Using Emulators to Approximate Predicted Performance Indicators in Complex Micro-Simulation and Multi-Agent Models of Travel Demand[1]

Soora Rasouli and Harry Timmermans

Eindhoven University of Technology
Urban Planning Group
Box 513
5600 MB Eindhoven
The Netherlands

Abstract. The very high computing times associated with micro-simulation and agent-based models of travel demand preclude their application in real-time policy development sessions in which discussions require fast turnaround times of model simulations. This paper suggests and explores the use of emulators to increase the possibilities of using large scale simulation models in real time. An emulator is a statistical approximation of the large scale model, depicting the relationship between a set of input variables and an output variable of the large scale model. To illustrate the approach, an emulator is built for the Albatross model system. Results are satisfactory. Implications and avenues of future research are discussed.

1. Introduction: motivation and objectives

Recent progress in modeling travel demand has evidenced increased application of micro-simulation and multi-agent models (e.g. Vovsha et al., 2005, 2008, Bekhor et al., 2011; Castiglione, 2011). As argued by Ziems et al. (2011), the primary motivation behind the use of microsimulation approaches concerns their ability to simulate decision-processes and choice behavior at the level of the individual decision-maker while accounting for the constraints, opportunities, and dependencies that characterize individual behaviour. The application of these micro simulation models first involves creating a synthetic population that describes each individual (and household) in the study area in terms of its socio-demographic profile. Next, this profile of each individual or agent is input to the (sub)models that together make up the activity-based model of travel demand. As these models tend to be stochastic, the simulation often involves Monte Carlo sampling from error distributions. To separate simulation error from other sources of variability, several runs of the model system are required.

As a result, the application of micro-simulation and agent-based models is computationally expensive. Especially if the process model underlying the model system is complex, a single run of the model may take many hours. To take Albatross (Arentze and Timmermans, 2004) as an example, a single run takes 1 night. Although such computing times may be less of a concern if the model is used in sequential project-type policy development settings, long computing times certainly preclude the application of complex micro-simulation and agent-based models in *real-time* transportation planning sessions.

In this paper, we argue that in this policy context emulators may be useful. An emulator is a statistical model depicting the relationship between the input variables and the

[1] Paper submitted to the 4th Conference on Innovations in Travel Modelling Conference, Tampa, USA, April 2012.

output variable of interest. It replaces and approximates the underlying process model of individual and household decision making and choice behavior and characterizes a particular feature of the complex process model. The approach has recently gained some attention in climate and environmental modelling, where computing times are even higher (e.g., Margvelashvili and Campbell, 2011; Borgonova et al., 2012).

The aim of the paper is to explore the use of emulators to reduce the computational burden in applications of complex activity-based model of travel demand, using Albatross (Arentze and Timmermans, 2004) as an example. To that end, we first briefly summarize the Albatross model system and the study area. Next, we discuss principles of building an emulator, which in this case is meant to forecast one of the performance indicators, generated by Albatross: kilometres travelled per day per person. This is followed by a discussion of the results. The paper is completed with a discussion of research issues in the development and application of emulators in forecasting travel demand.

2. Methodology

2.1 Introduction

To explore the applicability and potential of emulators in applied transportation planning we built a simple response curve statistical model of the kilometre travelled per day per person performance indicator generated by the Albatross system, a large scale activity-based model of travel demand, using agents. The purpose of this analysis was to examine whether a simple response curve model, which is considerably less demanding in computation times, could successfully represent the relationships between a set of socio-demographic input variables and total kilometers travelled performance indicator, generated by Albatross. Note that total kilometers travelled per day per person is not the quantity observed in the activity travel diaries, used in the data collection, but rather the *predicted* performance indicator of the Albatross system.

2.2 Albatross

Albatross is a multi-agent rule-based system, developed for the Dutch Ministry of Transportation that predicts which activities are conducted where, when, for how long, with whom and the transportation mode involved. The model is based on a *process* model, which mimics a priority-based scheduling process in which mandatory activities are scheduled first and discretionary activities are scheduled next. Choices made at a particular stage depend on previous choices and in addition on future possibilities. The formalism of decision trees (DT's) is used to derive the rules that underlie observed activity-travel choices as observed in empirical activity-travel diaries. CHAID-based and F-based tree induction methods are used to derive these decision making rules. Activity-travel choices are simulated *probabilistically* given a set of spatial-temporal, institutional and household constraints. The system includes 27 decision tables. Thus, the activity-travel behaviour of every member of a fraction of the created synthetic population is simulated by activating each of these decision tables sequentially using (i) person, household, institutional and environment information, (ii) simulated decision outcomes of previous scheduling steps and (iii) multiple constraints as input. Because decisions are made probabilistically, processing all individuals requires a substantial amount of computing time. Because the model system is based on the underlying process model, predicted activity-travel patterns *emerge* from the scheduling process; observed patterns are not direct input for the estimation of the model system.

The output of the Albatross system is a series of performance indicators, such as distance travelled, emissions, trip-tour ratios, market share of various transport modes, etc.

The analysis, reported in this paper, is based on one of these indicators: distance travelled per day per person.

2.3 Study area

The larger Rotterdam area serves as the study area for the present analysis. It is the second largest city in the Netherlands with a population of approximately 600,000 people in 2005. A synthetic population was created for this area using the approach described in Arentze and Timmermans (2008). This approach is like many other population synthesizers based on the iterative proportional fitting algorithm (e.g. Beckman et al., 1996), but differs in the sense that the consistency between individual and households is maintained by using relation matrices. The created synthetic population consists of 41668 individuals.

2.4 Building the emulator

Constructing an emulator depends strongly on the purpose of analysis, the nature of the variables of the model and the underlying processes. In the present study, to explore the application of emulators, the following simple emulator was built. The purpose was to predict a performance indicator Y (distance travelled per day per person) as a function of input variables while avoiding the high computing times that are associated with an application of the Albatross system. The following procedure was followed:

1. Draw a random sample s from the synthetic population;
2. Obtain for each sample s, the configuration of K input variables $\mathbf{x}_s = \{x_{s1}, x_{s2}, ..., x_{sK}\}$;
3. Run the agent-based model N_R times to obtain a stable estimate of $E(Y)_s$.
4. Estimate an appropriate statistical response curve model $E(Y) = f(\mathbf{x})$

In the present study, N_R = 100 and N = 41668

3. Results

Two statistical representations of the emulator were considered: a standard multiple regression model and a multivariate adaptive regression spline model. Because these two representations did not differ much, we report the results of the multiple regression model. The dependent variable of this model is the average total daily kilometers travelled per person, predicted by the Albatross model. The independent variables consist of a set of input variables of the Albatross model. For illustration, we selected three socio-demographic variables that would typically be used in applications of the activity-based model to simulate the effects of population change scenarios on travel demand, measured in terms of total kilometers travelled. In particular, household composition, socio-economic status (income) and age of children in the household were chosen. Each of these input variables was effect-coded. This means that L-1 indicator variables were constructed for each categorical variable of L categories. A category is coded by 1 on the corresponding indicator variable and 0 on all other indicators. One of the categories is chosen as the base and coded as -1 on all L-1 indicator variables. Because the purpose of the emulator is to obtain a good statistical representation of the response variable, all first order interaction effects were estimated as well. The results are shown in Table 1.

Table 1: Results of the emulator

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Constant	40.384	.777		52.006	.000
C0: single, no work	-27.735	2.207	-0.846	-12.567	.000
C1: single, work	-9.968	1.285	-.341	-7.759	.000
C2; two adults, one work	12.863	1.285	.440	10.012	.000
C3: two adults, two worker	30.458	1.435	1.008	21.222	.000
S0 income=min<16250	4.156	1.200	.154	3.465	.002
S1 income=low16250<<23750	-1.028	1.209	-.037	-.851	.403
S2 income=med 23750<<38750	-2.534	1.245	-.091	-2.035	.052
Ch0-no children younger than 18	1.857	1.141	.069	1.628	.116
Ch1-Child<6 years old	.884	1.141	.033	.775	.445
Ch2 - 6<<11	2.086	1.348	.073	1.547	.134
Interactions					
c0s0	1.922	2.743	.042	.701	.490
c0s1	-1.196	2.743	-.025	-.436	.666
c0s2	-.785	3.051	-.015	-.257	.799
c1s0	-1.941	2.140	-.047	-.907	.373
c1s1	.987	2.145	.024	.460	.649
c1s2	2.650	2.166	.062	1.223	.232
c2s0	-3.908	2.140	-.095	-1.826	.079
c2s1	-.975	2.145	-.024	-.454	.653
c2s2	4.127	2.166	.097	1.905	.068
c3s0	4.717	2.490	.110	1.895	.069
c3s1	-1.003	2.464	-.023	-.407	.687
c3s2	-3.397	2.213	-.080	-1.535	.137
c0ch0	3.786	2.768	.085	1.367	.183
c0ch1	-1.137	2.768	-.025	-.411	.685
c0ch2	-1.501	3.136	-.028	-.479	.636
c1ch0	2.906	2.108	.073	1.379	.180
c1ch1	1.640	2.108	.041	.778	.444
c1ch2	-6.175	2.227	-.144	-2.772	.010
c2ch0	-3.406	2.108	-.085	-1.616	.118
c2ch1	2.023	2.108	.051	.960	.346
c2ch2	6.501	2.227	.152	2.919	.007

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
c3ch0	-4.043	2.203	-.098	-1.835	.078
c3ch1	-.438	2.203	-.011	-.199	.844
c3ch2	1.225	2.724	.026	.450	.657
s0ch0	-5.041	1.880	-.134	-2.682	.013
s0ch1	.160	1.880	.004	.085	.933
s0ch2	1.909	2.021	.048	.945	.354
s1ch0	-1.199	1.886	-.032	-.636	.530
s1ch1	-1.229	1.886	-.032	-.652	.520
s1ch2	2.931	2.278	.071	1.287	.209
s2ch0	3.436	1.909	.091	1.800	.083
s2ch1	-2.314	1.909	-.060	-1.212	.236
s2ch2	.141	2.356	.003	.060	.953

The coefficient of correlation for the estimated model is .986, which is a satisfactory result. The predictions of the emulator are closely related to the daily kilometers travelled, predicted by the Albatross model system. Although less relevant in the present context, estimated coefficients of the multiple regression model show that predicted kilometers travelled are highest for two adults, two worker households and lowest for single, no worker households. Ceteris paribus, predicted total distance travelled tends to increase with an increasing number of young children (up to 11 years old), but decreases if there are older children in the household. The reason may be that children older than 12 are less escorted by their parents to or from school. The main effects for income are highest for the lowest income category, tend to decrease from the second lowest to the next higher income category, and then increase again. Upon closer examination, the estimated interactions suggest that this tendency only applies to the single, no work group.

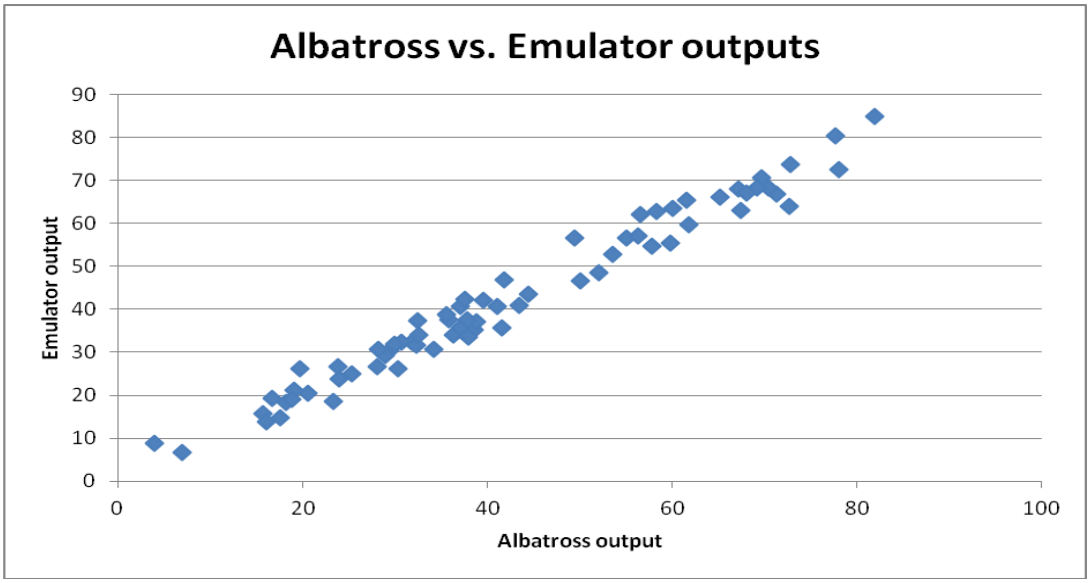

Figure 1: Albatross versus Emulator outputs

To further assess the value of the emulator, km travelled predicted by the emulator was graphed against observed km travelled (i.e. predicted by the full Albatross model). Figure 1 shows that the emulator behaves well – most data points are nicely positioned along the 45 degree line – there is no evidence of any systemic bias. Only few data points appear to be relative outliers – these points relate to single, no worker households of the lowest income category.

4. Discussion: implications for practice of travel modelling

We have argued that emulators may be potentially valuable tools in real time transportation planning sessions, where short turn-around times of model runs are critical. Emulators in this context are statistical representations of input-output relationships associated with a particular *process* model of activity-travel behaviour that uses micro-simulation or agent technology. By focusing on these relationships, large computing times required to perform the Monte Carlo simulations on the chain of stochastic sub-models are substantially reduced. The large scale, computationally intense, process model of activity-travel demand needs to be run only once for a fraction of the synthetic population to build the emulator, which can then be applied in subsequent scenario/policy development and assessment sessions with significantly lower computing times.

We have illustrated a simple emulator using Albatross as an example. We selected km travelled per day per person as an example of a performance indicator generated by the Albatross system. This application thus concerned point estimates of the relevant statistic. However, as discussed, in principle every summary measure of the sample can be emulated. Uncertainty analysis would be another major field of application. Summary statistics capturing uncertainty could be analyzed in a similar fashion. Of course, in this case, the model of activity-travel demand would need to be run a larger number of times to obtain to uncertainty distribution of the output variables. However, using an emulator would then be even more computationally efficient.

Whereas the present case has focused on model uncertainty, a similar approach could be developed for data uncertainty. If the underlying data are continuous, further decisions need to be made. These decisions relate to the sampling process (classic Monte Carlo or more dedicated sampling schemes such as Latin hypercubes), and to the nature of the data generating distributions (Gaussian or other distributions).

It should be realised that as a “model of a model”, emulators introduce another source of error. It is critical therefore that the emulator provides a high goodness-of-fit of the input-output relationships in the emulated model. Any discrepancies should be valued against the level of accuracy deemed required for the planning problem and purpose of the transportation planning session at hand.

The concept of emulator also has some interesting ramifications for methodological discussions in the field of activity-based modelling. The quintessence of the concept is to replace an underlying behavioural process model with a statistical model of input-output relationships. What if the model of travel demand is already based on such input-output relations – would the concept still be useful? Although activity-based models using discrete choice model specifications have been founded in random utility theory, they directly relate observations of activity-travel choices or full patterns to the explanatory variables of the model. In other words, arguably, one statistical input-output model would then be replaced by another statistical input-output model.

One may wonder the need to have the highly complex background model in the first place. The primary reason should be that one should be convinced that the highly complex model is behaviorally superior in predicting activity-travel behavior at different time resolutions. The

emulator would then approximate the aggregate results of the micro-simulation model, which can be used in scenario development sessions in which quick results are critical. In addition, it should be realized that the emulator does not provide detailed information; it just gives an initial estimate of aggregated results and is thus most useful when one needs to compare various alternatives in short time, avoiding too much effort and computational expenses.

Acknowledgement
This work has been funded by the European Commission, under the Seventh Framework Programme, by Contract no. 248488 within project The Uncertainty Enabled Model Web. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

References

Arentze, T.A. and H.J.P. Timmermans (2004), A learning-based transportation oriented simulation system, *Transportation Research Record* 38**,** 613-633.

Arentze, T.A. and H.J.P. Timmermans (2008), ALBATROSS: overview of the model, application and experiences, *Innovations in Travel Modeling Conference*, Portland, Oregon

Beckman, R., K. Baggerly and M. Mckay (1996), *Creating Synthetic Baseline Populations*. Elsevier Science Ltd, Great Britain, pp. 415-429

Bekhor, S., C. Dobler and K.W. Axhausen (2011), Integration of activity-based with agent-based models: Example from Tel Aviv Model and MATSim, *Proceedings Annual Meeting of the Transportation Research Board*, Washington DC.

Borgonova, E., W. Castaings and S. Tarantola (2012), Model emulator and moment-independent sensitivity analysis: An application to environmental modeling, to appear in *Environmental Modeling and Software*, to appear.

Castiglione, J. (2011), Design of integrated, advanced travel demand model with fine-grained, time-dependent network for small urban area, *Proceedings of the Annual Meeting of the Transportation Research Board*, Washington DC.

Margvelashvili, N. and E.P. Campbell (2011), Sequential data assimilation in fine-resolution models using error-subspace emulators: Theory and preliminary evaluation. To appear in *Journal of Marine Systems*, 90, 1, 13-22.

Vovsha, P., R. Donnelly and S. Gupta. (2008), Network equilibrium with activity-based microsimulation models: The New York experience. *Transportation Research Record* 2054, 102-109.

Vovsha, P., M. Bradley and J.L. Bowman (2005), Activity-based travel forecasting models in the United States: Progress since 1995 and prospects for the future. In: H.J.P. Timmermans (ed.), *Progress in Activity-Based Analysis*, Elsevier Science, Oxford, pp. 389-414

Ziems, S., S. Bhargava, J. Plotz and R.M. Pendyala (2011), Stochastic variability in microsimulation modeling and convergence of corridor-level characteristics. To appear in *Transportation Research Record* 3560