THE USEFULNESS OF THE SEQUENCE ALIGNMENT METHODS IN VALIDATING RULE-BASED ACTIVITY-BASED MODELS

George Sammour, Tom Bellemans, Koen Vanhoof, Davy Janssens, Geert Wets*

Transportation Research Institute (IMOB)
Faculty of Applied Economics
Hasselt University
Wetenschapspark 5 Box 6
B-3590 Diepenbeek
Belgium

Fax: +32 11 26 91 99
Tel.: +31 11 26 91 58

Email: 
{george.sammour,tom.bellemans,koen.vanhoof,davy.janssens,geert.wets}@uhasselt.be

* Corresponding author

Number of Words = 5994
Number of Tables = 3
Number of Figures = 3
Words counted: 5994 + 3 * 250 + 3 * 250 = 7494 words
ABSTRACT

The aim of this paper is to achieve a better understanding of rule-based activity-based models, by proposing a new level of validation on the process model level in the ALBATROSS model. To that effect, the work activity process model, which includes six different decision steps, is investigated. Each decision step is evaluated during the prediction of individuals’ schedules. The comportment of execution in the process model contains activation dependency. This branches the execution and evaluation of each agent under examination. And yields a sequence of decisions for each agent, where the Sequence Alignment Method (SAM) is employed to evaluate how similar/dissimilar predicted with observed decision sequences are. SAM utterly fits for assessing the analysis of decision sequences on this level. The original CHAID decision trees at each decision step utilized in ALBATROSS are compared with other well known induction methods chosen to appraise the purpose of the analyses. Additionally, the performance of the models is compared at three existing validation levels: the classifier or decision step level using confusion matrix statistics. The work activity trips Origin-Destination (OD) matrix level and time of day work activity start time level, using the correlation coefficient. The results of validation on the proposed process model level show conformance to those already existing, with additional information to help in better understanding the process model’s behaviour.

INTRODUCTION

In the past few decades, many studies have been conducted in order to try to understand the nature of travel demand. Travel demand is derived from the human needs to participate in activities that are distributed in time and space. Models that simulate travel demand using an activity-based approach have been gaining growing attention in recent times due to their strong behavioral foundation and insightful theoretical demand. Recognizing that travel is a demand derived from individuals' needs to perform activities, researchers in travel demand modeling have become increasingly interested in analyzing and predicting individuals' decisions about activity participation. Activity-scheduling models share the objective to predict the sequence of decisions that leads to an observed activity pattern of households/individuals. Activity-based models aim at predicting on a daily basis and for individuals which activities are conducted, by whom, for how long, at what time, the location, and which transport mode is used when traveling is involved (1). The data requirements for activity-based models are in general demanding compared to conventional travel demand models. This is obvious specially that this type of micro-simulation models should be able to predict the travel behaviour in detail including how the activities are selected and scheduled. And so the validation of behavioural models becomes a difficult task. Rule-based activity-based models are no exception as the validation process can be performed on several levels hence, validating the model on an additional or new level may incur extra knowledge to further calibrate and improve its performance. An existing and fully operational rule based activity based model is the ALBATROSS model (1); it is a computational process model, where schedules are predicted using CHAID based induction tree method.

The validation of the ALBATROSS model is performed on many different levels, in their original work (1) considered model performance on three levels: (i) the choice facet or
the decision tree induction level, by measuring the predictive accuracy of each decision rule
in the scheduling process. (ii) At the activity pattern level, sequence alignment methods are
used to assess the correspondence between the observed and predicted activity sequences (2).
(iii) At the trip matrix level, using correlation coefficients calculated to measure the degree of
correspondence between the observed and the predicted Origin-Destination matrices.
Decision trees derived from survey data may become large, complex and difficult to
interpret. In several experimental and analytical studies using the ALBATROSS model,
examples as in (4) and (5) performed validation on three levels, choice facet, activity pattern
and trip matrix levels.

The objective of this study is to investigate and assess the performance and predictive
behaviour of activity-based models on the decision process level of rule based activity based
models. The process model level is a core component of the scheduler engine in
ALBATROSS, which may reveal extra information on the model. And consequently assess
in more understanding the effect of using a specific induction method and in return improve
model performance. By further analyzing the process model the sequence alignment method
(SAM) was selected to measure how similar predicted to observed decision sequences are.
To this end, this work attempts to prove that evaluating activity-based models on this new
level expose information helps in additional understanding of the model.

The remaining part of this paper is organized as follows, in the next section the
ALBATROSS model and the FEATHERS framework used to implement the model for
Flanders are described, followed by a discussion of the diary data used for training the
model. The analyses and the process model are further discussed explaining the induction
methods and elaborating on the usefulness of adapting SAM in process models. Then
experiments design and discussion of results are discussed, followed by the conclusion and
future works.

THE FEATHERS / ALBATROSS SYSTEM

ALBATROSS is a fully operational rule based activity-based model that incorporates
household-level decision making (1) (4). In ALBATROSS, rules are used to predict activity-
travel choices of individuals and households. The decision rules are formalized from the
training of decision trees by using a CHAID decision tree induction method on surveyed
activity-travel diary data. In ALBATROSS, to generate a schedule for each person for each
day a sequential decision process is assumed, in which the rules are derived from 26 decision
trees, and the activity scheduling process model consists of four components or sub models
(14). The first component is responsible for generating primary work activities and their start
time, duration of each work episode if more than one episode is predicted, and their location,
and finally the transport mode for the work trip. The second component is used to generate
secondary fixed activities, usually work-related such as bring/get, business or other
mandatory activities. In addition it decides which type of activities performed, the number of
episodes for each activity, and their start time and duration. The third component is similar to
the second component, except it determines the flexible activities part of the schedule. The
fourth and last component is in charge of predicting the transport mode of secondary fixed
and flexible activities, as the transport mode of the primary work is already decided by the
first component. It is important to note that in ALBATROSS the activity travel behavior of
the two heads only is captured. A full account of the Albatross model system is given in (1).
The analysis performed in this work is performed on the first component dealing with work activity scheduling excluding the transport mode decision step. Figure 1 depicts the work activity decision process model used in ALBATROSS. Each numbered rectangle refers to a decision tree model derived from activity diary data. The index $j$ used in the figure refers to the number of work episodes, if more than one work activity episode is predicted.

The first decision step evaluates whether the individual’s schedule contains a work activity, if so, the duration of the work activity is predicted next. Followed by the number of work activity episodes, subsequently the ratio between work episodes and the break time duration is decided. And finally the work activity start time is predicted. Decision steps 1 and 3 are discrete choice decisions, whereas, decision steps 2, 4, 5 and 6 are continuous choice decisions. It is noteworthy that if decision step 1 infers no work episode for the individual under consideration then decision steps 2-6 will not be executed. Similarly, if decision step 3 evaluates to not including a second work episode, then decision steps 4 and 5 will not be evaluated. This implies that there is an activation dependency in the execution of this process model.

The analysis in this work is developed within the FEATHERS (Forecasting Evolutionary Activity-Travel of Households and their Environmental RepercussionS) framework (6). The FEATHERS framework is developed to facilitate the development of modular activity-based models for transportation demand in Flanders (Belgium). The scheduling engine that is currently implemented in the FEATHERS framework is based on the scheduling model that is present in the ALBATROSS system (1). The framework is fully operational at the level of Flanders. The scheduling is based on CHAID decision trees (8), trained based on the Onderzoek VerplaatsingsGedrag Vlaanderen (OVG) travel survey data. The modular design of FEATHERS allows for ease of adaptation of classification methods other than CHAID decision trees, such as Bayesian networks (9), simple classifiers (5), and association rules (10). Taking the above in account, the analysis in this work was conducted based on the ALBATROSS model that was implemented in the FEATHERS framework. However, for research purposes the FEATHERS framework is extended to conduct experiments using alternate induction methods, such as decision trees, logistic regression and OneR (11) (work is still going on to add more methods). This additional functionality allows one to train models outside FEATHERS, using data mining packages that can export Predictive Model Markup Language (PMML) (12). PMML is an XML based language to annotate data mining model parameters in textual form with meta-data for re-use. And thus, using this functionality, the CHAID induction method was replaced by alternatives such as C45 (13), Logistic regression and OneR (12) then integrated within the scheduling model.

**FLEMISH ACTIVITY TRAVEL DIARY DATA FOR MODEL TRAINING**

The data sets used for training the models in the work activity process model and all the 26 decision trees originates from the OVG survey. The survey is a trip-based survey method. The travel survey was conducted based on a random sample from the national register. These persons involved in a survey that was perform primarily through face-to-face interviews. Table 1 shows the situational and socio-demographic variables that are used as prediction variables in FEATHERS/ALBATROSS.

The variables that relate to the household level attributes are urban density, household composition, the presence of youngest children in the household, socio-economic class, and car ownership. The gender, driver license, work status and work status of the person’s partner
are variables related to the individual attributes. In addition, variables such as, the number of employees with daily-good and non-daily good, number of households within a specific distance from home location of a household, the distances (in decameters) of the nearest daily and non-daily good sector and the nearest distance of employees within a ranges are related to the measures of accessibility given the home location of the household.

![Work activity process model in ALBATROSS, adapted from (7).](image)

Finally, variables labeled with a (*) are captured and kept from previous decisions and included in the next decision step, only during the decision process. Continuous variables such as duration, duration Ratio, break time duration and start time of work activity episodes are discretised by using Equal Frequency Interval (EFI) method. A 70-30% training-test split was made on the data. As mentioned above, the datasets for decision steps 1 and 3 are discrete choice models; with minority class is 28% and 13% respectively.
TABLE 1 Work activity pattern datasets description

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urb</td>
<td>Urban density</td>
<td>0: highest density, 4: lowest density</td>
</tr>
<tr>
<td>Comp</td>
<td>Household composition</td>
<td>0: single without children, 1: single with children, 2: single with parents, 3: partner without children, 4: partner with children</td>
</tr>
<tr>
<td>Child</td>
<td>Presence of the youngest children</td>
<td>0: no children, 1: &lt;6, 2: 6-12, 3: &gt;12 years</td>
</tr>
<tr>
<td>Day</td>
<td>Day of the week</td>
<td>0: Monday to 6: Sunday</td>
</tr>
<tr>
<td>pAge</td>
<td>Age category</td>
<td>0: &lt;35, 1: 35-55, 2: 55-&lt;65, 3: 65-&lt;75, 4: &gt;75 years</td>
</tr>
<tr>
<td>Ncar</td>
<td>Number of cars in household</td>
<td>0: no cars, 1: 1 car, 2: 2 or more cars</td>
</tr>
<tr>
<td>Gend</td>
<td>Gender</td>
<td>0: female, 1: male</td>
</tr>
<tr>
<td>Driver</td>
<td>Driving license of person</td>
<td>0: is not a driver, 1: is driver</td>
</tr>
<tr>
<td>wstat</td>
<td>Work status of person</td>
<td>0: no work, 1: part time, 2: full time</td>
</tr>
<tr>
<td>Pwstat</td>
<td>Work status of person’s partner</td>
<td>0: no work, 1: part time, 2: full time</td>
</tr>
<tr>
<td>Xdag</td>
<td>Number employees daily-good sector within 3.1 km from home</td>
<td>0: &lt;0,115], 1: &lt;115,253], 2: &lt;253,307], 3: &lt;307,507], 4: &lt;507,675], 5: &gt;675</td>
</tr>
<tr>
<td>Xn-dag</td>
<td>Number employees non-daily-good sector within 4.4 km from home</td>
<td>0: &lt;0,395], 1: &lt;395,635], 2: &lt;635,762], 3: &lt;762,938], 4: &lt;938,2525], 5: &gt;2525</td>
</tr>
<tr>
<td>Xarb</td>
<td>Number employees within 4.4 km from home</td>
<td>0: &lt;0,8785], 1: &lt;8785,12995], 2: &lt;12995,16120], 3: &lt;16120,20199], 4: &lt;20199,70314], 5: &gt;70314</td>
</tr>
<tr>
<td>Xpop</td>
<td>Number households within 3.1 km from home</td>
<td>0: &lt;0,5050], 1: &lt;5050,8845], 2: &lt;8845,13217], 3: &lt;13217,16833], 4: &lt;16833,22884], 5: &gt;22884</td>
</tr>
<tr>
<td>Ddag</td>
<td>Distance (dm) to nearest 160 employees daily-good sector</td>
<td>0: &lt;0,71], 1: &lt;71,127], 2: &lt;127,165], 3: &lt;165,202], 4: &lt;202,346], 5: &gt;346</td>
</tr>
<tr>
<td>Dn-dag</td>
<td>Distance (dm) to nearest 260 employees non-daily-good sector</td>
<td>0: &lt;0,92], 1: &lt;92,145], 2: &lt;145,176], 3: &lt;176,258], 4: &lt;258,334], 5: &gt;334</td>
</tr>
<tr>
<td>Darb</td>
<td>Distance (dm) to nearest 4500 employees total</td>
<td>0: &lt;0,92], 1: &lt;92,128], 2: &lt;128,201], 3: &lt;201,274], 4: &lt;274,360], 5: &gt;360</td>
</tr>
<tr>
<td>Dpop</td>
<td>Distance (dm) to nearest 5200 households</td>
<td>0: &lt;0,0], 1: &lt;0,105], 2: &lt;105,126], 3: &lt;126,163], 4: &lt;163,278], 5: &gt;278</td>
</tr>
<tr>
<td>Dur*</td>
<td>Total duration (min.) of work activity</td>
<td>0: &lt;0,395], 1: &lt;395,495], 2: &lt;495,526], 3: &lt;526,565], 4: &gt;565</td>
</tr>
<tr>
<td>Nep*</td>
<td>Number of work episodes</td>
<td>0: one, 1: two</td>
</tr>
<tr>
<td>Ratio*</td>
<td>Ratio (%) between first and second work episodes. Duration (min.) of break time between first and second work episodes</td>
<td>0: &lt;0,40], 1: &lt;40,48], 2: &lt;48,52], 3: &lt;52,60], 4: &gt;60</td>
</tr>
<tr>
<td>Inter*</td>
<td></td>
<td>0: &lt;0,25], 1: &lt;25,47], 2: &lt;47,60], 3: &lt;60,95], 4: &gt;95</td>
</tr>
</tbody>
</table>

* Included only if known in stage of the decision process.

ANALYSIS

To be able to analyse the behaviour of the work activity process model only decision steps 1 and 3 are replaced by alternative classification methods. Because at these decision steps the execution pattern of the process model is affected. While the continuous decision steps (2, 4,
5 and 6) are kept unchanged using the original CHAID based tree induction. The analysis was performed using four different induction methods that are appropriate for assessing the proposed validation level. The first method is the original CHAID tree method. The second technique is the C45 decision tree method for two reasons, (a) C45 is a benchmarking method in the data-mining community, (b) in a case study, Wets et al (16) found approximately equal performance of CHAID and C45 decision tree algorithms in terms of goodness of fit. The third technique is the Logistic Regression classification method, which will be referred to as Logit throughout this paper. The Logit method was selected because it generally outperforms decision tree methods in terms of classification accuracy, especially for small size data sets, as shown by (17). Moreover, Logit can produce probability estimates. The fourth and last method is OneR induction, which is a very simple classifier that provides a rule based on the value of a single attribute. And given the unbalanced nature of the discrete class data sets it is expected that this method will be biased towards the majority class.

In the next subsections the induction methods used in the analyses are described, followed by an elucidation of the SAM similarity measure. And in the next section the proposed validation method on the process model level is discussed in details.

**Decision Tree induction methods general concepts**

Decision trees are techniques which are used to make decisions from a set of training cases. To use a decision tree for prediction, a rule is specified that assigns a class of the condition attribute to each case classified by the tree. ALBATROSS uses a probabilistic action-assignment rule, for both discrete and continuous choice induction, instead of a deterministic assignment rule, because this results in a better prediction of the aggregate distributions. And so, each rule is assigned a probability distribution that is derived from the frequency distribution over the classes of the condition attribute in the training set for each leaf. An important issue in decision tree learning is over-fitting. The concept of over-fitting occurs when the induction algorithm generates a decision tree that perfectly fits the data in the training data set but lacks the capability of generalization of instances not present in the training set. To avoid over-fitting the minimum number of cases at leaf nodes was set to 30 for both CHAID and C45 decision tree models (18).

**The CHAID decision tree**

The CHAID was introduced by (8), it originated from the automatic interaction detection (AID) method. The CHAID based induction tree method is able to generate trees with more than two branches attached to the same node at any level of the tree and mainly suited for the analysis of large data sets. It is based on the chi-squared ($\chi^2$) statistic to identify the best split of the data set on condition variables into homogenous partitions with respect to the class variable. In addition the CHAID based tree induction method allows for specifying a threshold ($\alpha$) for splitting based on the significance level and the minimum number of cases at leaf nodes. The tree building algorithm is performed by recursively iterating through the condition variables to test for each variable the pair of categories whether there is no statistically significant difference within the pair with respect to the class variable. The split with the highest significance value across condition variables is selected. This procedure is
repeated until no significant splits are found or the maximum number of cases at leaf nodes is reached.

C4.5 decision tree

There are two stages for building a classification decision tree in the C4.5 algorithm (11). The first stage involves generating the decision tree based on the training data set, where the second stage has to do with pruning the decision tree based on a validation or test data set that is left out from the training set. The algorithm works as follows. Assume we have a data set \( S \) of training cases or samples, where each case consists of \( n \) condition or explanatory variables \( x_1, x_2, \ldots, x_n \) and a class or response variable \( C_i \), for \( i = \{1, 2, \ldots, p\} \) classes. C4.5 first grows an initial tree using the divide-and-conquer technique by splitting the training set into homogeneous subsets \( S_1, S_2, \ldots, S_p \), until the leaf nodes contain only cases from a single class. An important issue in learning classification trees is over-fitting on the data. Therefore to avoid over-fitting C4.5 adopts a pruning strategy, where the decision tree is simplified by removing one or more sub-trees and replacing them with leaves. For a detailed description, the interested reader is referred to (11).

Logistic regression

Logistic regression (18), sometimes referred to as Logit, is an alternative regression technique naturally suited to categorical data. Logit fits an S-shaped curve to the data. Let \( X, Y \) be a dataset with a binary response or class variable, where \( X \) is a vector of \( k \) independent variables \( (x_1, x_2, \ldots, x_k) \) for each case \( x_i \) in \( X \) the response or dependent variable is either \( y_i = 1 \) or \( y_i = 0 \) then, the logistic model predicts the Logit of \( Y \) from \( X \). The Logit is the natural logarithm \( ( \ln ) \) of odds of \( Y \), and odds are ratios of probabilities \( \pi \) of \( Y \) happening (i.e., a work activity exists in an individual’s schedule at a specific day) to probabilities \( (1 - \pi) \) of \( Y \) not happening (i.e., a work activity does not exists in an individual’s schedule). The simple logistic model has the following form:

\[
\text{logit}(Y) = \ln \left( \frac{\pi}{1 - \pi} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{1}
\]

Where \( \ln \) is the natural logarithm, \( \pi \) is the probability of the class variable \( Y = 1 \), \( \alpha \) is the \( Y \) intercept, and \( \beta_1, \beta_2, \ldots, \beta_k \) are the regression coefficients. The probability \( (\pi) \) that the class variable \( Y = 1 \) is computed by:

\[
\pi(y = 1) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)} \tag{2}
\]

The \( \alpha \) and \( \beta_1, \beta_2, \ldots, \beta_k \) are typically estimated by the maximum likelihood method.

One R

One R is a very simple classifier that provides a rule based on the value of a single attribute. According to (12) the algorithm may compete with state-of-the-art techniques used in the field (12). Similar to other algorithms, One R takes as input a set of several attributes and a
class variable. Its goal is to infer a rule that predicts the class given the values of the attributes. The One R algorithm chooses the most informative single attribute and bases the rule exclusively on this attribute. Full details can be found in (12). The algorithm assumes that the attributes are discrete. If not, they must be discretised.

THE USEFULLNESS OF SAM FOR THE WORK ACTIVITY PROCESS MODEL

The Sequence Alignment Methods (SAM)

The work related to sequential analysis of activity patterns in activity-based models reached a new milestone, when the Sequence Alignment Method (SAM) was introduced in transportation research by Wilson (20). The interesting characteristic of the SAM is that it makes use of biological distance rather than geometric (Euclidean) distance as the basic concept of comparison (21). Mainly in Activity-based models, the SAM methods are used to measure the goodness of fit, in terms of how similar/dissimilar the observed and the predicted activity sequences are. This is done by calculating the effort required to make the two sequences identical using insertion, deletion, and substitution operators. Insertion and deletion operations require the same cost of one unit, whereas substitution requires twice that cost. The lower the SAM measure, the more similar the two sequences are. In the context of this work, the SAM measure will be used on the process model level rather than the activity pattern level. The approach in which the SAM is adopted on the process level and the rationale behind choosing SAM is explained in the next section.

The adaptation of SAM on the decision process model level

The validation of the ALBATROSS model, as mentioned in the introduction, is performed on mainly three levels, the choice facet or the decision tree induction level, the activity pattern level, and the trip (O-D) matrix level. These levels provide goodness-of-fit measures either on individual classifiers, or on the system outputs. However, they do not provide information on the activation dependency and its effect on the model’s performance. Therefore, to be able to assess and analyze the behaviour of the decision process model in ALBATROSS, a validation method on the process model level is required. And a measure was needed to appraise the quality of prediction at each decision step. Considering the characteristics of decision outcomes at each decision step in the work activity process model, as shown in Figure 2, the process actually output a sequence of decision outcomes or as will be called in the remainder of the paper the decision sequence. Thus, the SAM measure is the best fit for the purpose of assessing the validity on the process model level. The generation of the decision sequences involves the following definitions and assumptions, for each individual:

- A predicted decision outcome sequence \( [D1 D2 D3 D4 D5 D6]_{\text{Pred}} \) is generated, and similarly,
- An observed sequence \( [D1 D2 D3 D4 D5 D6]_{\text{Obs}} \) that is extracted from diary data is generated accordingly.
- The length of the predicted and observed sequences can be 1, when no work activity inclusion, 3 when only one work episode is conducted or 6 when two work activity episodes are captured.
Another point of concern, with regard to the proposed level of analysis is the approach in which the SAM measure is calculated. Will the SAM measure be calculated on a one-to-one (on the single decision step level)? All-to-all, taking the whole decision sequence after the process model finishes execution? Or in a stepwise manner, which entails calculating the SAM after each decision step taking in account the previous decision outcome, as the process model is executing.

Using the one-to-one approach, the SAM distance will be measured for each decision step separately. This will serve as an accuracy measure for the individual decision step or the classifier level itself. Moreover, in the all-to-all approach, only one SAM distance is measured, which indicates how similar the two decision sequences are. Nevertheless, using this approach will not capture the activation dependency behaviour. And finally, using the stepwise approach, the SAM is evaluated after each decision step keeping the previous decision. And this entails that at each decision step the SAM distance is measured for the observed and predicted decision sequences preserving previous decision symbols as the execution of the process model continues.

FIGURE 2 Work activity process model decision outcomes in ALBATROSS.

DESIGN OF EXPERIMENTS AND RESULTS

The aim of this study is to validate and assess the performance of activity based models on the process model level, and further validate that the proposed method on three existing validation levels, the classifiers’ level, the work activity Origin-Destination (OD) matrices
level (spatial resolution), and the work activity start time distribution throughout the day (temporal resolution). This will allow for assessing the performance of the work activity process model. The C45 approach was trained using WEKA’s J48/C4.5 implementation. The OneR approach was also trained using WEKA. The Logit models were trained using the Rattle package for R (22). The models were exported to PMML and a decisionMaker class is implemented in the FEATHERS framework to deploy PMML decision trees as well as Logit models. The experiments were setup by running FEATHERS for the simulation of cases and generating schedules for both the training and test sets in four different settings, where in each setting a different classifier for decision steps 1 and 3 is used for prediction of work activities in the process model for each day.

Work activity process model level accuracy analysis

The analyses on the process level were conducted by capturing the decision output at each decision step and calculate the stepwise SAM distance between predicted and observed decision sequences. This implies that the decision sequence grows in length (depending on the activation dependency) as the execution of decision steps continues. So the amount of increase in the SAM distance within the same model approach points out the effect of a decision step on the previous decision step. The average length of the observed sequence is 1.9 (1.4) symbols with standard deviation between brackets, whereas for CHAID, C45, Logit and OneR the average lengths are 1.8 (1.5), 1.9 (1.5), 2.04 (1.6), and 1 (0) respectively. The average length of the OneR approach is 1 with a variance of 0 because the model always predicts no work activity and hence the decision sequence contains only one symbol. It is observed that the CHAID and C45 approaches predict similar decision sequence lengths. On the other hand, the Logit approach predicts longer decision sequences, note that is due to the activation dependency execution of decisions, and in order to measure the similarity between predicted and observed decision sequences requires more effort in terms of deletion.

Figure 3 depicts the stepwise SAM distance for the training and test sets, at each decision step in the work activity process model represented in a line chart. The chart illustrates that the CHAID and C45 approaches reported similar performance and behaviour. Despite the fact that at decision step 1 the Logit approach performed best, the CHAID and C45 reported a close decision sequence incremental SAM distance starting from decision step 2, which means that when evaluating the two steps all together, decision trees outperform the Logit approach.

Classifier level
Discrete choice models

The evaluation criteria of the discrete choice models are presented using two accuracy measures, the confusion matrix (also called contingency table) accuracy measure, since both discrete choice classifiers are binary. And the Brier score (23) because of the probabilistic action assignment rule used in scoring the models.

The confusion matrix records correctly and incorrectly recognized examples for each class. The following accuracy statistics can be derived from the confusion matrix:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]
Sensitivity = \frac{TP}{TP + FN} \hspace{1cm} (4)

Specificity = \frac{TN}{TN + FP} \hspace{1cm} (5)

F - Measure = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \hspace{1cm} (6)

FIGURE 3 Stepwise SAM distances for the work activity process model.
Where, TP: number of true positive values, FP: the number of false positive values, TN: number true negative values and FN: false negative values. The precision in the F-Measure can be computed as: \( \text{precision} = \frac{TP}{TP+FP} \). Accuracy is not a preferred performance measure for imbalanced datasets (24). When working with a high imbalance, a classifier classifying everything as a majority class sample will result in a high predictive accuracy. Sensitivity approximates the probability of the positive class being correctly classified, and specificity estimates the probability of correctly predicting the negative class. The F-measure focuses more on the dropout class by consideration of sensitivity and precision as it is the weighted average of the precision and recall. An F-measure value reaches its best value at 1 and its worst value at 0.

The Brier score (BS) is a metric related to the mean-squared-error often used in statistical fitting as a measure of model goodness. It is a descriptive measure often used in the literature on prediction accuracy. The Brier score is calculated as follows:

\[
BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2
\]  

Where \( p_i \) is the predicted probability and \( o_i \) is the observed value of the instance \( i \) (0 if negative and 1 if positive). The BS measures the average squared deviation between predicted probabilities for a set of events and their outcomes. So a lower score represents a higher accuracy. Table 2 provides the results of the analysis to assess model performance. The results suggest that for decision step 1, the Logit model outperforms all other methods specially in predicting the positive class value (\( yWo \)). As expected, CHAID and C45 show similar performance with a slight increase in performance in favor of C45. The predictive performance (sensitivity) for the (\( yWo \)) class variable, which is the minority class, is notably higher in the Logit approach and this can be explained by the fact that Logit outperforms decision tree approaches for small size datasets (18). The OneR approach prediction outcome was always no work since the distribution of the class variable in this dataset is skewed (72% no work), with a Brier score equal to the percentage of the minority class in the dataset. Results also suggest that the drop in the accuracy in the test set was not significant, while there was a slight increase in accuracy for the CHAID approach.

Considering the performance of decision step 3, again CHAID and C45 confirmed similar performance but outperform the Logit and OneR approaches, the reason for the weaker performance of the Logit approach is that the data set at decision step 3 is highly skewed 87% and this leads to underestimating the rare class calculated by Equation 2 as reported by (24). Finally the OneR model always predicts the majority and so the predictive power of the minority class is zero. The NA in the OneR approach indicates that the measure cannot be computed since the TP and FP values used to calculate the precision for this approach are zero.

### Continuous choice models

The continuous choice models where trained using only the CHIAD tree induction method used originally in ALBATROSS were kept the same for the analyses performed using alternative discrete choice models. The performance of continuous choice models was assessed by means of the Relative Absolute Error (RAE) which gives an indication of how
good a predicted value is relative to the observed value. The reason for selecting this measure is that it can be reported as a percent error measure for numeric or continuous predictions. The RAE is calculated by dividing the sum of the absolute difference between the predicted and observed values by the observed cases. Results showed fairly good results with 21%, 22.4% and 9% for decision steps 2, 4 and 6 respectively for training sets, and 20%, 20.4% and 10% for test sets, while for decision step 5 the RAE reported 64% for training and 61% for test set.

**TABLE 2** Accuracy statistics for discrete choice models (classifier level)

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>Brier Score</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td>CHAID</td>
<td>0.11766</td>
<td>0.54065</td>
<td>0.841026</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>0.113781</td>
<td>0.813008</td>
<td>0.839448</td>
</tr>
<tr>
<td></td>
<td>C45</td>
<td>0.114957</td>
<td>0.59248</td>
<td>0.84497</td>
</tr>
<tr>
<td></td>
<td>OneR</td>
<td>0.279625</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>Brier Score</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td>CHAID</td>
<td>0.112366</td>
<td>0.554371</td>
<td>0.851653</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>0.115959</td>
<td>0.791045</td>
<td>0.83628</td>
</tr>
<tr>
<td></td>
<td>C45</td>
<td>0.115108</td>
<td>0.556503</td>
<td>0.825519</td>
</tr>
<tr>
<td></td>
<td>OneR</td>
<td>0.264972</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>Brier Score</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td>CHAID</td>
<td>0.106202</td>
<td>0.1875</td>
<td>0.890315</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>0.0979974</td>
<td>0.125</td>
<td>0.985998</td>
</tr>
<tr>
<td></td>
<td>C45</td>
<td>0.108202</td>
<td>0.242188</td>
<td>0.866978</td>
</tr>
<tr>
<td></td>
<td>OneR</td>
<td>0.129949</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>Brier Score</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td></td>
<td>CHAID</td>
<td>0.136115</td>
<td>0.155844</td>
<td>0.885496</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>0.144773</td>
<td>0.0649351</td>
<td>0.959288</td>
</tr>
<tr>
<td></td>
<td>C45</td>
<td>0.134237</td>
<td>0.194805</td>
<td>0.903308</td>
</tr>
<tr>
<td></td>
<td>OneR</td>
<td>0.1638</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Work activity trip matrix and trips start time level accuracy analysis (spatial and temporal resolutions)**

At the work activity trip matrix level (spatial resolution), the observed and predicted OD matrices, for training and test sets, were compared. An activity OD matrix contains the frequency of work activity trips for each combination of origins (rows) and destinations (columns). The frequency of trips at each zone in Flanders was aggregated forming a one dimensional array with work activity trip counts at each zone. The correlation is calculated between observed and predicted matrix entries $\rho$(observed, predicted).

The work activity start time level (temporal resolution) was also analysed by calculating the correlation between the observed and predicted work activity start times for each hour of the day, the reason this analysis was conducted to further investigate the larger increase in the SAM distance on the process model level at decision step 6 (Start time) as can
be observed in Figure 2. The indication of NA in Table 3 for the OneR approach indicates that the correlation is not available since no work activities were predicted using this approach. The results in Table 3 indicate that the correlation coefficients are similar with the Logit approach having a slightly lower correlation coefficient than the CHAID and C45 approaches.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CHAID</th>
<th>Logit</th>
<th>C45</th>
<th>OneR</th>
<th>CHAID</th>
<th>Logit</th>
<th>C45</th>
<th>OneR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.832</td>
<td>0.802</td>
<td>0.835</td>
<td>NA</td>
<td>0.896</td>
<td>0.873</td>
<td>0.899</td>
<td>NA</td>
</tr>
<tr>
<td>Test</td>
<td>0.816</td>
<td>0.799</td>
<td>0.82</td>
<td>NA</td>
<td>0.827</td>
<td>0.771</td>
<td>0.803</td>
<td>NA</td>
</tr>
</tbody>
</table>

CONCLUSION AND FUTURE WORK

From a data mining perspective, rule based activity based models are validated on mainly three major levels, namely, the classifier accuracy level, where single rules are evaluated and analysed, the generated activity pattern level using the sequence alignment (SAM) distance measure by calculating how similar the observed and the predicted activity sequences are. And the trip matrix level by assessing the correlation coefficient to measure the degree of correspondence between the observed and the predicted origin-destination matrices. The work reported in this study proposed a methodology to validate rule-based activity-based models on the process model level. The proposed analyses suggested that conducting an investigation on the process model level, provides additional information on how the model performs when using a specific classifier at a specific decision step. The results obtained from the analyses, conform to other levels of validation. Plus extra information indicating that, despite the fact that a classifier’s predictive performance is compelling, yet the activation dependency of the process model affects the overall model performance and accuracy. Additionally, the results showed that the branching of decision steps at 1 and 3 is a critical issue for the outcome of the model. And perhaps changing the order of such decisions might lead to a better model.

Future work will be directed towards approaches related to changing the order of decision steps in the process model. And training the models with and without the inclusion of additional features in the training data sets in subsequent models and investigate the performance and behaviour of the model for each setting.

REFERENCES


