# Data Driven Validation Framework for Multi-agent Activity-based Models

Jan Drchal, Michal Čertický, and Michal Jakob

Faculty of Electrical Engineering
Czech Technical University in Prague
drchajan@fel.cvut.cz
{certicky,jakob}@agents.fel.cvut.cz

**Abstract.** Activity-based models, as a specific instance of agent-based models, deal with agents that structure their activity in terms of (daily) activity schedules. An activity schedule consists of a sequence of activity instances, each with its assigned start time, duration and location, together with transport modes used for travel between subsequent activity locations. A critical step in the development of simulation models is validation. Despite the growing importance of activity-based models in modelling transport and mobility, there has been so far no work focusing specifically on statistical validation of such models. In this paper, we propose a six-step *Validation Framework for Activity-based Models (VALFRAM)* that allows exploiting historical real-world data to assess the validity of activity-based models. The framework compares temporal and spatial properties and the structure of activity schedules against real-world travel diaries and origin-destination matrices. We confirm the usefulness of the framework on three activity-based transport models.

## 1 Introduction

Transport and mobility have recently become a prominent application area for multi-agent systems and agent-based modelling [Chen and Cheng, 2010]. Models of transport systems offer an objective common ground for discussing policies and compromises [de Dios Ortúzar and Willumsen, 2011], help to understand the underlying behaviour of these systems and aid in the actual decision making and transport planning.

Large-scale, complex transport systems, set in various socio-demographic contexts and land-use configurations, are often modelled by simulating the behaviour and interactions of millions of autonomous, self-interested agents. Agent-based modelling paradigm generally provides a high level of detail and allows representing non-linear patterns and phenomena beyond traditional analytical approaches [Bonabeau, 2002]. Specific subclass of agent-based models, called *activity-based models*, address particularly the need for realistic representation of travel demand and transport-related behaviour. Unlike traditional trip-based models, activity-based models view travel demand as a consequence of agent's needs to pursue various activities distributed in space and understanding of

travel decisions is secondary to a fundamental understanding of activity behaviour [Jones et al., 1990].

Gradual methodological shift towards such a behaviourally-oriented modelling paradigm is evident. An early work on the topic is represented by the CARLA model, developed as part of the first comprehensive assessment of behaviourally-oriented approach at Oxford [Jones et al., 1983]. Later work is represented by the SCHEDULER model – a cognitive architecture producing activity schedules from long- and short-term calendars and perceptual rules [Gärling et al., 1994] or ALBATROSS – the first model of complete activity scheduling process automatically estimated from data [Arentze and Timmermans, 2000].

In order to produce dependable and useful results, the model needs to be *valid*[1] enough. In fact, validity is often considered the most important property of models [Klügl, 2009]. The process of quantifying the model validity by determining whether the model is an accurate representation of the studied system is called *validation* and the validation process needs to be done thoroughly and throughout all phases of model development [Law, 2009].

Despite the growing adoption of activity-based models and the generally acknowledged importance of model validation, a validation process for activity-based models in particular has not yet been standardized by a detailed methodological framework. Validation techniques and guidelines are addressed in most modelling textbooks [Law, 2007] and have even been instantiated in the form of a validation process for general agent-based models [Klügl, 2009]; however, such techniques are still too general to provide concrete, practical methodology for the key validation step: statistical validation against real-world data.

In this paper, we address this gap and propose a validation framework entitled VALFRAM (Validation Framework for Activity-based Models), designed specifically for statistically quantifying the validity of *activity-based* transport models. The framework relies on the real-world transport behaviour data and quantifies the model validity in terms of clearly defined validation metrics. We illustrate and demonstrate the framework on several activity-based transport models of a real-world region populated by approximately 1 million citizens.

## 2 Preliminaries

### 2.1 Activity-based Models

Activity-based models [Ben-Akivai et al., 1996] are multiagent models in which the agents plan and execute so-called *activity schedules* – finite sequences of *activity instances* interconnected by *trips*. Each activity instance needs to have a specific *type* (e.g. `work`, `school` or `shop`), *location*, desired *start time* and *duration*. Trips between activity instances are specified by their main transport mode (e.g. `car` or `public transport`).

[1] *Valid* model is a model of sufficient *accuracy*. We use these terms interchangeably in the following text.

### 2.2 Validation Methods

Validation methods in general are usually divided into two types:

- *Face validation* subsumes all methods that rely on natural human intelligence such as expert assessments of model visualizations. Face validation shows that model's behaviour and outcomes are reasonable and plausible within the frame of the theoretic basis and implicit knowledge of system experts or stake-holders. Face validation is in general incapable of producing quantitative, comparable numeric results. Its basis in implicit expert knowledge and human intelligence also makes it difficult to standardize face validation in a formal methodological framework. In this paper, we therefore focus on statistical validation.
- *Statistical validation* (sometimes called *empirical*) employs statistical measures and tests to compare key properties of the model with the data gathered from the modelled system (usually the original real-world system).

From a higher-level perspective, VALFRAM can be viewed as an activity-based model-focused implementation of the *statistical validation* step of a more comprehensive validation procedure for generic agent-based models, introduced in [Klügl, 2009], as depicted in Figure 1. Besides the face and statistical validation, this procedure features other complementary steps such as calibration and sensitivity analysis.

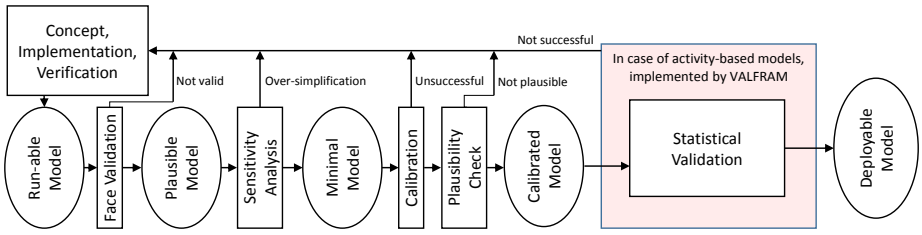



Fig. 1: Higher-level validation procedure for agent-based models in general, introduced in [Klügl, 2009]. VALFRAM implements the statistical validation step specifically for activity-based models.

Being set in the context of activity-based modelling, the VALFRAM framework is concerned with the specific properties of activity schedules generated by the agents within the model. These properties are compared to historical real-world data in order to compute a set of numeric similarity metrics.

## 3 VALFRAM description

In this section a detailed description of VALFRAM is given. We cover validation data, validation objectives and finally measures defined by VALFRAM.

### 3.1 Data

A requirement for statistical validation of any model is data capturing the relevant aspects of the behaviour of modelled system, against which the model is validated. To validate an activity-based model, the VALFRAM framework requires two distinct data sets gathered in the modelled system:

1. *Travel Diaries*: Travel diaries are usually obtained by long-term surveys (taking up to several days), during which participants log all their trips. The resulting data sets contain anonymized information about every participant (usually demographic attributes such as age, gender, etc.), and a collection of all their trips with the following properties: *time* and *date, duration, transport mode(s)* and *purpose* (the activity type at the destination). More detailed travel diaries also contain the *locations* of the origin and the destination of each trip.
2. *Origin-Destination Matrix* (O-D Matrix): The most basic O-D matrices are simple 2D square matrices displaying the number of trips travelled between every combination of origin and destination locations during a specified time period (e.g. one day or one hour). The origin and destination locations are usually predefined, mutually exclusive zones covering the area of interest and their size determines the level of detail of the matrix.

### 3.2 VALFRAM Validation Objectives

The VALFRAM validation framework is concerned with a couple of specific properties of activity schedules produced by modelled agents. These particular properties need to correlate with the modelled system in order for the model to accurately reproduce the system's transport-related behaviour. At the same time, these properties can actually be validated based on available data sets – travel diaries and O-D matrices. In particular, we are interested in:

A. *Activities* and their:
    1. *temporal* properties (start times and durations),
    2. *spatial* properties (distribution of activity locations in space),
    3. *structure* of activity sequences (typical arrangement of successive activity types).
B. *Trips* and their:
    1. *temporal* properties (transport mode choice in different times of day; durations of trips),
    2. *spatial* properties (distribution of trip's origin-destination pairs in space),
    3. *structure* of transport mode choice (typical mode for each destination activity type).

| | **A.Activities** | | **B.Trips** | |
|---|---|---|---|---|
| | Task | Data set | Task | Data set |
| **1.Time** | Compare the distributions of *start times* and *durations* for each activity type using Kolmogorov-Smirnov (KS) statistic. | Travel Diaries | Compare the distribution of selected *modes* by *time of day* and the distribution of *travel times* by *mode* using $\chi^2$ and KS statistics. | Travel Diaries |
| **2.Space** | Compare distribution of each activity type in 2D space using RMSE. Plot heat maps for additional feedback. | Space-aware Travel Diaries | Compute the distance between generated and real-world O-D matrix using RMSE. | Origin-Destination Matrix |
| **3.Structure** | *i)* Compare activity counts within activity schedules using $\chi^2$ statistics. *ii)* Compare distributions of activity schedule subsequences as n-grams profiles using $\chi^2$ statistics. | Travel Diaries | Compare the distribution of selected *transport mode* for each type of *target activity* type using $\chi^2$ statistics. | Travel Diaries |

Table 1: Six validation steps of VALFRAM framework and corresponding validation data sets needed for each of them.

### 3.3 VALFRAM Validation Metrics

To validate these properties of interest, we need to perform six validation steps (A1, A2, A3, B1, B2, B3), as depicted in Table 1 and detailed in the rest of this section. In each validation step, we compute specific numeric metrics (statistics). For all metrics, higher values of these statistics indicate a larger difference between the model and validation set, i.e., lower accuracy.

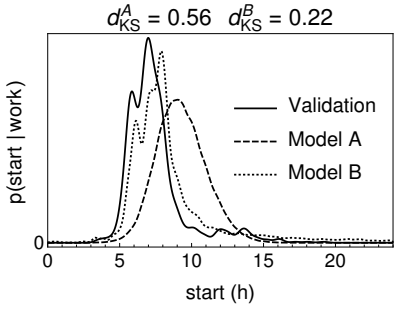
**A1. Activities in Time:** The comparison of activity distributions in time is realized by means of a well-established Kolmogorov-Smirnov two-sample statistic[2] [Hollander et al., 2013]. VALFRAM applies the method to start time distributions $p(\text{start}|\text{act. type})$ as well as to duration distributions $p(\text{duration}|\text{act. type})$.

The statistic is defined as the maximum deviation between the empirical cumulative distribution functions $F_M$ and $F_V$ which are based on the model and validation data distributions: $d_{KS} = \sup_x |F_M(x) - F_V(x)|$. The values lie in the interval $[0, 1]$.

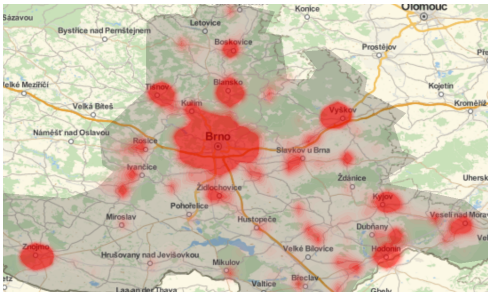
Figure 2a shows an example application of the Kolmogorov-Smirnov statistic comparing two different models to validation data.

**A2. Activities in Space:** The comparison of activity distributions in space is performed separately for every activity type. Unlike in the previous step, the distributions are two-dimensional (latitude, longitude or projected coordinates). The process consists of the following steps. First, bivariate empirical cumulative distribution functions (ECDFs) $F_M$ and $F_V$ are constructed using coordinate

[2] We have also experimented with related Anderson-Darling and Cramér–von Mises statistics getting similar results. Kolmogorv-Smrnov was finally selected as it is widely known and easier to get insight into.




(a) `work` activity start time




(b) Modeled area, `sleep` activity

Fig. 2: Start time distributions for `work` activity shown for validation data and two different models (a) including Kolmogorov-Smirnov statistics. Modelled area including `sleep` activity spatial PDF visualized as a heat map (b).

data for both model and validation data, respectively. Second, $F_M$ and $F_V$ are regularly sampled getting matrices $E^M$ and $E^V$ both having $m$ rows and $n$ columns. Third, Root Mean Squared Error (RMSE) of the two matrices is computed using $d_{ecdf} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(E^M_{ij} - E^V_{ij}\right)^2/(mn)}$. As $E^M_{ij} \leq 1$ and $E^V_{ij} \leq 1$, the measure $d_{ecdf}$ is again limited to the $[0, 1]$ interval.

Figure 2b shows the spatial probability distribution function (PDF) of `sleep` type activities on the validation set visualized as a heat map. The probability distribution was approximated from data using Gaussian kernels. Similar heat maps might be helpful when developing a model as they can show where problems or imprecisions are.

**A3. Structure of Activities:** In the previous steps, we examined the activity distributions in time and space. In this step, we consider the activity composition of the entire activity schedules. We propose a measure which compares distributions of *activity counts* in activity schedules as well as a measure comparing the distribution of possible *activity type sequences*.

*Activity Count:* The comparison of activity counts in activity schedules is based on a well-known Pearson's chi-square test [Sokal and Rohlf, 1994]. The procedure is performed separately for each activity type. First, frequencies $f^M_i$ and $f^V_i$ for the count $i$ are collected for both model and validation data. Validation data frequencies $f^V_i$ are then used to get count proportions $p^V_i$ and in turn validation frequencies $s^V_i$ scaled to match the sum of model's frequencies ($\sum_i s^V_i = \sum_i f^M_i$). Using $f^M_i$ and $s^V_i$ chi-square statistic is computed as $\chi^2 = \sum_i \left(f^M_i - s^V_i\right)^2 / s^V_i$.

*Activity Sequences:* We also compare activity sequence distributions. The method is based on the well-established text mining techniques [Manning, 1999]. Particularly, we compare *n-gram profiles* using chi-square statistic. N-gram is a continuous subsequence of the original sequence having a length exactly $n$. Consider an example activity schedule consisting of the following activity sequence:

$\langle$`none`, `sleep`, `work`, `leisure`, `sleep`, `none`$\rangle$[3]. The set of all 2-grams (bigrams) is then: {$\langle$`none`, `sleep`$\rangle$, $\langle$`sleep`, `work`$\rangle$, $\langle$`work`, `leisure`$\rangle$, $\langle$`leisure`, `sleep`$\rangle$, $\langle$`sleep`, `none`$\rangle$}. We create an *n-gram profile* by counting frequencies of all n-grams in a range $n \in \{1, 2, \cdots, k\}$ for all activity schedules. All the $N$ n-grams are then sorted by their counts in a decreasing order so that the counts are $f_i \geq f_j$ for any two n-grams $i$ and $j$ where $1 \leq i < j \leq N$ (for a tie $f_i = f_j$ one should sort in the lexicographical order). We only work with a proportion $P$ of n-grams having the highest count in the profile. More precisely, we take only the first $M$ n-grams, where $M$ is the highest value for which $\sum_{i=1}^{M} f_i \leq P \sum_{i=1}^{N} f_i$ is true.

In order to compare n-gram profiles of model and validation data, we employ chi-square statistic matching both profiles by the corresponding n-grams (only n-grams found in both profiles are considered).

**B1. Trips in Time:** The validation of trips in time consists of two sub-steps: a comparison of mode distributions for a given time of day and a comparison of travel time distributions for selected modes.

*Modes by Time of Day:* The comparison of mode distributions for a given time of day, i.e., $p$(mode|time range), is based on exactly the same approach which we used to compare activity counts (validation step A3): the $\chi^2$ statistic is computed for mode frequencies of trips starting in a selected time interval. We suggest computing $\chi^2$ statistic for twenty four one-hour intervals per day, although other partitionings are possible.

*Travel Times per Mode:* Travel time distributions for modes $p$(travel time|mode) are validated in the same way as activities in time (see validation step A1) using Kolmogorov-Smirnov statistic $d_{KS}$.

**B2. Trips in Space:** In order to validate trip distributions in space, we propose a symmetrical dissimilarity measure based on O-D matrix comparison. The algorithm is realized in three consecutive steps. First, O-D matrices are rearranged to use a common set of origins and destinations. Second, both matrices are scaled to make trip counts comparable. Third, RMSE for all elements which have non zero trip count in either of the matrices is computed.

The algorithm starts with two O-D matrices: model matrix $M$ and validation matrix $V$. Each element $M_{ij}$ (or $V_{ij}$) represents a count of trips between origin $i$ and destination $j$. The positional information (i.e., latitude/longitude or other type of coordinates) is denoted $m_i, m_j \in C_M$ for model and similarly $v_i, v_j \in C_V$ for validation data where $C_M$ and $C_V$ are sets of all possible coordinates (e.g., all traffic network nodes).

Note that in most practical cases $C_M \neq C_V$. As an example we can have precise GPS coordinates generated by the model, however, only approximate or aggregated trip locations from validation travel diaries. As we have to work with the same locations in order to compare the O-D matrices, we need to select a common set of coordinates $C$. In practice, this would be typically the validation data location set ($C = C_V$) while all locations from $C_M$ must be projected to

[3] Note, that `none` activities are added to the beginning and end of the activity schedule in order to preserve information about initial/terminal activity.

it by replacing each $m_i$ by its closest counterpart in $C$. This might eventually lead to resizing of the O-D matrix $M$ as more origins/destinations might get aggregated into a single row/column.

In many cases the total number of trips in $M$ and $V$ can be vastly different. The second step of the algorithm scales both $M$ and $V$ to a total element sum of one: $M'_{ij} = \frac{M_{ij}}{\sum_i \sum_j M_{ij}}$ and $V'_{ij} = \frac{V_{ij}}{\sum_i \sum_j V_{ij}}$. Each element of both $M'_{ij}$ and $V'_{ij}$ now represents a relative traffic volume between origin $i$ and destination $j$.

Finally, we compute the O-D matrix distance using the following equation:

$$d_{OD} = \sqrt{\frac{\sum_i \sum_j \left(M'_{ij} - V'_{ij}\right)^2}{\left|\left\{(i,j) : M'_{ij} > 0 \vee V'_{ij} > 0\right\}\right|}}. \quad (1)$$

Note that the equation is RMSE computed over all origin-destination pairs which appear either in $M'_{ij}$, $V'_{ij}$ or in both. We have decided to ignore the elements which are zero in both matrices as these might represent trips which might not be possible at all (i.e., not connected by the transport network). Possible values of $d_{OD}$ lie in interval $[0, 1]$ (the upper bound is given by $M'_{ij} \leq 1$ and $V'_{ij} \leq 1$).

**B3. Mode for Target Activity Type:** The validation of the mode choice for target activity type is again based on $\chi^2$ statistic. Here, we collect counts per each mode for each target activity of choice.

## 4 VALFRAM Evaluation

In general, we expect a statistical validation framework to meet three key conditions. First, the framework quantifies the accuracy of the validated models in a way which allows comparing model's accuracy in replicating different aspects of the beahviour of the modelled system. Second, data required for validation are available. Third, validation results produced by the framework correlate with the expectations based on expert insight and face validation.

VALFRAM meets conditions 1 and 2 for activity-based models by explicitly expressing the spatial, temporal and structural properties of activities and trips, using only travel diaries and O-D matrices. To evaluate it with respect to condition 3, we have built three different activity-based models, formulated hypotheses about them based on our expert insight and used VALFRAM to validate both of them.

### 4.1 Evaluation Models

The first model, denoted $\mathrm{M_A}$ (model A), is a rule-based model inspired by ALBATROSS[4] [Arentze and Timmermans, 2000]. The second model, denoted $\mathrm{M_B}$, is a

[4] Although we call $\mathrm{M_A}$ the rule-based model, it estimates activity count, durations and occasionally start times using linear-regression models based on data. All other activity schedule properties are based on rules constructed using expert knowledge.

fully data-driven model based on Recurrent Neural Networks (RNNs). More specifically, the model employs fully-connected Long-Short Term Memory (LSTM) units [Hochreiter and Schmidhuber, 1997] and several sets of softmax output units. Given the training dataset based on travel diaries, the model is trained to repetitively take current activity type and its end time as input in order to produce a trip (including trip duration and main mode) and the following activity (defined by type and duration). As $M_B$ is currently unable to generate spatial component of the schedules (e.g., activity locations), VALFRAM steps A2 and B2 are evaluated on a predecessor of $M_A$ denoted $M'_A$ (model A′). $M'_A$ uses a less sophisticated approach to select activity locations.

All $M_A$, $M_B$ and $M'_A$ models were used to generate a sample of 100,000 activity schedules. Our validation set $V$ contained approximately 1,800 schedules. Such a disproportion is typical in reality, since obtaining real-world data tends to be more costly than obtaining synthetic data from model. All the data used in this study cover a single workday. An overview of the modelled area is depicted in Figure 2b.

In the following text we present five hypotheses based on our insight of models. Note that all VALFRAM steps A1 through B3 are performed in order to evaluate them.

### 4.2 Test Hypotheses

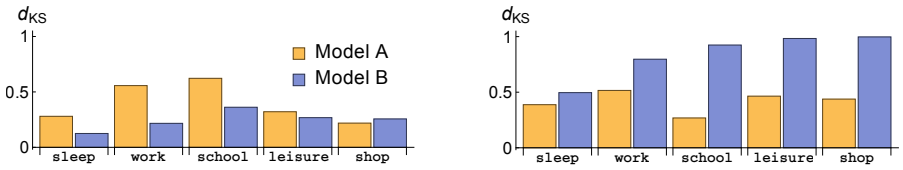
**Hypothesis 1:** The rule-based model $M_A$ uses very simple linear classifier for decisions on activity start times, so it will likely perform worse than the RNN-based model in their assignment. On the other hand, the activity scheduler in $M_A$ performs schedule optimization, during which it adapts activity durations according to rules psychologically plausible. This should produce more realistic behaviour than the purely data-driven RNN model[5].

Step A1 of VALFRAM confirms the hypothesis. Figure 3a depicts the distributions $p(\text{start}|\texttt{work})$ for validation data V and models $M_A$ and $M_B$. The values $d^A_{KS} > d^B_{KS}$ indicate the higher accuracy of the RNN model, with the most significant difference in the case of `work` and `school` activities. On the other hand, Figure 3b shows that $M_A$ outperforms $M_B$ in terms of activity durations.

**Hypothesis 2:** Activity sequences of real-world system tend to be harder to replicate using simple rule-based models than robust data-driven approaches.

Results of the step A3 (*activity counts*) for all the activity types are shown in Table 2. The data-driven model $M_B$ outperforms $M_A$ with the exception of the `leisure` activity (which we later found to be insufficiently covered by the RNN training data). Note that both $M_A$ and $M_B$ give the same $\chi^2$ value for the `sleep` activity which is caused by the fact that both models generate daily schedules having strictly two `sleep` activities in the current setup. For the step A3 (*activity sequences*) we got the following results for both models using the

[5] At least given the limited size of the RNN training dataset.




(a) start time (b) duration

Fig. 3: An example of activity in time comparison. The values of $d_{KS}$ are shown for both models $M_A$ and $M_B$. $M_B$ outperforms $M_A$ on start times while the situation is the opposite for durations.

proportion $P = 0.9$ and $k = 11$ (longest sequence in data): $\chi^2 \approx 8.4 \times 10^5$ for $M_A$ and $\chi^2 \approx 2.6 \times 10^5$ for $M_B$ showing superiority of the RNN model.

| Model | `sleep` | `work` | `school` | `leisure` | `shop` |
|---|---|---|---|---|---|
| $M_A$ | 21468.1 | 2889.3 | 542.2 | **1750.3** | 974.2 |
| $M_B$ | 21468.1 | **255.7** | **293.8** | 4625.7 | **773.8** |

Table 2: Activity counts for selected activities ($\chi^2$ statistic). Model $M_B$ outperforms model $M_A$ with the exception of the `leisure` activity type.

**Hypothesis 3:** While rule-based model optimizes the whole daily activity plans, RNN-based model works sequentially and schedules new activity based only on the previous ones. Therefore, it will be less accurate towards the end of the day.

By a further analysis of step A3 (*activity sequences*), which involved the comparison of a set of n-grams having highest frequency difference, we have, indeed, found that the RNN model tends to be less accurate towards the end of the generated activity sequence resulting in schedules not ended by the `sleep` activity in a number of cases. Moreover, Figure 4 shows a comparison of *mode by time of day* selection $\chi^2$ values (step B1) for $M_A$ and $M_B$ showing that although $M_B$ is initially more accurate it eventually degrades and the rule-based model $M_A$ prevails.

**Hypothesis 4:** Unlike the rule-based model, the RNN model has no access to trip-planning data (i.e., transport network, timetables) which will decrease its performance in selecting trip modes.

For the step B1 (*travel times per mode*) we got $d^A_{KS} = 0.22 < d^B_{KS} = 0.31$ for `car` and $d^A_{KS} = 0.37 < d^B_{KS} = 0.43$ for `public transport` modes. Results of the step B3 are summarized in Table 3 also supporting the superiority of $M_A$ in modelling mode selection.

**Hypothesis 5:** Model $M'_A$ will be inferior to $M_A$ as it uses an oversimplified activity location selection.

For the step A2 this is clearly demonstrated in Figure 5 by $d^A_{ecdf} < d^{A'}_{ecdf}$ for the `leisure` and `shop` activities (only activity types affected by the algorithm

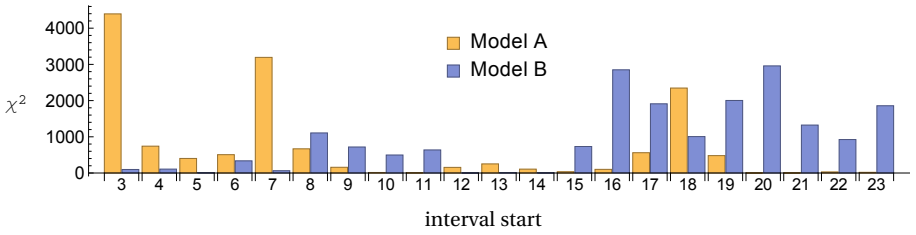



Fig. 4: Modes by the time of day. The figure shows a comparison of $\chi^2$ values for `car` and `public transport` modes for one hour intervals between 3:00 and 23:00.

| Model | `sleep` | `work` | `school` | `leisure` | `shop` |
|---|---|---|---|---|---|
| $M_A$ | **562** | **1371.7** | **1120** | 12817.3 | **5** |
| $M_B$ | 2875.2 | 3437.9 | 7286.2 | **475.1** | 2507.3 |

Table 3: Transport mode selection for target activity type ($\chi^2$ statistic). Model $M_A$ outperforms model $M_B$ in four out of five activity types.

selecting activity locations). For the step B2 we get $d^A_{OD} = 3.7 \times 10^{-4} < d^{A'}_{OD} = 4.8 \times 10^{-4}$ which again supports the hypothesised improvement of A over A′.

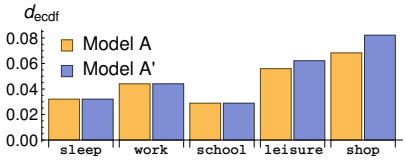



Fig. 5: Activities in space: comparison of Model A to Model A′. $M'_A$ is inferior to $M_A$ for flexible activities ($d^A_{ecdf} < d^{A'}_{ecdf}$) based on $18 \times 31$ ECDF matrices.

## 5 Conclusion

We have introduced a detailed methodological framework for data-driven *statistical validation* of multiagent activity-based transport models. The VALFRAM framework compares activity-based models against real-world *travel diaries* and *origin–destination matrices* data. The framework produces several validation metrics quantifying the *temporal*, *spatial* and *structural validity* of activity schedules generated by the model. These metrics can be used to assess the accuracy of the model, guide model development or compare the model accuracy to other models. We have applied VALFRAM to assess and compare the validity of three activity-based transport models of a real-world region comprising around 1 million inhabitants. In the test application, the framework correctly identified strong

and weak aspects of each model, which confirmed the viability and usefulness of the framework.

## Acknowledgement

This publication was supported by the European social fund within the framework of realizing the project "Support of inter-sectoral mobility and quality enhancement of research teams at Czech Technical University in Prague", CZ.1.07/ 2.3.00/30.0034, period of the project's realization 1.12.2012 – 30.6.2015. This publication was further supported by the Technology Agency of the Czech Republic (grant no. TE01020155).

## References

[Arentze and Timmermans, 2000] Arentze, T. and Timmermans, H. (2000). *Albatross: a learning based transportation oriented simulation system*. Eirass Eindhoven.

[Ben-Akivai et al., 1996] Ben-Akivai, M., Bowman, J. L., and Gopinath, D. (1996). Travel demand model system for the information era. *Transportation*, 23(3):241–266.

[Bonabeau, 2002] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287.

[Chen and Cheng, 2010] Chen, B. and Cheng, H. H. (2010). A review of the applications of agent technology in traffic and transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 11(2):485–497.

[de Dios Ortúzar and Willumsen, 2011] de Dios Ortúzar, J. and Willumsen, L. (2011). *Modelling Transport*. Wiley.

[Gärling et al., 1994] Gärling, T., Kwan, M.-p., and Golledge, R. G. (1994). Computational-process modelling of household activity scheduling. *Transportation Research Part B: Methodological*, 28(5):355–364.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hollander et al., 2013] Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric Statistical Methods*. Wiley, 3rd edition.

[Jones et al., 1983] Jones, P. M., Dix, M. C., Clarke, M. I., and Heggie, I. G. (1983). *Understanding travel behaviour*. Number Monograph.

[Jones et al., 1990] Jones, P. M., Koppelman, F., and Orfeuil, J.-P. (1990). Activity analysis: State-of-the-art and future directions. *Developments in dynamic and activity-based approaches to travel analysis*, pages 34–55.

[Klügl, 2009] Klügl, F. (2009). *Agent-based simulation engineering*. PhD thesis, Habilitation Thesis, University of Würzburg.

[Law, 2007] Law, A. M. (2007). *Simulation modeling and analysis, 4th edition*. McGraw-Hill New York.

[Law, 2009] Law, A. M. (2009). How to build valid and credible simulation models. In *Simulation Conference (WSC), Proceedings of the 2009 Winter*, pages 24–33. IEEE.

[Manning, 1999] Manning, C. D. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

[Sokal and Rohlf, 1994] Sokal, R. R. and Rohlf, F. J. (1994). *Biometry: The Principles and Practices of Statistics in Biological Research*. W. H. Freeman, 3rd edition.