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

OPTIMIZING COPIOUS ACTIVITY TYPE CLASSES BASED ON CLASSIFICATION ACCURACY AND ENTROPY RETENTION

Wim Ectors, Corresponding Author
Hasselt University, Transportation Research Institute (IMOB),
Agoralaan, BE-3590 Diepenbeek, Belgium
Tel.: +32(0)11 26 91 14; Fax.: +32(0)11 26 91 99; Email: wim.ectors@uhasselt.be

Sofie Reumers
Hasselt University, Transportation Research Institute (IMOB),
Agoralaan, BE-3590 Diepenbeek, Belgium
Tel.: +32(0)11 26 91 60; Fax.: +32(0)11 26 91 99; Email: sofie.reumers@uhasselt.be

Won Do Lee
Manchester Metropolitan University, Crime and Well-being Big Data Centre,
All saints, M15 6BH Manchester, England
Tel.: + 44(0)161 247 6538; Email:w.lee@mmu.ac.uk

Bruno Kochan
Hasselt University, Transportation Research Institute (IMOB),
Agoralaan, BE-3590 Diepenbeek, Belgium
Tel.: +32(0)11 26 91 47; Fax.: +32(0)11 26 91 99; Email: bruno.kochan@uhasselt.be

Davy Janssens
Hasselt University, Transportation Research Institute (IMOB),
Agoralaan, BE-3590 Diepenbeek, Belgium
Tel.: +32(0)11 26 91 28; Fax.: +32(0)11 26 91 99; Email: davy.janssens@uhasselt.be

Tom Bellemans
Hasselt University, Transportation Research Institute (IMOB),
Agoralaan, BE-3590 Diepenbeek, Belgium
Tel.: +32(0)11 26 91 27; Fax.: +32(0)11 26 91 99; Email: tom.bellemans@uhasselt.be

Geert Wets
Hasselt University, Transportation Research Institute (IMOB),
Agoralaan, BE-3590 Diepenbeek, Belgium
Tel.: +32(0)11 26 91 58; Fax.: +32(0)11 26 91 99; Email: geert.wets@uhasselt.be

Word count: 5988 words text + 4 tables x 250 words + 2 figures x 250 words = 7488 words

Paper submitted: July 31, 2017
Revised version submitted: October 31, 2017

ABSTRACT

Despite the advantages, big transport data are characterized by a considerable disadvantage as well. Personal and activity-travel information are often lacking, making it necessary to deduce this information with data mining techniques.

However, some studies predict many unique activity type classes (ATCs), while others merge multiple activity types into larger ATCs. This action enhances the activity inference estimation, but destroys important activity information. Previous studies do not provide a strong justification for this practice. An objectively optimized set of ATCs, balancing model prediction accuracy and preserving activity information from the original data, becomes essential.

Previous research developed a classification methodology in which the optimal set of ATCs was identified by analyzing all possible ATC combinations. However, for the US National Household Travel Survey (NHTS) 2009 data set which comprises 36 ATCs (home activity excluded), this approach is practically impossible in a finite amount of time since there would be 3.82*10^30 unique combinations.

The aim of this paper is to optimize which original ATCs should be grouped into a new class, and this for data sets for which it is impossible or impractical to simply calculate all ATC combinations. The proposed method defines an optimization parameter U (based on classification accuracy and information retention) which is maximized in an iterative search algorithm. The optimal set of ATCs for the NHTS 2009 data set was determined. A comparison finds that this optimum is considerably better than many expert opinion activity type classification systems. Convergence was confirmed and performance gains were benchmarked.

Keywords: Activity type classification, (Big) transport data annotation, optimal set of activity types, local search algorithm, classification accuracy, entropy indices

INTRODUCTION

These days big data sets are collected continuously and in real time, making large amounts of data that are temporally and spatially referenced available to researchers (*1*). Furthermore, advancements in ICT and the improvement of location-aware technologies facilitate the collection of transport data, e.g. daily trajectories. The new transport data-collection methods support researchers with refined, detailed data sets of real-time data. These large collections of spatio-temporal information offer research opportunities, i.e. they enable a better investigation and understanding of human travel behavior.

Due to the availability of temporal information (e.g. time stamps), big transport data are very effective in exploring individual mobility patterns. Despite the advantages, big transport data are characterized by a considerable disadvantage as well. Personal and activity-travel information are often lacking (*2*), making it necessary to deduce this information from the available travel patterns.

In order to overcome this shortcoming, behavioral data mining techniques are frequently used to infer activity types (sometimes otherwise denoted as trip or travel purposes, activity classes, activity categories or activity encoding) from behavioral attributes, such as temporal attributes and spatial information (e.g. (*3–5*)). In recent studies on activity-travel data mining, different inference techniques are investigated. However, in these researches different classifications of activity types exist. Some studies infer many activity classes, while others aggregate or group several activity types, limiting the number of activity type classes (ATCs) (*6*). As argued in (*6*), in none of these studies a strong justification is established. The activity type classification in the majority of researches merely relies on the travel survey design, due to a lack of clear standards for ATCs which is grounded by a theoretical background (*7*). The ATCs (and the size of this set of classes) strongly affect the classification accuracy. Often, activity types are aggregated in order to enhance the activity inference estimation. However, by aggregating activity types, and thus enhancing the activity inference estimation, important activity information is lost. Therefore, the need for a standardized method for activity categorization arises. An optimal set of activity types is an essential prerequisite for a robust and sound transport data annotation. The proposed method is an objective alternative to the subjective ATCs based on intuition.

Previous research (*6*) developed a classification methodology using a rule-based heuristic algorithm in which the optimal grouping of ATCs was identified. The optimization method searches for an optimal balance between improving model accuracy and preserving activity information from the original data set. This method focused solely on the temporal attributes (i.e. activity start time and activity duration) from household travel survey (HTS) data sets, in order to develop a generic categorization method which is applicable to as many big data sources as possible. Other types of attributes, e.g. spatial information, can however also be used in the proposed method. The method was applied to two HTSs, i.e. the Seoul HTS and the Flanders (Belgium) HTS called OVG.

The optimization method in (*6*), however, might not be appropriate when the initial data set contains too many unique ATCs. The optimization strategy comprises three stages, where in the first stage all possible combinations of ATCs are generated. This brute-force approach calculates approximately 117,000 unique sets of combinations of classes for both the OVG and Seoul HTS as both HTS data sets consist of only 10 distinct activity types, when the 'Being at home' activity is excluded from the experiment. The home activity was excluded from the experiment, because this activity type is quite easy to classify and is mostly predicted with a very high accuracy (e.g. (*8*)). Additionally, due to a large share of home activities in the data set, its good classification capability obscures the suboptimal or bad classifications of out-of-home activities. In the second stage of the optimization strategy, classifiers are trained and tested on the data that were transformed according to the ATC combinations of the first stage. Finally, the optimal set of ATCs is defined in the third stage of the optimization method. On a server equipped with two intel Xeon EQ-2643 v2 processors (running at approximately 80% capacity, i.e. 20 threads) estimating 117,000 classifiers took roughly 30 hours of computation time. However, for the US National Household Travel Survey (NHTS) 2009 data set (*9*) which comprises 36 ATCs (home activity excluded), calculating classifiers for all possible grouping combinations is impossible since the increase in distinct combinations is exponential. In other words, a

large number of initial activity types (*n* in Table 1) which are considered for aggregation will result in an extremely large set of grouping combinations that needs to be processed as shown in Table 1. Therefore, the computation time of the second stage of the optimization method would rise up to 1.13*10^23 years for the US NHTS data set on the same server using 20 threads. Note that the age of the universe is only 13.8*10^9 years (*10*).

TABLE 1 Number of possible activity type class (ATC) combinations as a function of the number of activity types, *n*

n	# of ATC combinations	n	# of ATC combinations
1	1	21	4.748698E+14
2	2	22	4.506716E+15
3	5	23	4.415201E+16
4	15	24	4.459589E+17
5	52	25	4.638590E+18
6	203	26	4.963125E+19
7	877	27	5.457170E+20
8	4,140	28	6.160539E+21
9	21,147	29	7.133980E+22
10	115,975	30	8.467490E+23
11	678,570	31	1.029336E+25
12	4.213597E+06	32	1.280647E+26
13	2.764444E+07	33	1.629596E+27
14	1.908993E+08	34	2.119504E+28
15	1.382959E+09	35	2.816002E+29
16	1.048014E+10	36	3.819715E+30
17	8.286487E+10	37	5.286837E+31
18	6.820768E+11	38	7.462899E+32
19	5.832742E+12	39	1.073882E+34
20	5.172416E+13	40	1.574506E+35

TABLE 2 Examples of household travel survey data sets with their number of distinct activity type classes (ATCs)

Data set	Country (or region) of origin	Number of person days surveyed	Number of ATCs (home activity excluded)
AUS VISTA 2007 & 2009 (*11*, *12*)	Australia	67,060	12
BEL Beldam 2010 (*13*)	Belgium	11,279	11
BEL OVG 3.0-4.5 (*14*)	Belgium (Flanders)	13,522	10
CHE Thurgau 2003 (*15*)	Switzerland	8,522	25
DEU Mobidrive 1999 (*16*)	Germany	13,244	22
FIN HLT 2010-2011 (*17*)	Finland	10,137	19
FRA ENTD 2008 (*18*)	France	17,996	31
GBR NTS 2009-2014 (*19*)	United Kingdom	551,234	22
IRL NTS 2009 (*20*)	Ireland	5,023	9
KOR Seoul HTS 2010 (*21*, *22*)	South Korea	219,269	10
NLD OViN 2013 (*23*)	The Netherlands	34,710	13
SVN Ljubljana 2013 (*24*)	Slovenia	3,426	12
SWE RVU 2011-2014 (*25*)	Sweden	31,457	25
USA NHTS 2009 (*9*)	United States of America	257,586	36

Considering that the US NHTS is not the only data set which consists a large number of activity types, this computation time issue will surface for other travel data sets. In the UK HTS data set (*26*), for example, 22 distinct activity types are employed. In Table 2, several travel data sets are listed, together with the number of activity types that are considered in each case.

To overcome this process time issue, the research in this paper proposes an update of the optimization categorization methodology using a '*local search'* algorithm. The local search algorithm starts from a predefined ATC grouping combination and iteratively tries to optimize this group by applying random changes, and thus reducing the required computation process time. The remainder of this paper is structured as follows. The next section describes the data and clarifies the methodology. Subsequently, the results of the convergence of the local search algorithm are presented, followed by the optimal ATCs for annotation. Finally, a conclusion is formulated.

METHODOLOGY

Data description

Two HTSs were used in this research. The first HTS, the Seoul HTS, was organized in the Seoul Metropolitan Area (SMA), Republic of Korea, in 2010. This data set consists of self-reported daily household activity-travel data from approximately 76,000 individuals. As reported in Table 2, this data set contains 11 distinct trip motives (or activity types), of which the 'home' activity will be excluded as justified in the introduction. The Seoul HTS was included in this study to confirm the correct convergence of the proposed search algorithm to the optimum which was found in (*6*), and to benchmark the algorithm's performance gains. The convergence on this data set will be discussed and the performance of the algorithm will be compared to the approach in (*6*), justifying the need and benefits of the iterative search approach. The optimum set of ATCs of this data set will however not be discussed here. Interested readers may find a thorough analysis in (*6*).

The second HTS used in this study is the NHTS from the USA in 2009. It contains surveyed information from 308,901 individuals. This massive data set contains detailed trip information of approximately 1.17*10^6 trips, of which the trip purpose is encoded in 37 distinct classes. After excluding trips having the 'home' trip purpose, approximately 768,000 records remain to train activity type classifiers. The copious activity types in this data set are the reason for the development of the proposed methodology, as explained in the introduction. To the author's best knowledge, this is the richest activity type encoding in a HTS (not considering time-use surveys); see also Table 2. It is therefore a challenge to find the optimal set of activity types, which may be used in any activity type inference or annotation research. Additionally, this data set is employed in many studies to train their models. Finding and using an optimal set of activity types may enable the seamless consolidation of multiple research outcomes.

As mentioned earlier in the introduction, only temporal variables such as activity start time and duration are used to train classifiers in this study. All other variables in the data are disregarded. This choice was made in order to make this research as compatible as possible with other study areas. Additionally, many applications start from e.g. GPS recordings, smart card data etc. for which classification based on temporal variables gives already good results (*27*).

The data was split in a train set (75%) and test set (25%). According to common practice, the train set is used to train a classifier, whilst the test set is used to evaluate its prediction accuracy on 'new' data.

Grouping of activity types

This section discusses the combinatorial challenge of grouping or aggregating of activity types into new classes. For example, in the set of ATCs [[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]], activity types 3 and 6 may be merged into a new class as such: [[1], [2], [3, 6], [4], [5], [7], [8], [9], [10]]. The number of possible ATCs grows exponentially with the number of distinct activity types: n. This is the result of all the permutations of activity types across possible groups and the different combinations of possible

group sizes. The order of activity types within a group, and the order of the groups among themselves does not matter. The possible group size combinations for a given n may be obtained by computing the integer partitions. For example, for $n = 4$ the integer partitions are {1+1+1+1, 2+1+1, 2+2, 3+1, 4}. Each element in these partitions represents a group's size. The first partition represents the case where no activity types are merged, the final one represents the case where all 4 activity types are grouped in one group of size 4. For each partition i: $g_1 + g_2 + \cdots + g_j = n$, there are x_i number of ways to distribute n activity types across the groups:

$$x_i = \frac{n!}{\prod_j \left(g_j! \cdot \sqrt[\frac{1}{f_j}]{f_j!} \right)} \quad (1)$$

where f_j the frequency of a particular element in the partition (which represents a group's size). For example, in the partition 2+1+1, element '2' has a frequency of 1. In partition 2+2, element 2 has a frequency of 2. The factor $\prod_j \left(\sqrt[\frac{1}{f_j}]{f_j!} \right)$ corrects x_i for the permutations of equal-sized groups as the order of these equal-sized groups is unimportant, and should not increase x_i (that is, $2_a + 2_b = 2_b + 2_a$). The total number of possible ATC combinations is the sum of all x_i for a given n and its integer partitions i. These values are listed in Table 1. One observes how the increase of possible combinations increases exponentially, hereby strengthening the justification for the need of the proposed methodology.

Optimization through local search

In order to optimize the ATCs, the proposed method combines some of the original activity types into a new class, and subsequently calculates the classification accuracy and entropy of the activity type variable. The classification accuracy represents the performance of predicting an ATC, and the entropy represents the amount of information such a prediction is giving. The entropy (or embedded information) is greatest when no activity types are merged into a new class, yet the classification accuracy increases when activity types are merged into new classes (as there are fewer classes to predict). The aim of this paper is to optimize which original activity types should be grouped into a new class, and this for data sets for which it is impractical or impossible to simply calculate all ATC combinations (due to an extremely large amount of combinations). The proposed method defines an optimization parameter U which is maximized in an iterative search algorithm.

At the heart of the optimization strategy in (*6*) is the optimization parameter which may be calculated using Equation *(2)*

$$U = \frac{A_i - A_0}{R_A} - a\frac{E_0 - E_i}{R_E} \quad (2)$$

where A_i is the test set accuracy and E_i the activity type entropy of a particular combination of ATCs i. A_0 and E_0 are, respectively, the test set accuracy and activity type entropy of the reference case of no activity type aggregation into new classes. $R_A = A_{max} - A_{min}$ is the range in test set accuracy improvement and $R_E = E_{max} - E_{min}$ is the range in entropy reduction, observed within the set of results of all ATC combinations. a can be used to give a relative weight to either the classification accuracy improvement or to the entropy retention if there exists such an intrinsic bias for one of these indices. A sensitivity analysis of this parameter is described in (*6*). The entropy may be calculated with Equation *(3)*:

$$E = -\sum_i p_i \log_2(p_i) \quad (3)$$

where p_i is the probability on class i.

However, Equation *(2)* can only be used when the results from all ATC combinations are known, as R_E depends on the minimum entropy E_{min}, and R_A requires the maximum classification accuracy A_{max} to be known. Note that the maximum entropy E_{max} and minimum classification accuracy A_{min} can however be obtained from the reference case in which no activity types are grouped into a new class. In (*6*), the optimization parameter U was calculated only *after* the entropy and classification accuracy for all approx. 117,000 ATC combinations were calculated. Since calculating the entropy and classification accuracy for all possible combinations of ATCs is impossible given a large number of distinct activity types in the USA NHTS 2009 (see Introduction), E_{min} and A_{max} need to be substituted.

The answer consists of allowing the trivial solution, that is the case when all activity types are grouped into a single large class. In this trivial case, the entropy is zero (all activity type information is lost) and the classification accuracy is 100% (as only one class remains to predict). The results in (*6*) reveal that in practice E_{min} and A_{max} are in fact very close to, respectively, zero and one, thus supporting the proposed measure. Doing so, Equation *(2)* may be simplified to the following form in which all parameters (except A_i and E_i) can be calculated from the start:

$$U = \frac{A_i - A_0}{A_{max} - A_{min}} - a\frac{E_0 - E_i}{E_{max} - E_{min}} = \frac{A_i - A_0}{1 - A_0} - a\frac{E_0 - E_i}{E_0} \quad (4)$$

As a result, U may be calculated without the need to calculate the classification accuracy and entropy for all possible ATC combinations beforehand. An iterative optimization approach is now possible. The proposed optimization algorithm performs the following steps:

1. Start without grouping activity types into new classes (all activity types form their own group). This is the reference set of ATCs, e.g. this set of ten distinct activity types: [[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]]. For now this is the best grouping scheme.
2. Generate a new grouping scheme based on the above 'best' grouping scheme, but with some random changes, e.g. the case where activity type '2' and '10' are merged into a new class: [[1], [2, 10], [3], [4], [5], [6], [7], [8], [9]]. A random change is defined as the exchange of one activity type from one group to another (this can be an empty group). The number of random changes that are applied are according to an exponential distribution: the probability of a single change is 64.4%, that of two changes 23.7%, that of three changes 8.7% etc. and this up to a maximum of ten random changes. Using 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. Multiple blocks of ATCs can arise without biasness.
3. For this new set of ATCs, train a decision tree (DT) on the train set and calculate activity classification accuracy based on the test set, and calculate entropy retention in the data. Compute U (Equation *(4)*) while taking $a = 1$ for this study.
4. If the newly calculated U is larger than the U 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 is possibly equal to the global optimum). For the Seoul HTS

2010 data set, iterations stopped after 100 cycles without a change in best U, 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). Each case could potentially converge to a different local optimum. If however consistently the same solution is found, this may be considered evidence for a global optimum.

The C4.5 (J48 in Weka (*28*)) DT classification algorithm yields an excellent classification accuracy and requires only a short time to train it (*6*). This was the classifier of choice in step 3. In step 5 it was explained that after a predefined number of iterations without change of the optimum, the algorithm would stop. This predefined number was chosen after initial experimentation and may not be optimal. It is however critical that this number is chosen sufficiently large for cases of copious distinct activity types. This is important since more combinations of classes are possible and thus the optimal becomes more difficult to find. The algorithm needs sufficient time to try random different combinations before one can conclude convergence.

In the experiments described in this paper, the algorithm was run for 10 (Seoul HTS 2010) or 15 (NHTS 2009) times as described above in step 6. Due to the random changes applied in step 2, each run had a different path of convergence. Yet, as will be discussed in the results section, consistently the same optimum was found giving evidence for a global optimum

RESULTS

Convergence of the local search algorithm

First the proposed algorithm was run for the Seoul HTS 2010 data set, similar as in (*6*). The intention of this experiment is to confirm that the proposed algorithm works, and that it yields major improvements in performance. Compared to (*6*), slightly different values for U are expected since an adapted formula is used in this study. The algorithm ran for 10 times (independently) and converged each time to the same optimum, which was reassuringly also the same as was found in (*6*). Figure 1 illustrates the convergence of these runs. Although each run started at a different U value due to the random change at the start of the algorithm, they all converged to the same optimal U value. Notably, this exact same result could be found *in just a couple of minutes*, whilst in the approach of (*6*) approximately 30 hours on 20 threads of a high-end server were needed. As also concluded in that study, this optimum is considerably better than many 'expert opinion' activity type classification systems being used.

After having confirmed the excellent performance of the method on the Seoul HTS, the experiment was repeated on the NHTS 2009 data set. Within 15 parallel runs, approximately 97,000 distinct combinations were calculated in total. Table 3 lists a selection of all those combinations, including also the most optimal set of ATCs as the first entry in the table (see also next section). The first run found the optimum in just over 40.5 hours (after 4,324 iterations), the last one in just under 92 hours (10,072 iterations). Mind that in all runs the final 2000 iterations were part of the stopping criterion. Note that this is a mere fraction of the 3.82*10^30 sets of ATC combinations that would have to be analyzed with the method of (*6*). Compared to the Seoul HTS, processing time for NHTS 2009 took considerably longer. This is a consequence of the increased time to transform the ATCs in the data and subsequently train the DTs on this large data set.

Figure 2 illustrates the convergence of the 15 runs. One immediately notices how all seem to have converged to the same maximum U value. However, there are two runs who failed to converge to this maximum value, of which one run is clearly visible up to approximately iteration 6,000. The U value at which these runs reached the stopping criterion is inferior to the one at which the 13 other runs converged. This may be caused by too few iterations before the stopping criterion is fulfilled (set to 2,000 in this experiment) or that these runs were stuck in a local optimum. The latter is unlikely, since by allowing up to 10 random changes on the previous best scheme (see Methodology section) it would be very likely that any local optimum could be avoided, on the condition that the algorithm was given

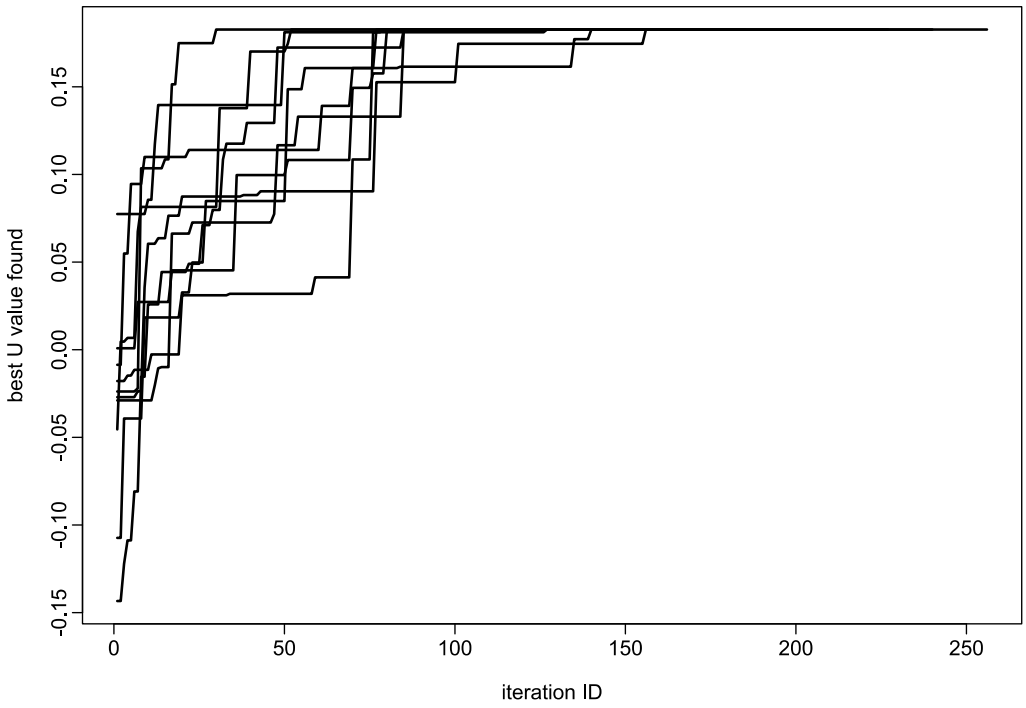

FIGURE 1 Convergence plot for 10 independent runs on the Seoul HTS 2010 data set. The slowest run finished in just under 13 minutes (Intel Core i5-4210M CPU @ 2.60GHz) and needed 255 iterations (100 part of the stopping criterion). All 15 found the same optimum

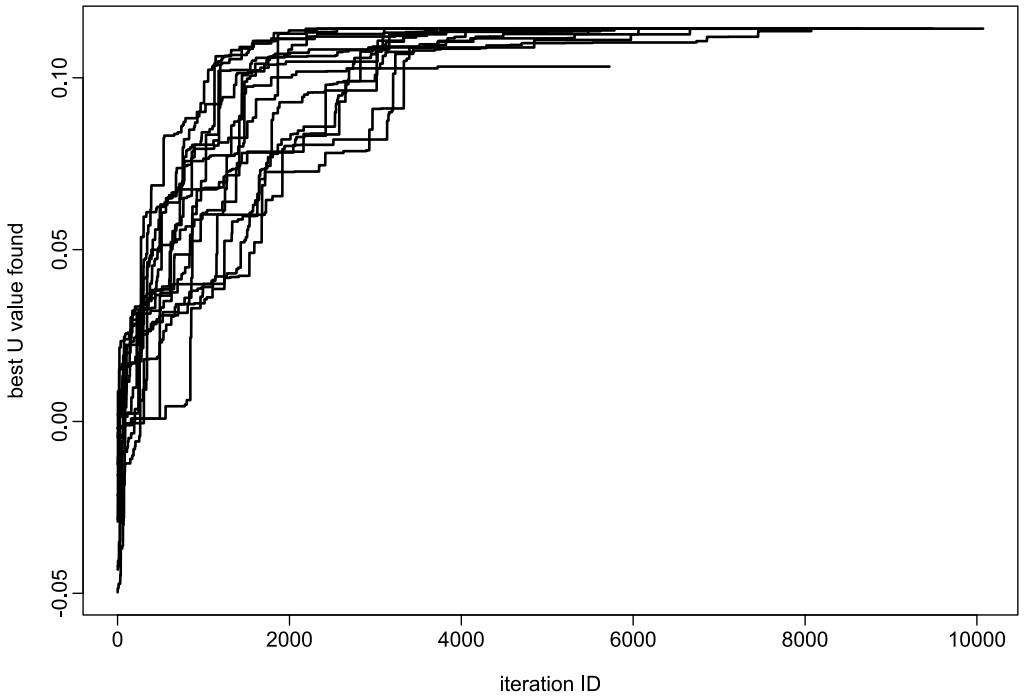

FIGURE 2 Convergence plot for 15 independent (parallel) runs on the NHTS 2009 data set. The last thread finished in just under 92 hours (Intel Xeon E5-2660 v4 CPU @ 2.00GHz). All but two found the same optimum

TABLE 3 Most optimal sets of the combined results of the 15 parallel runs on the NHTS 2009 data set and some interesting sets to compare with. Table 4 lists the encoding of the activity types. The best set of activity classes (1st row in table) was the end result in 13 out of 15 runs. The *italic sets of activity classes represent multiple variations with 10: 'Work' (see text)**

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]	0.734	2.216	0.114272
*[10, 23], [22, 30, 40, 41, 42, 43, 51, 53, 71, 73, 82] **	0.734	2.216	0.114268
[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
*[10, 62], [23, 70], [22, 30, 40, 41, 42, 43, 51, 53, 71, 73, 82] **	0.734	2.214	0.113751
[10, 11, 12, 13, 14], [20, 21, 22, 23, 24], [30, 40, 41, 42, 43, 50, 51, 52, 53, 54, 55, 60, 61, 62, 63, 64, 65, 70, 71, 72, 73, 80, 81, 82, 83, 97] (ref.: (*29*))	0.851	0.977	0.001754
Reference case (original 36 activity types)	0.340	4.276	0
[10, 11, 12, 13, 14], [20, 21, 22, 23, 24, 30, 40, 41, 42, 43, 50, 51, 52, 53, 54, 55, 60, 61, 62, 63, 64, 65, 70, 71, 72, 73, 80, 81, 82, 83, 97] (ref.: (*4*))	0.895	0.618	-0.014185
[10, 11, 12, 13, 14], [23, 24, 30, 40, 41, 42, 43, 50, 51, 52, 53, 54, 55, 63, 64, 80, 81, 82, 83, 97], [20, 21, 22, 60, 61, 62, 65, 70, 71, 72, 73] (ref.: e.g. (*30*))	0.733	1.271	-0.107685
[10, 11, 12, 13, 14], [20, 21, 22, 23, 24], [40, 41, 42, 43], [50, 51, 52, 53, 54, 55], [60, 61, 62, 63, 64, 65], [70, 71, 72, 73], [80, 81, 82, 83] (ref.: (*9*) (first digit NHTS codes))	0.476	2.754	-0.150825
[24, 30, 40, 41, 42, 43, 61, 64, 65, 82], [10, 11, 12, 13, 14, 20, 21, 70, 71, 72, 73], [22, 23, 50, 51, 52, 53, 54, 55, 60, 62, 63, 80, 81, 83, 97] (ref. e.g. (*31*))	0.632	1.539	-0.197553
[10, 11, 12, 13, 14], [20, 21, 22, 23, 24], [30, 40, 41, 42, 43], [50, 51, 52, 53, 54, 55, 60, 61, 62, 63, 64, 65, 70, 71, 72, 73, 80, 81, 82, 83, 97] (ref.: (*32*))	0.599	1.741	-0.200993
[40, 41, 42, 43], [70, 71, 72, 73], [10, 11, 12, 13, 14], [20, 21, 22, 23, 24], [50, 51, 52, 53, 54, 55], [30, 60, 61, 62, 63, 64, 65, 80, 81, 82, 83, 97] (ref.: (*33*))	0.485	2.429	-0.213240

enough time. Both runs would also converge to the same optimum as the others with a better setting of the stopping criterion. The other 13 runs, which started at different random variations on the initial scheme, converged to the same optimum which is greater than that of the two which did not converge to this *U* value. This gives confidence for a globally optimal set of ATCs.

Optimal activity type classes for annotation

Table 3 lists some interesting results from the experiments on the NHTS 2009 data. The first entry is the most optimal ATC combination: 10, 11, 12, 13, 14, 20, 21, 23, 24, 50, 52, 54, 55, 60, 61, 62, 63, 64, 65, 70, 72, 80, 81, 83, 97, [22, 30, 40, 41, 42, 43, 51, 53, 71, 73, 82]. Compared to the reference case, its test set classification accuracy has more than doubled from 34.0% to 73.4% . This comes at a cost of

losing 2.06 bits of information. It suggests these 26 distinct classes (excl. 'Home') are more optimal compared to the original 36. It merges the following activity types into a new class:

- 22: Go to religious activity
- 30: Medical/dental services
- 40: Shopping/errands
- 41: Buy goods: groceries/clothing/hardware store
- 42: Buy services: video rentals/dry cleaner/post office/car service/bank
- 43: Buy gas
- 51: Go to gym/exercise/play sports
- 53: Visit friends/relatives
- 71: Pick up someone
- 73: Drop someone off
- 82: Get/eat meal

This is a class which is hard to define. Some are flexible in nature (buying goods,…) yet others have obligations to third parties and are not flexible (e.g. picking up or dropping off someone). However, none of them usually have a very long duration and could in theory occur at almost any time within a day. All activity types occur at a relatively high frequency (see Table 4). Many of these activities are likely to be chained together: picking up or dropping off people whilst visiting friends/relatives or going to play sports, or getting something to eat before (or after) doing some shopping etc. This makes it hard to distinguish between these activity types based on only temporal profiles, and hence it makes sense to merge them into a single class.

The next ATC schemes in the list combines 10: 'Work' with all other activity types which are not in the large group of the most optimal scheme (24 distinct combinations, e.g. [10, 23]; [10, 70]; [10, 62]; etc.) and finally it also joins the large group. Because of space constraints, only the best performing of all those variations is listed in italics in Table 3. From Table 4 one observes that activity type 10 is slightly peculiar, as its weighted frequency is many orders of magnitude smaller than other activity types. It is clearly different from 'Go to work' as the latter has a frequency which is approximately 10^5 times larger. The exact definition of the 'Work' activity could not be found. Because of the very low frequency, the impact of this activity type on the classification accuracy and entropy retention is very small. This experiment concludes that in practice these variations with activity type 10 may not be different from the most optimal scheme and one could most likely ignore them.

Subsequently in Table 3 one finds the scheme where, in addition to the large group from before, also 23: 'Go to library: school related' and 70: 'Transport someone' are merged into a single class. This could make sense as this experiment used only time-related variables to train the DTs, and one could intuitively think the temporal distributions of both activity types may be similar. Again different combinations with activity type 10 are listed afterwards. The schemes discussed so far perform similar as the most optimal scheme. One has to be cautious when interpreting the rank in Table 3 as the algorithm does not guarantee to find all ATC combinations.

Next in Table 3 are seven interesting activity class combination schemes from literature to compare with the optimal scheme. An attempt was made to merge ATCs in a similar fashion as in these studies. The most obvious comparison may be made with an ATC scheme based on the first digit of the NHTS codes (*9*). Even though there are much fewer activity classes to predict compared to the most optimal scheme, its classification accuracy is much lower at 47.6% compared to 73.4%. This deficiency outweighs the fact that this scheme retains slightly more information than the optimal scheme. The scheme based on (*33*) performs similarly. The ones inspired by (*31*) and (*32*) perform worse than the optimal scheme on *both* the classification accuracy and information retention. The schemes inspired by (*4*), (*29*) and (*30*) have similar of better classification accuracies compared to the optimal scheme, however these lost a major portion of their information content as a consequence.

Depending on the research, there might exist a reason for employing one of the suboptimal schemes (e.g. some activity types *need* to be predicted and may not be merged, or a predefined number of ATCs is required). Yet, without such justification, this work suggests one should strongly consider using the revealed most optimal set of ATCs in order to simultaneously maximize the prediction accuracy and the information in that prediction.

TABLE 4 Trip motive codes in NHTS 2009 which were used in this study's optimization of activity type classes. There are 37 distinct codes (including 'Home')

NHTS 2009 codes	Description of trip motive	Weighted frequency
1	Home	1.35E+11
10	Work	2.16E+05
11	Go to work	3.11E+10
12	Return to work	5.73E+09
13	Attend business meeting/trip	1.07E+09
14	Other work related	7.90E+09
20	School/religious activity	1.13E+09
21	Go to school as student	1.18E+10
22	Go to religious activity	6.98E+09
23	Go to library: school related	4.54E+08
24	OS - Day care	8.29E+08
30	Medical/dental services	6.30E+09
40	Shopping/errands	7.10E+09
41	Buy goods: groceries/clothing/hardware store	4.40E+10
42	Buy services: video rentals/dry cleaner/post office/car service/bank	1.12E+10
43	Buy gas	6.60E+09
50	Social/recreational	3.78E+09
51	Go to gym/exercise/play sports	1.34E+10
52	Rest or relaxation/vacation	3.28E+09
53	Visit friends/relatives	1.76E+10
54	Go out/hang out: entertainment/theater/sports event/go to bar	6.84E+09
55	Visit public place: historical site/museum/park/library	1.85E+09
60	Family personal business/obligations	4.48E+09
61	Use professional services: attorney/accountant	1.11E+09
62	Attend funeral/wedding	6.68E+08
63	Use personal services: grooming/haircut/nails	1.47E+09
64	Pet care: walk the dog/vet visits	2.94E+09
65	Attend meeting: PTA/home owners association/local government	1.61E+09
70	Transport someone	3.09E+08
71	Pick up someone	1.10E+10
72	Take and wait	1.19E+09
73	Drop someone off	1.20E+10
80	Meals	7.92E+08
81	Social event	2.49E+09
82	Get/eat meal	2.04E+10
83	Coffee/ice cream/snacks	2.98E+09
97	Other reason	2.59E+09

CONCLUSION

As demonstrated in previous research (*6*), there is a strong need for activity categorization standards in the domain of trip purpose annotation (activity type classification). An optimal set of activity type classes (ATCs) is an essential prerequisite for a robust and sound transport data annotation, or any modeling exercise. Most existing researches use a suboptimal set of ATCs in their methodology (without providing a justification), leading to high classification accuracies, but low information in the prediction. An optimization strategy that was proposed in previous research (*6*), has shown a limitation: the issue of copious distinct ATC combinations and its associated long computation time.

The aim of this paper is to optimize which original activity types should be merged into a new class, and this for data sets for which it is impractical or impossible to simply calculate all ATC combinations due to an extremely large amount of combinations. The paper suggests a revision of the optimization method in (*6*). The proposed method defines an optimization parameter U, based on classification accuracy and information retention, which is maximized in an iterative search algorithm.

To confirm the correct convergence of the search algorithm and to benchmark the performance gains needed, the algorithm was run for ten times (independently) on the Seoul household travel survey in 2010. These converged to the same optimum as in (*6*) *in just a couple of minutes*, whilst in the approach of (*6*) approximately 30 hours on 20 threads of a high-end server were needed.

In fifteen parallel runs on the very large national household travel survey (NHTS) of the U.S. in 2009, approximately 97,000 distinct combinations were calculated. Thirteen runs found the same most optimal set of ATCs in merely 40.5 hours (after 4,324 iterations) to 92 hours (10,072 iterations) instead of an estimated 1.13*10^23 years (3.82*10^30 sets of ATC combinations) using the method of (*6*). The two remaining runs reached the stopping criterion prematurely. The most optimal set of ATCs for the NHTS 2009 creates only a single group, in which the following NHTS 2009 activity types are merged into a new class:

- 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

This is a class which is hard to define, but the activity types have in common that they usually don't have a very long duration and that they could in theory occur at almost any time within a day. All activity types occur at a relatively high frequency. Many of these activities are likely to be chained together. Additionally merging the ATCs 'Go to library: school related' and 'Transport someone' into a second group is also acceptable as this forms the second most optimal set of ATCs found.

An attempt was made to merge the original ATCs in a similar fashion as in other studies, in order to compare those approaches to the optimal set of ATCs found in this study. All tested combinations are inferior to the revealed optimal one, either by classification accuracy, by retained information or by both indices simultaneously.

Depending on the research, there might however exist a reason for employing one of the suboptimal schemes (e.g. some activity types *need* to be predicted and may not be merged, or a predefined number of ATCs is required). Yet, without such justification, this work suggests one should

strongly consider using the objectively determined most optimal set of ATCs of the NHTS 2009 in order to simultaneously maximize the prediction accuracy and the information in that prediction.

Future research will include also spatial and regional variables to apply the methodology to a big transport data activity type annotation problem. Furthermore, the application of data fusion based on annotated optimized ATCs will be investigated. Models based on traditional ATCs and optimized ones can be compared.

REFERENCES

1. Kitchin, R. Big Data and Human Geography: Opportunities, Challenges and Risks. *Dialogues in Human Geography*, Vol. 3, No. 3, 2013, pp. 262–267. https://doi.org/10.1177/2043820613513388.

2. Lee, W. Do, K. Choi, T. Bellemans, J. H. Hwang, and S. Cho. Data Mining Method for Smart Card Data Using Household Travel Survey: A Pilot Study of Public Transportation in Suwon, South Korea. 2015.

3. Feng, T., and H. J. P. Timmermans. Detecting Spatial and Temporal Route Information of GPS Traces. In *Geoinformatics for Intelligent Transportation* (I. Ivan, I. Benenson, B. Jiang, J. Horák, J. Haworth, and T. Inspektor, eds.), Springer International Publishing, Cham, pp. 61–75.

4. Lu, Y., and L. Zhang. Imputing Trip Purposes for Long-Distance Travel. *Transportation*, Vol. 42, No. 4, 2015, pp. 581–595. https://doi.org/10.1007/s11116-015-9595-0.

5. Montini, L., N. Rieser-Schüssler, A. Horni, and K. W. Axhausen. Trip Purpose Identification from GPS Tracks. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2405, 2014, pp. 16–23. https://doi.org/10.3141/2405-03.

6. 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. 13, No. 8, 2017, pp. 1–50. https://doi.org/10.1080/23249935.2017.1331275.

7. Ahern, A., G. Weyman, M. Redelbach, A. Schulz, L. Akkermans, L. Vannacci, E. Anoyrkati, and A. Von grinsven. *Analysis of National Travel Statistics in Europe*. 2013.

8. Reumers, S., F. Liu, D. Janssens, and G. Wets. The Annotation of Global Positioning System (GPS) Data with Activity Purposes Using Multiple Machine Learning Algorithms. In *Mobile Technologies for Activity-Travel Data Collection and Analysis* (S. Rasouli and H. Timmermans, eds.), IGI Global, pp. 119–133.

9. U.S. Department of Transportation, and Federal Highway Administration. 2009 National Household Travel Survey. http://nhts.ornl.gov.

10. NASA/WMAP Science Team. WMAP- Age of the Universe. https://map.gsfc.nasa.gov/universe/uni_age.html. Accessed Jul. 17, 2017.

11. Department of Economic Development; Jobs; Transport and Resources (DEDJTR). Victorian

Integrated Survey of Travel and Activity 2007. www.economicdevelopment.vic.gov.au/vista.

12. Department of Economic Development; Jobs; Transport and Resources (DEDJTR). Victorian Integrated Survey of Travel and Activity 2009. www.economicdevelopment.vic.gov.au/vista.

13. Cornelis, E., M. Hubert, P. Hyunen, K. Lebrun, G. Patriarche, A. De Witte, L. Creemers, K. Declercq, D. Janssens, M. Castaigne, L. Hollaert, and F. Walle. La Mobilité En Belgique En 2010 : Résultats de L'enquête BELDAM. *SPF Mobilité & Transports*.

14. Janssens, D., K. Declercq, and G. Wets. *Onderzoek Verplaatsingsgedrag Vlaanderen 4.5 (2012-2013)*. 2014.

15. Loechl, M. Stability of Travel Behaviour: Thurgau 2003. *Travel Survey Metadata Series*. Volume 16. http://archiv.ivt.ethz.ch/vpl/publications/tsms/tsms16.pdf.

16. Chalasani, V. S., and K. W. Axhausen. Mobidrive: A Six Week Travel Diary. *Travel Survey Metadata Series*. Volume 2. https://www.ethz.ch/content/dam/ethz/special-interest/baug/ivt/ivt-dam/vpl/tsms/tsms2.pdf.

17. Liikennevirasto - Finnish Transport Agency. *National Travel Survey 2010–2011*.

18. Armoogum, J., J.-P. Hubert, D. Francois, B. Roumier, M. Robin, and S. Roux. *Enquête Nationale Transports et Déplacements 2007-2008 (ENTD 2007-2008) (Rapport Technique)*. 2011.

19. Department for Transport. National Travel Survey, 2002-2014 [computer File]. 9th Edition.

20. Central Statistics Office. *National Travel Survey 2009*. Dublin, Ireland, 2011.

21. Metropolitan Transport Authority. *The Report of Household Travel Survey in Seoul Metropolitan Area [In Korean]*. Seoul, 2012.

22. Korea Transportation Institute. *National Transportation Demand Survey and Database Establishment in 2010: Passenger O/D Survey on the National Area*. 2011.

23. Centraal Bureau voor de Statistiek (CBS), and Rijkswaterstaat (RWS). Onderzoek Verplaatsingen in Nederland 2013 - OViN 2013. http://dx.doi.org/10.17026/dans-x9h-dsdg.

24. Klemenčič, M., M. Lep, B. Mesarec, and B. Žnuderl. Potovalne Navade Prebivalcev v Mestni Občini Ljubljana in Ljubljanski Urbani Regiji.

25. Trafik Analys. *RVU Sverige 2011–2014 - Den Nationella Resvaneundersökningen (RVU Sweden 2011-2014 - National Travel Survey)*. Stockholm, 2015.

26. Department for Transport. National Travel Survey, 2002-2015. *UK Data Service*.

27. Kusakabe, T., and Y. Asakura. Behavioural Data Mining of Transit Smart Card Data: A Data

Fusion Approach. *Transportation Research Part C: Emerging Technologies*, Vol. 46, 2014, pp. 179–191. https://doi.org/10.1016/j.trc.2014.05.012.

28. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software. *ACM SIGKDD Explorations*, Vol. 11, No. 1, 2009, pp. 10–18. https://doi.org/10.1145/1656274.1656278.

29. Lee, S., and M. Hickman. Trip Purpose Inference Using Automated Fare Collection Data. *Public Transport*, Vol. 6, No. 1-2, 2014, pp. 1–20. https://doi.org/10.1007/s12469-013-0077-5.

30. Kochan, B. *Implementation, Validation and Application of an Activity-Based Transportation Model for Flanders*. 2012.

31. Bradley, M., and P. Vovsha. A Model for Joint Choice of Daily Activity Pattern Types of Household Members. *Transportation*, Vol. 32, No. 5, 2005, pp. 545–571. https://doi.org/10.1007/s11116-005-5761-0.

32. Shen, L., and P. R. Stopher. A Process for Trip Purpose Imputation from Global Positioning System Data. *Transportation Research Part C: Emerging Technologies*, Vol. 36, No. November, 2013, pp. 261–267. https://doi.org/10.1016/j.trc.2013.09.004.

33. Lu, Y., S. Zhu, and L. Zhang. Imputing Trip Purpose Based on GPS Travel Survey Data and Machine Learning Methods. *Transportation Research Board*, Vol. 1250, 2013.