCs, train a decision tree (DT) on the train set and subsequently calculate activity classification accuracy A_i based on the test set, and calculate entropy retention E_i in the data. Compute U_i.

④ If the newly calculated U_i is larger than U_{best} of the best grouping scheme, replace the best grouping scheme with the newly found grouping scheme.

⑤ Repeat step 2 to 4 until a stopping criterion is satisfied, indicating that the algorithm converged to a (local) optimum (which should be equal to the global optimum). For the Seoul HTS 2010 data set, iterations stopped after 100 cycles without a change in U_{best}, whilst for the NHTS 2009 data set this threshold was set to 2000 cycles.

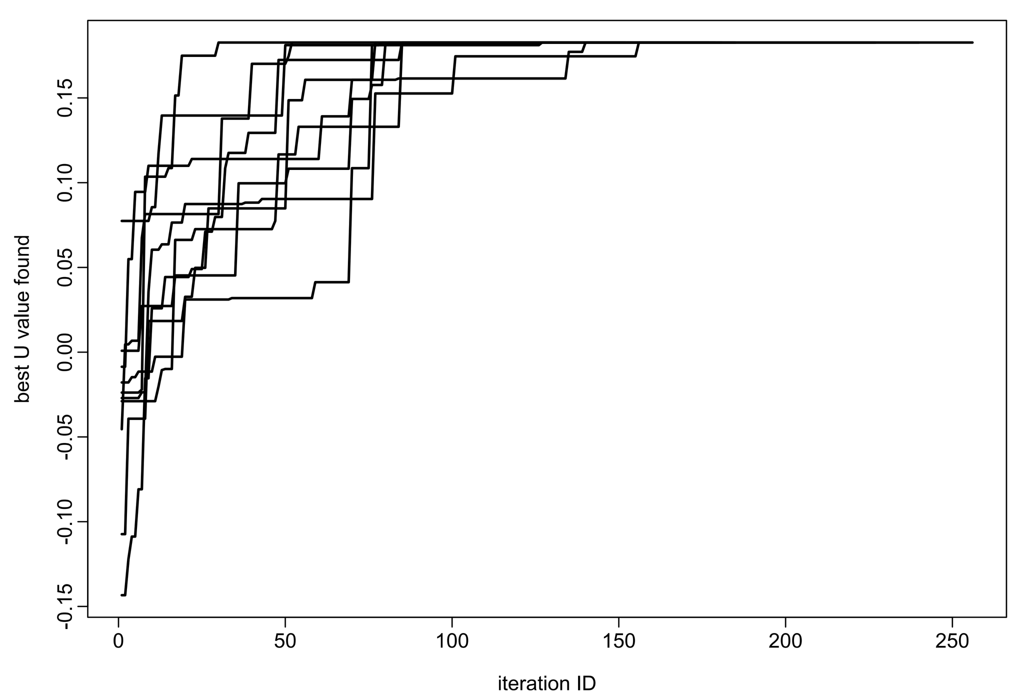
⑥ Step 1 to 5 can be repeated (optionally with different `seed' set). When consistently the same solution is found, this may be considered evidence for a global optimum.

Methods in this research:

- Confirm correct convergence (cfr. Ectors et al. (2017)) and performance improvements with the Seoul HTS
  - 10 independent runs of the local search optimization
- Find the optimal set of ATCs for the USA NHTS (which is the `ultimate' challenge or worst-case scenario)
  - 15 independent runs of the local search optimization

Results

- **Confirm correct convergence (cfr. Ectors et al. (2017)) and performance improvements with the Seoul HTS:**

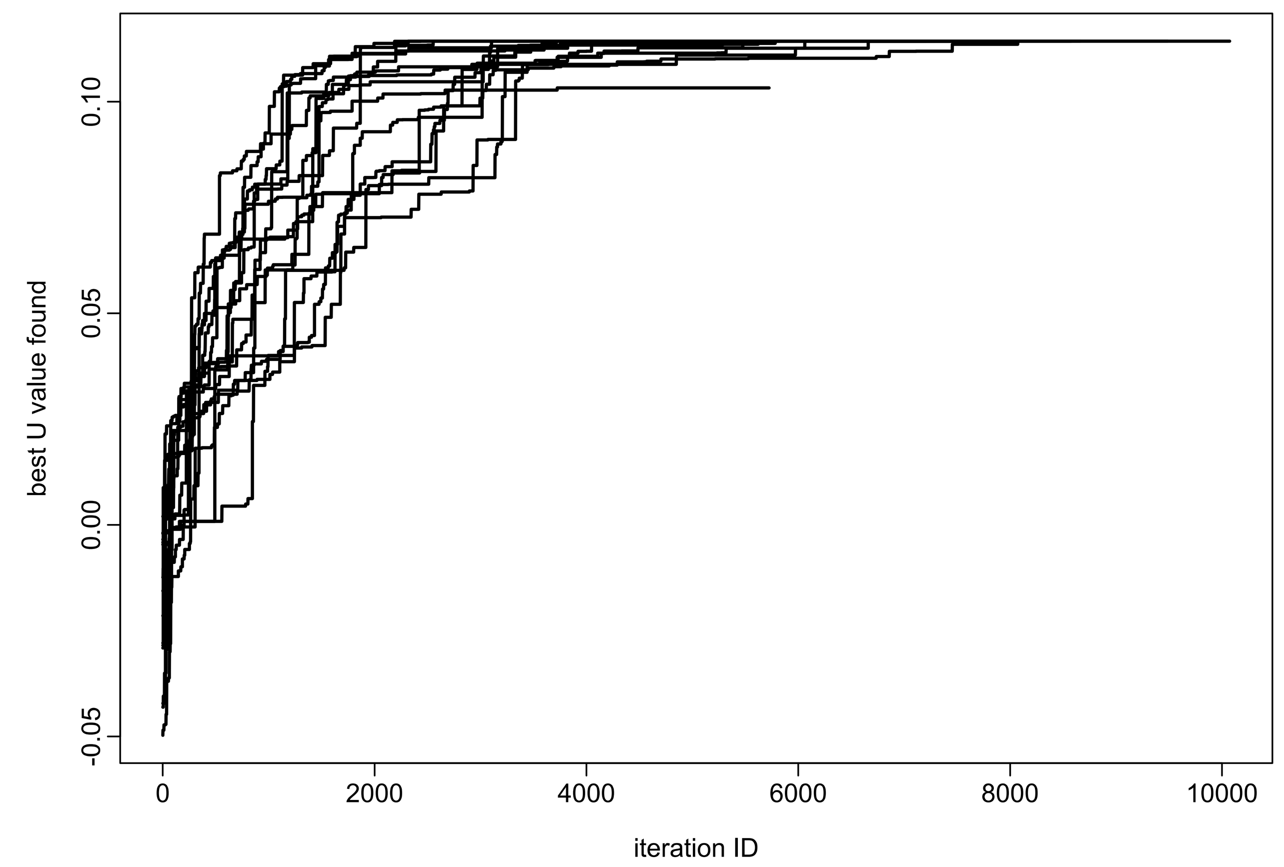


✓ Convergence to the same optimum for all 10 runs

✓ The optimum is the same set of ATCs as in Ectors et al. (2017)

✓ Instead of 30 hours of computation time (115,975 classifiers were trained in 20 parallel threads on a server using the method of Ectors et al. (2017)) **only a *few minutes* were needed with the local search algorithm**

- **Find the optimal set of ATCs for the USA NHTS:**



Main results for USA NHTS 2009:

✓ Convergence criterion reached for each run in 40.5h (4,324 iterations) to 92h (10,072 iterations) (a mere fraction of what would have been needed in Ectors et al. 2017)

✓ **The optimal set of ATCs was found**: out of 37 original ATCs group these activities in a single group

- Go to religious activity
- Medical/dental services
- Shopping/errands
- Buy goods: groceries/clothing/ hardware store
- Buy services: video rentals/dry cleaner/ post office/car service/ bank
- Buy gas
- Go to gym/exercise/play sports
- Visit friends/relatives
- Pick up someone
- Drop someone off
- Get/eat meal

✓ *Better* than `expert opinion' ATCs (see table below for comparison)

✓ Convergence to the same optimum for 13 out of 15 runs

Sets of activity classes (only grouped activity types are shown)	Test Set Accuracy	Entropy	U (↓)
[22, 30, 40, 41, 42, 43, 51, 53, 71, 73, 82] (the optimal group)	0.734	2.216	0.114272
[23, 70], [22, 30, 40, 41, 42, 43, 51, 53, 71, 73, 82]	0.734	2.214	0.113756
[10, 23, 70], [22, 30, 40, 41, 42, 43, 51, 53, 71, 73, 82]	0.734	2.214	0.113751
ref.: (Lee and Hickman, 2014)	0.851	0.977	0.001754
Reference case (original 36 activity types)	0.340	4.276	0
ref.: (Lu and Zhang, 2015)	0.895	0.618	-0.014185
ref.: e.g. (Kochan, 2012)	0.733	1.271	-0.107685
Grouped by first digit of NHTS codes	0.476	2.754	-0.150825
ref. e.g. (Bradley and Vovsha, 2005)	0.632	1.539	-0.197553
ref.: (Shen and Stopher, 2013)	0.599	1.741	-0.200993
ref.: (Lu et al., 2013)	0.485	2.429	-0.213240

Concluding remarks

- 2 out of 15 converged to a `suboptimal' set of ATCs because:
  - of too few iterations before the stopping criterion is fulfilled?
  - ~~these runs were stuck in a local optimum?~~

  The latter can be rejected since some runs that *did* successfully converge to the optimum encountered the suboptimal set of ATCs during their iterations, meaning that there *was* a direct path that could lead to the same optimum. *By chance* (i.e. too few iterations before stopping criterion was fulfilled) such a path was not found for 2 out of 15 runs.
- The grouped ATCs have similar *temporal* properties (usually not a long duration; could occur at any time of the day; likely to be chained together)
- Unless the research demands a particular set of ATCs, one should consider using the optimized ATCs

Contact information: Wim Ectors, Hasselt University, Transportation Research Institute (IMOB), Agoralaan, 3590 Diepenbeek, Belgium. E-mail: wim.ectors@uhasselt.be