Activity-Based Travel Demand Modeling Framework FEATHERS: Sensitivity Analysis with Decision Trees [Link](#) 

Peer-reviewed author version

Made available by Hasselt University Library in [Document Server@UHasselt](#)

**Reference** (Published version):
Bao, Qiong; Kochan, Bruno; Shen, Yongjun; Bellemans, Tom & Janssens, Davy (2016) Activity-Based Travel Demand Modeling Framework FEATHERS: Sensitivity Analysis with Decision Trees. In: Travel Demand Forecasting, Volume 2, p. 89-98.

DOI: 10.3141/2564-10
Handle: http://hdl.handle.net/1942/21120
Sensitivity analysis on decision trees in activity-based travel demand modeling framework FEATHERS

Qiong Bao*, Bruno Kochan, Yongjun Shen, Tom Bellemans, Davy Janssens, and Geert Wets

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

E-mails:
{qiong.bao, bruno.kochan, yongjun.shen, tom.bellemans, davy.janssens, geert.wets}@uhasselt.be

*Corresponding author
Tel: +32(0)11269146
Fax: +32(0)11269198

Word count: 5224+9*250=7474
Number of Figures: 5
Number of Tables: 4
Submission date: 14/03/2016
Abstract
The technique of decision trees is increasingly applied in activity-based travel demand modeling. It owns the strength of representing the full complexity of interactions between different variables. However, this complexity on the other hand often hinders an interpretation in terms of the relative impacts of these variables on the activity travel choice. In this study, a sensitivity analysis is performed on decision trees in FEATHERS, an activity-based micro-simulation modeling framework, with the purpose of quantitatively measuring the relative impact of input variables involved in the given decision trees on the choice variable. Both of the local and global sensitivity analysis approaches are investigated: i) a one-at-a-time approach which predicts the choice frequency distribution by varying selected input variables one after another, and keeping all other variables as observed; and ii) the improved Sobol’ method which evaluates the effect of an input variable while all other variables are varied as well. By applying these two approaches to two representative decision trees concerning work-related activity (i.e., commute trip) choice and transport mode choice for work-related activities in the FEATHERS framework, consistent results about the key input variables for these two decision trees are derived, and some extra insights are gained from each of these two approaches.
1 INTRODUCTION
Decision trees, which can be used to reproduce human decision making concerning daily activity plans, have gained increased interest to modeling spatial-temporal choice behavior \((1)\). Different from the discrete choice models that predict individual’s activity and travel choices mainly based on utility maximization \((2,3)\), the computational process models such as decision trees assume that choice behavior is driven by heuristic decision rules. Thus, a heuristic rule does not necessarily produce optimal outcomes, but rather actions that have yielded satisfactory outcomes under similar conditions in the past \((4)\). FEATHERS (the Forecasting Evolutionary Activity-Travel of Households and their Environmental RepercussionS), developed by the Transportation Research Institute of Hasselt University, is an activity-based micro-simulation modeling framework currently implemented for the Flanders region of Belgium \((5)\). In this framework, a rule-based approach based on a sequence of decision trees is used in the scheduling process, and decisions are based on a number of attributes of the individual (e.g., age, gender), of the household (e.g., number of cars), and of the geographical zone (e.g., population density, number of shops). In general, the main strength of applying the decision trees technique for modeling activity travel choice is its ability to represent the full complexity of interactions between different attributes. However, this complexity on the other hand makes the interpretation in terms of the relative impacts of these attributes on the choice variable difficult \((6)\). Consequently, a systematic thorough sensitivity analysis as a form of model assessment and validation is sorely required.

Sensitivity analysis is the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input \((7)\). The objective of the sensitivity analysis is to identify the most significant parameters in the model and to quantify how the input uncertainty influences the outputs. A thorough sensitivity analysis helps to interpret the model, increases its credibility across a range of input scenarios and can uncover underlying errors. In general, sensitivity analysis techniques consist mainly of local approaches and global approaches. A local approach addresses sensitivity relative to point estimates of parameter values, in which inputs are varied one at a time by a small amount around some fixed point and the effects of individual variation on the output are calculated, such as partial derivative, one-at-a-time sensitivity measures, and the sensitivity index \((8)\). A local sensitivity analysis is easy to perform, and even can be calculated without detailed knowledge of the parameter distribution and without the use of random sampling schemes or large computer programs. However, these methods are inefficient when the number of parameters is large and only a few of them are influential, and they cannot take into account interactions resulting from the simultaneous variation of multiple parameters. A global sensitivity analysis approach, on the other hand, evaluates the effect of a parameter while all other parameters are varied as well and thus the entire effect on the output and interactions between input parameters can be assessed \((9)\). Moreover, the global sensitivity method can be applied to arbitrary nonlinear functions. Some typical global sensitivity analysis approaches are, for instance, regression based approaches \((10)\), regionalized sensitivity analysis \((11)\), the Morris method \((12)\), Fourier amplitude sensitivity test (FAST) \((13)\) and its extended version (eFAST) \((14)\), as well as the Sobol’ method\((15,16)\). In travel demand and agent-based modeling research, different sources of uncertainty were reviewed by \((17)\), and various sensitivity analysis approaches were applied. See e.g., \((18,19)\).

In this study, we apply both the local and global approaches for sensitivity analysis of decision trees in the FEATHERS framework. The main objective is to quantitatively measure the relative impact of input variables (condition variables) involved in the given decision trees on the
choice variable. The one-at-a-time (OAT) sensitivity measure proposed in the ALBATROSS model (6) and the improved Sobol’ method (20) are the two selected approaches that are applied respectively to two representative decision trees concerning work-related activity (i.e., commute trip) choice and transport mode choice for work-related activities in the FEATHERS framework. More specifically, after having a decision tree structure derived from a training set, the OAT sensitivity analysis consists in predicting the choice frequency distribution by varying selected input condition variables one after another, and keeping all other variables as observed. Then, impact is measured in terms of the size of differences in frequency distributions between conditions. The improved Sobol’ method, as a global sensitivity approach, involves the variation of values for all selected input variables simultaneously using samples from pre-defined probability distributions, and variance based indices are calculated subsequently (9). In short, this method is based on a decomposition of the variance into terms of increasing dimensionality, and the sensitivity of each input and input interaction are assessed based on their contribution to the total variance.

The remaining of this paper is structured as follows. In Section 2, we briefly introduce the FEATHERS framework and the decision trees we investigate in this study. In Section 3, we elaborate on the two sensitivity analysis approaches. The detailed experiment execution and the corresponding results are provided in Section 4. The paper ends with concluding remarks and future research in Section 5.

2 DECISION TREES IN FEATHERS FRAMEWORK

FEATHERS is an activity-based micro-simulation modeling framework developed for transport demand forecasting (5). Currently, it is fully operational for the Flanders region of Belgium. In this framework, an activity-based scheduler adapted from the ALBATROSS model (6) is embedded, in which a rule-based approach based on a sequence of 26 decision trees is used to simulate the way individuals build schedules. The output of the model consisting of predicted activity schedules, describes for a given day which activities are conducted, at what time (start time), for how long (duration), where (location), and, if travelling is involved, the transport mode used and chaining of trips. More specifically, the model starts with an empty schedule or diary where after it will evaluate whether or not work activities will be included. If this is the case, then the number of work activities will be estimated together with their beginning times and durations. In a second step the locations of the work activities are determined. The system sequentially assigns locations to the work activities in order of schedule position. This is done by systematically consulting a fixed list of specific decision trees. During the third step the model proceeds with the next decision steps, i.e., selection of work related transport modes, inclusion and time profiling of non-work fixed and flexible activities, determination of fixed and flexible activity locations and finally determination of fixed and flexible activity transport modes. As these activity-based schedules constitute the output of the model, they can then be used to infer Origin-Destination (OD) matrices for traffic assignment.

Technically, the decision trees in FEATHERS framework are derived from activity diary data by means of the CHi-squared Automatic Interaction Detector (CHAID) algorithm. Most decisions involve a choice between discrete alternatives, e.g., transport mode. Observations of choice outcomes are taken from the activity diary together with characteristics of the pattern as far as known in the concerned stage of the activity schedule.

To determine when to stop splitting in a given branch of the tree, the induction method uses the minimum number of cases at a leaf node and an alpha level of significance of the best
possible split, in which, the minimum number of cases is defined based on the sample size, and
the alpha significance level is set to 5%. Splitting at a leaf node stops if none of the possible
splits (in terms of the minimum number of cases) is significant (in terms of alpha). Figs. 1 and 2
show the results of two representative decision trees concerning work activity (i.e., commute trip)
implementation and private car mode choice for this activity in the FEATHERS framework.

<Figures 1 and 2 here>

To quantitatively measure the relative impact of input variables involved in the above two
decision trees on the choice variable, i.e., going to work or not, respectively using a private car or
not when going to work, we first specify the input variables for each decision under
consideration that can be used in the prediction stage. That is, 6 input condition variables (i.e.,
wstat, Gend, pAge, SEC, Day, and Xn-dag) are involved in the decision tree concerning work
activity implementation, and 10 variables (i.e., cadist, Near, trcon, SEC, Gend, Driver, TRvona,
Wdu2, TRcoff, and Urb) are considered as determinant inputs for the private car mode choice.
The detailed definition of each variable and its corresponding condition values are given in Table
1.

<Table 1 here>

3 METHODOLOGY
3.1 One-at-a-time Sensitivity Measures
To perform sensitivity analysis on an activity-based model, Arentze and Timmermans (2005) (6)
proposed a method to derive condition variable impact measures for a given decision tree in the
ALBATROSS model. The principle of the algorithm is as follows: After having a decision tree
structure derived from a training set, the tree is used to predict the choice frequency distribution
for the same set of cases multiple times under different simulated setting of the condition
variables. Then, impact is measured in terms of the size of differences in frequency distributions
between conditions.

Mathematically, let’s consider the following condition variables

\[ X_j = (x_{1j}, x_{2j}, \ldots, x_{kj}) \]

where \( k \) is the number of condition variables and \( j \) an index of cases. Assume that each condition
variable is a discrete variable with levels \( x_{kj} = 1, 2, \ldots, U(k) \), and the same goes to the choice
variable, which is denoted as \( Y_j = 1, 2, \ldots, I \).

To measure the impact of condition variable \( k (k=1, 2, \ldots, K) \), a \( U(k) \times I \) frequency table is
derived as:

\[ f_{u(k)} = \sum_j p_i (X_{u(k),j}), \quad u(k) = 1, 2, \ldots, U(k) \]  \hspace{1cm} (1)

where \( X_{u(k),j} = (x_{1j}, x_{2j}, \ldots, u(k), \ldots, x_{kj}) \) is a manipulated vector of condition values and \( p_i \) is the
predicted overall probability of \( i \).

In order to derive the above impact frequency table, a prediction run is repeated \( U(k) \)
times replacing for each case the observed value of the \( k^{th} \) condition variable by a varying value
\( u(k) = 1, 2, \ldots, U(k) \) and keeping all other variables as observed.
Based on the frequency table resulting from this procedure, information about the impact of condition variable \( k \) on choice outcomes can be obtained. The following sensitivity measures are derived from the table:

\[
IS = \chi^2(f_{u(k)i})
\]

(2)

\[
IS_i = \sum_{u(k) = 2} f_{u(k)i} - f_{u(k)-1,i}
\]

(3)

\[
DS_i = \sum_{u(k) = 2} (f_{u(k)i} - f_{u(k)-1,i})
\]

(4)

\[
MS_i = \frac{DS_i}{IS_i}
\]

(5)

The first measure represents the Chi-square of the frequency table and therefore, indicates the overall impact of the condition variable on the choice variable. The second and third measure both calculate the sum of differences in choice frequency \( i \) across successive pairs of condition levels. The distinction is that the former measure sums up absolute differences while the latter measure takes the sign of the differences into account. Thus, if the condition variable levels have a monotonically increasing or decreasing impact on choice \( i \), \( IS_i = |DS_i| \) and ratio \( MS_i = \pm 1 \). On the other hand, if the curve would have a perfect U or inverted U shape, \( DS_i = 0 \) and \( IS_i > 0 \). Therefore, the ratio \( MS_i \) represents a coefficient of monotonicity of impacts across levels on the scale of [-1,1]. In general, the whole procedure is easy to perform but rather computationally intensive. It requires repeating prediction run \( U(k) \) times for each condition variable and each decision tree.

3.2 The Improved Sobol’ Method

The Sobol’ method, originally proposed by Sobol’ in 1990 (15), is a variance-based method, which allows for simultaneous variation of the values of all input variables, in contrast to the simple one-at-a-time sensitivity analysis discussed in Section 3.1. The most important features of this method are: First, it is model independent. That is, the sensitivity measure is model-free and thus it could be applied to identify the most influential factors even when the model is complex or unknown; Moreover, it is a global method capable of capturing the influence of each input factor on the full range of output variation. That is, the total effect index accounts for how the variance of a certain output depends not only on variations of the single input (first-order effect), but also on its interaction effects with the other inputs (higher-order effects). In addition, the interpretation of the results is very intuitive and straightforward.

Mathematically, assume that a model output \( Y \) can be written as a function of its input variables \( X \):

\[
Y = f(X) = f(x_1, x_2, \ldots, x_k)
\]

(6)

where each input variable \( x_i (i = 1, 2, \ldots, k) \) has a range of variation that might lead to some uncertainty of the model output. Such a uncertainty could be represented as the unconditional variance \( V(Y) \). According to (15,21), the decomposition of this variance can be expressed as:

\[
V(Y) = \sum_{i=1}^k V_i + \sum_{i=1}^k \sum_{j=1}^k V_{ij} + \cdots + V_{12\ldots k}
\]

(7)

\[
V_i = V\left[ E(Y | X_i = x_i^*) \right]
\]

(8)
\[ V_{ij} = V \left( E \left( Y \mid X_i = x_i^*, X_j = x_j^* \right) \right) - V_i - V_j \]  
where \( V_i \) is the main effect of \( X_i \) on \( Y \) given that the \( i^{th} \) input \( X_i \) has a fixed value of \( x_i^* \); \( V_j \) is the main effect of \( X_j \) on \( Y \) given that the \( j^{th} \) input \( X_j \) has a fixed value of \( x_j^* \); \( V_{ij} \) is the joint effect of the pair \( (X_i, X_j) \) on \( Y \) given that the inputs \( X_i \) and \( X_j \) have fixed values of \( x_i^* \) and \( x_j^* \), respectively; \( V_{12\ldots k} \) is the joint effect of the inputs \( (X_1, X_2, \ldots, X_k) \).

The definition of the first-order and the total-effect sensitivity indices \( S_i \) and \( S_{iT} \) for a given factor \( X_i \) can then be calculated as:

\[ S_i = \frac{V_i}{V(Y)} = \frac{V \left( E \left( Y \mid X_i \right) \right)}{V(Y)} \]  
\[ S_{iT} = S_i + \sum_{j \neq i} S_{ij} = 1 - \frac{V \left( E \left( Y \mid X_{\sim i} \right) \right)}{V(Y)} \]

where \( S_i \) quantifies the effect of varying \( X_i \) alone and its value is between 0 and 1. A large value means a higher degree of the variable importance. \( S_{iT} \) includes the variance derived from \( X_i \) and also from its any combination with the other variables. \( S_{ij} \) is the second-order index represents the interaction effect of pairs of inputs. \( X_{\sim i} \) denotes all of the input variables other than \( X_i \).

Although it is effective, the main drawback of variance-based measures is their computational cost. Given a base sample of \( N \), the total cost of the brute-force method is therefore \( N^2 \) runs of the model.

In order to reach a numerically efficient way for the calculation, Saltelli (2002) (20) proposed an improved Sobol’ method, which is a Monte-Carlo based implementation, capable of computing sensitivity measures (the full set of first-order and total-effect indices) for arbitrary groups of factors. The numerical procedure for the calculation is summarized as follows.

We first create two independent input variable sampling matrices \( A \) and \( B \) with dimension \((N, k)\), where \( N \) is the sample size and \( k \) is the number of input variables. Thus, each row in matrices \( A \) and \( B \) represents a possible value of \( X \).

\[
A = \begin{bmatrix}
x_1^{(1)} & x_2^{(1)} & \cdots & x_i^{(1)} & \cdots & x_k^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \cdots & x_i^{(2)} & \cdots & x_k^{(2)} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_1^{(N-1)} & x_2^{(N-1)} & \cdots & x_i^{(N-1)} & \cdots & x_k^{(N-1)} \\
x_1^{(N)} & x_2^{(N)} & \cdots & x_i^{(N)} & \cdots & x_k^{(N)} \\
x_{k+1}^{(1)} & x_{k+2}^{(1)} & \cdots & x_{k+i}^{(1)} & \cdots & x_{2k}^{(1)} \\
x_{k+1}^{(2)} & x_{k+2}^{(2)} & \cdots & x_{k+i}^{(2)} & \cdots & x_{2k}^{(2)} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{k+1}^{(N-1)} & x_{k+2}^{(N-1)} & \cdots & x_{k+i}^{(N-1)} & \cdots & x_{2k}^{(N-1)} \\
x_{k+1}^{(N)} & x_{k+2}^{(N)} & \cdots & x_{k+i}^{(N)} & \cdots & x_{2k}^{(N)}
\end{bmatrix}
\]

\[
B = \begin{bmatrix}
x_1^{(1)} & x_2^{(1)} & \cdots & x_i^{(1)} & \cdots & x_k^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \cdots & x_i^{(2)} & \cdots & x_k^{(2)} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_1^{(N-1)} & x_2^{(N-1)} & \cdots & x_i^{(N-1)} & \cdots & x_k^{(N-1)} \\
x_1^{(N)} & x_2^{(N)} & \cdots & x_i^{(N)} & \cdots & x_k^{(N)} \\
x_{k+1}^{(1)} & x_{k+2}^{(1)} & \cdots & x_{k+i}^{(1)} & \cdots & x_{2k}^{(1)} \\
x_{k+1}^{(2)} & x_{k+2}^{(2)} & \cdots & x_{k+i}^{(2)} & \cdots & x_{2k}^{(2)} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{k+1}^{(N-1)} & x_{k+2}^{(N-1)} & \cdots & x_{k+i}^{(N-1)} & \cdots & x_{2k}^{(N-1)} \\
x_{k+1}^{(N)} & x_{k+2}^{(N)} & \cdots & x_{k+i}^{(N)} & \cdots & x_{2k}^{(N)}
\end{bmatrix}
\]

Next, we define a matrix \( C_i \) formed by all columns of \( B \) except the \( i^{th} \) column, which is taken from \( A \):
We now compute the model output for all the input values in the sample matrices $A$, $B$, and $C_i$, obtaining three vectors of model outputs with a dimension of $N \times 1$:

$$y_A = f(A) \quad y_B = f(B) \quad y_{C_i} = f(C_i)$$

(15)

The first-order and the total-effect sensitivity indices $S_i$ and $S_{Ti}$ can then be calculated by the following formulas:

$$S_i = \frac{(1/N) \sum_{j=1}^{N} y_{B(j)} \left( y_{C_i(j)} - y_{A(j)} \right)}{V(Y)}$$

(16)

$$S_{Ti} = \frac{(1/2N) \sum_{j=1}^{N} \left( y_{A(j)} - y_{C_i(j)} \right)^2}{V(Y)}$$

(17)

where

$$V(Y) = \frac{1}{N} \sum_{j=1}^{N} (y_{A(j)})^2 - \left( \frac{1}{N} \sum_{j=1}^{N} y_{A(j)} \right)^2$$

(18)

and $y_{A(j)}$, $y_{B(j)}$ and $y_{C_i(j)}$ denotes the model output based on the $j^{th}$ input values in the sampling matrices $A$, $B$, and $C_i$.

The most important advantage of applying the above formulas for sensitivity index calculation is that the computation is much faster via existing short cuts, and the total cost would be reduced to $N(k+2)$, which is much lower than the $N^2$ runs of the original model. For more detailed information on these formulas, we refer to (22).

Based on the estimated sensitivity indices, the impact of the different input variables and their variations on the model outcomes can be identified.

4 APPLICATION AND RESULTS

4.1 One-at-a-time Sensitivity Analysis And Results

To apply the OAT sensitivity analysis introduced in Section 3.1, we construct the frequency tables for each of the decision trees under study. Taking the decision tree concerning work activity implementation as an example, 6 input condition variables (see Fig. 1) are involved which collectively determine whether a work-related activity will be implemented or not. So, in order to measure the relative importance of each input variable on the choice variable, we compute the choice frequency distribution for each input variable by varying the value of the selected input variable and keeping all the others as observed for all possible cases. Taking the work status of each individual (wstat=0 means he/she has no work and wstat=1 means he/she has work) as an example, a frequency table is constructed by running the decision tree prediction with the variable wstat being equal to either 0 or 1 for each possible case, and the predicted overall probability of the choice variable is shown in Table 2.
It can be seen from Table 2 that if a person has no work, his/her overall probability of implementing a work-related activity is substantially lower than the probability of implementing a non-work-related activity. But this difference becomes smaller if he/she has work. Such a result is logical and also indicates implicitly that the work status of each individual has a great impact on whether a work-related activity will be implemented or not. To quantify such an impact, we calculate the Chi-square of the frequency table. Table 3 shows the results for all the 6 input variables, together with the sensitivity measures $IS_{work}$ and $MS_{work}$, and Table 4 shows the results for the decision tree concerning private car mode choice for work-related activities.

Based on the Chi-square of the frequency table presented in the second column of Table 3, we can clearly see that work status (wstat) is the most important variable for work activity choice, followed by day of the week (Day). Although the remaining four variables (i.e., gender (Gend), income (SEC), age of the person (pAge), and the number of employees within 4.4 km from home (Xn_dag)) also have some impact on the results, their importance level is much lower compared to the above two.

With regard to the directions of impact, here we only consider the variables with multiple condition levels (i.e., levels>2) in the decision tree (see Fig. 1). So, only day of the week (Day) is considered to calculate its $MS_{work}$ value. The result shows a high degree of monotonicity of this variable, and the sign indicates a tendency of decreasing choice of work-related activity as the week proceeds (note that days were numbered from Monday to Sunday), as can be expected.

Concerning the private car mode choice for work-related activities, Table 4 shows that car distance to location (cadist) is the most influential variable. The next most important variables are having a driving license (Driver), number of cars in the household (Ncar), and car travel time ratio of congested and free floating condition (TRcoff). While the remaining variables have less impacts on the choice here.

With regard to the directions of impact, we also only consider the variables with multiple condition levels (i.e., levels>2) in the decision tree (see Fig. 2), which are car distance to location (cadist) and urban density (Urb) in this study. Although the degree of monotonicity is not high for both of these variables, the sign of $MS_{var}$ suggests that the choice of driving a private car to work displays a tendency of increase with increasing travel distance (0: (0, 3.1], ..., 5: >45.8) and decreasing urban density (0: highest density, ..., 4: lowest density).

### 4.2 Global Sensitivity Analysis And Results

To apply the improved Sobol’ method, we need to first generate two independent input variable sampling matrices $A$ and $B$. To make the samples more homogeneously distributed in the whole range of variability of input variables, the Sobol’ quasi-random sequence (23) is selected with a size of $(N, 2k)$, where $N$ is the sample size and $k$ is the number of input variables. Thus, all the sampling points are uniformly distributed in the space of $(0,1)$. 

Next, given the fact that all the input variables in this study are condition variables, the randomly generated sampling points which are continuous in nature, have to be transferred into the corresponding discrete values. In doing so, we first calculate the distribution of each input variable involved in this study based on the full population files, and obtain the probability and cumulative probability for each condition level. Then, we select the condition level with the range on the cumulative probability array that includes the sampling point under study.

Having generated the input variable sets, we can now follow the steps described in Section 3.2 and repeat the simulation $N$ times to obtain the first-order and the total-effect sensitivity indices $S_i$ and $S_i^{TS}$. Figs. 3 and 4 show the results of 5,120 times of simulation for the two decision trees under study.

As we can see, most of the index scores converge to a stable value after certain amount of simulation times. In order to avoid coincidental fluctuation in one simulation and to improve reliability of the results, the average values of the last 1,000 times of simulation are calculated for the first-order and the total-effect sensitivity indices. The results are presented in Fig. 5.

For the decision tree concerning work activity implementation, both the first-order and the total effect indices show that work status (wstat) is the most influential variable to the variation of the predicted work activity choice, followed by day of the week (Day). While for the remaining four variables, their impact can be almost ignored. With regard to the decision tree concerning private car mode choice for work-related activities, car distance to location (cadist) and having a driving license (Driver) are the two most important variables, followed by number of cars in the household (Ncar), car travel time ratio of congested and free floating condition (TRcoff), and there is a train connection to location (trcon). The remaining variables have less impacts on the mode choice. Such a result is in line with the one obtained in the OAT sensitivity analysis.

However, by calculating the difference between the first-order and the total-effect indices, some new insights in the interactions occurring among all the input factors can be gained. For the work activity implementation, we find that work status (wstat) is the most influential single variable to the output of the model ($S_i$ is large while $S_i^{TS} - S_i$ is small), but day of the week (Day) has similar individual and collective effects on the output ($S_i$ and $S_i^{TS} - S_i$ are almost equal). With regard to the private car mode choice for work-related activities, we can see that almost all of the important input variables affect the output mainly through the interactions ($S_i^{TS} - S_i$ is much larger than $S_i$ for most of the input variables).

5 DISCUSSION AND CONCLUDING REMARKS

In this study, to quantify the relative impact of input variables involved in the decision tree-based scheduling algorithm on the choice variables in FEATHERS framework, two sensitivity analysis approaches were investigated. One included the one-at-a-time sensitivity measures representing the local sensitivity analysis approach, and the other was the improved Sobol’ method representing the global sensitivity analysis approach. In general, each approach has its own
advantages and disadvantages. By varying selected input variables one after another, and keeping all other variables as observed, the OAT approach is easy to understand and also easy to perform, and no detailed knowledge of the input distribution or random sampling scheme is needed. However, since only one input is varied each time, the OAT approach cannot take into account interactions resulting from the simultaneous variation of multiple input variables. Meanwhile, it is computationally intensive. For this study, a prediction was run $C_2^1C_2^1C_2^2C_4^3C_4^3C_4^4=3,360$ times to compute the choice frequency distributions for the decision tree concerning work activity implementation, and for the decision tree concerning private car mode choice, a total number of $C_6^1C_3^1C_4^1C_4^2C_4^3C_4^4C_4^5C_4^6C_4^7=414,720$ times was needed.

With regard to the improved Sobol’ method, it evaluates the effect of one input variable while all other variables are varied as well. Thus, the entire effect on the output and interactions between input variables are assessed. Moreover, although the formulas seem complex, a number of computer programs can be consulted to do the calculation, which facilitate the application of this approach to a great extent. However, to apply the improved Sobol’ method for this study, we have to estimate the distribution of each input variable first, and a sampling scheme has to be applied.

After performing these two approaches to the two representative decision trees concerning work-related activity choice and transport mode choice for work-related activities in the FEATHERS framework, both of the results showed that work status (wstat) was the most important variable for work activity choice, followed by day of the week (Day). While car distance to location (cadist) and having a driving license (Driver) were the two most influential variables for the variation of the predicted private car mode choice for work-related activities. Such a result was self-evident to some extent, which justified the effectiveness of the sensitivity analyses we performed. Also, the consistent result from the two approaches verified the key input variables for these two decision trees, and provides planners with valuable information to effectively allocate the limited resources for input data collection in the future.

In addition, some extra insights were also gained from each of these two approaches, respectively. More specifically, by calculating the extra OAT sensitivity measure, i.e., $MS$, which captures the degree of monotonicity of the impact, we found a tendency of decreasing choice of work-related activity as the week proceeds (from Monday to Sunday) and a tendency of increasing private car usage with increasing travel distance and decreasing urban density, as can be expected. On the other hand, by comparing the first-order and the total effect indices derived from the improved Sobol’ method, we found that work status (wstat) was the most influential single variable to the work activity choice, but day of the week (Day) had similar individual and collective effects on the output. While for the private car mode choice, almost all of the important input variables affected the output mainly through the interactions. It can be interpreted as follows: the work status of a person determines to a great extent whether he/she is going to implement a work-related activity, given the additional information about the day of the week. But to determine whether this person will use a private car for this activity or not, information on multiple variables is needed. Amongst others, the car distance to work location and whether he/she has a driving license are the two most important variables.

Having proven the effectiveness of applying both of the sensitivity analysis approaches to measure the relative impact of input variables involved in the given decision trees on the choice variable, the same procedures can be repeated for the remaining 24 decision trees in the FEATHERS framework. Moreover, apart from investigating activity travel choice, in the future, the same sensitivity analysis approaches can also be applied for parameter calibration and entire
model evaluation, such as to assess the robustness of the predicted travel demand and to understand how the individual components are related to each other.

REFERENCES


A List of Table and Figure Captions

TABLE 1 Input Variables Involved in the Two Decision Trees
TABLE 2 Choice Frequency Distribution Related to Work Status
TABLE 3 OAT Results of the Decision Tree Concerning Work Activity Implementation
TABLE 4 OAT Results of the Decision Tree Concerning Private Car Mode Choice for Work-related Activities

FIGURE 1 The decision tree concerning work activity implementation.
FIGURE 2 The decision tree concerning private car mode choice for work-related activities.
FIGURE 3 Iteration results of the first-order and the total-effect sensitivity index of each of the six input variables concerning work activity implementation.
FIGURE 4 Iteration results of the first-order and the total-effect sensitivity index of each of the ten input variables concerning private car mode choice for work-related activities.
FIGURE 5 GSA results of two decision trees.
TABLE 1 Input Variables Involved in the Two Decision Trees

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Condition values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day</td>
<td>Day of the week</td>
<td>0: Monday … 6: Sunday</td>
</tr>
<tr>
<td>pAge</td>
<td>Age of the person</td>
<td>0: &lt;35, 1: [35, 55), 2: [55, 65), 3: [65, 75), 4: &gt;=75</td>
</tr>
<tr>
<td>SEC</td>
<td>Income</td>
<td>0: [0, 1250), 1: [1250, 2250), 2: [2250, 3250), 3: &gt;=3250</td>
</tr>
<tr>
<td>Ncar</td>
<td>Number of cars in household</td>
<td>0: no cars, 1: 1car, 2: 2 or more cars</td>
</tr>
<tr>
<td>Driver</td>
<td>Having driving license</td>
<td>0: no license, 1: with license</td>
</tr>
<tr>
<td>Gend</td>
<td>Gender of individual</td>
<td>0: male, 1: female</td>
</tr>
<tr>
<td>wstat</td>
<td>Work status</td>
<td>0: no work, 1: work</td>
</tr>
<tr>
<td>Xn-dag</td>
<td>Number of employees within 4.4 km from home</td>
<td>0: (0,395], 1: (395,635], 2: (635,762], 3: (762,938], 4: (938, 2525], 5: &gt;2525</td>
</tr>
<tr>
<td>Urb</td>
<td>Urban density</td>
<td>0: highest density … 4: lowest density</td>
</tr>
<tr>
<td>cadist</td>
<td>Car distance to location (km)</td>
<td>0: (0, 3.1], 1: (3.1, 6.4], 2: (6.4, 12.7], 3: (12.7, 22.0], 4: (22.0, 45.8], 5: &gt;45.8</td>
</tr>
<tr>
<td>trcon</td>
<td>There is a train connection to location</td>
<td>0: no, 1: yes</td>
</tr>
<tr>
<td>TRvona</td>
<td>Ratio of access and egress time of total public transport travel time</td>
<td>0: (0,18], 1: (18,25], 2: (25,32], 3: (32,40], 4: (40,46], 5: &gt;46</td>
</tr>
<tr>
<td>Wdu2</td>
<td>Total duration of work/school in current schedule (min.)</td>
<td>0: (0,240], 1: (240,360], 2: (360,480], 3: &gt;480</td>
</tr>
<tr>
<td>TRcoff</td>
<td>Car travel time ratio of congested / free floating condition</td>
<td>0: (0,1.0], 1: (1.0, 1.53], 2: (1.53, 2.15], 3: (2.15, 2.87], 4: (2.87,3.87], 5: &gt;3.87</td>
</tr>
</tbody>
</table>
TABLE 2 Choice Frequency Distribution Related to Work Status

<table>
<thead>
<tr>
<th></th>
<th>Work-related activity</th>
<th>Non-work-related activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has no work</td>
<td>57.54</td>
<td>1622.46</td>
</tr>
<tr>
<td>Has work</td>
<td>934.92</td>
<td>745.08</td>
</tr>
<tr>
<td></td>
<td>IS</td>
<td>IS(_{work})</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>---------------</td>
</tr>
<tr>
<td>wstat</td>
<td>1100.79</td>
<td>877.38</td>
</tr>
<tr>
<td>Day</td>
<td>232.48</td>
<td>156.96</td>
</tr>
<tr>
<td>Gend</td>
<td>8.58</td>
<td>77.46</td>
</tr>
<tr>
<td>SEC</td>
<td>1.30</td>
<td>36.78</td>
</tr>
<tr>
<td>pAge</td>
<td>0.23</td>
<td>5.18</td>
</tr>
<tr>
<td>Xn_dag</td>
<td>0.02</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>$IS$</td>
<td>$IS_{\text{car}}$</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>------------------</td>
</tr>
<tr>
<td>cadist</td>
<td>52365.86</td>
<td>66227</td>
</tr>
<tr>
<td>Driver</td>
<td>29538.52</td>
<td>54180</td>
</tr>
<tr>
<td>Ncar</td>
<td>2572.07</td>
<td>11305</td>
</tr>
<tr>
<td>TRcoff</td>
<td>1210.91</td>
<td>3948</td>
</tr>
<tr>
<td>Gend</td>
<td>412.17</td>
<td>6400</td>
</tr>
<tr>
<td>trcon</td>
<td>458.49</td>
<td>6750</td>
</tr>
<tr>
<td>SEC</td>
<td>85.92</td>
<td>1461</td>
</tr>
<tr>
<td>Urb</td>
<td>81.19</td>
<td>3478</td>
</tr>
<tr>
<td>TRvona</td>
<td>22.96</td>
<td>534</td>
</tr>
<tr>
<td>Wdu2</td>
<td>0.58</td>
<td>138</td>
</tr>
</tbody>
</table>
FIGURE 1 The decision tree concerning work activity implementation.
FIGURE 2 The decision tree concerning private car mode choice for work-related activities.
FIGURE 3 Iteration results of the first-order and the total-effect sensitivity index of each of the six input variables concerning work activity implementation.
FIGURE 4 Iteration results of the first-order and the total-effect sensitivity index of each of the ten input variables concerning private car mode choice for work-related activities.
FIGURE 5 GSA results of two decision trees.