Variable selection in Land-Use and Transport Integrated models

Cedric Boittin*, Nicolas Gaud*, Vincent Hilaire*, and David Meignan**

* IRTES-SeT Laboratory
University of Technology of Belfort-Montbeliard
Belfort, France
** Universität Osnabrück
Fachbereich Mathematik / Informatik
Osnabrück, Germany

All around the world, the growth of urban areas has considerably increased in the last century. The management of these areas by controlling the transportation facilities development and the land-use evolution, has therefore become a critical task. In order to efficiently capture the economic, social, and environmental aspects, it is necessary to design systems which precisely model their causes and effects. In the last century, many studies tackled the two disciplines of transport and land-use, and emphasis has been set on their mutual relationships. It is now of common knowledge that land-use evolution induces a change in the transportation needs, and conversely. Since the 1950s, researchers have tried to simulate these multiple interactions by developing Land-Use Transport Integrated (LUTI) models. The primary objective in LUTI modelling is to support decision making, at a determined geographical scale (city, region, etc.). LUTI simulations aim to predict the evolution of the considered system, under constraints determined by the various scenarios the end-user wants to evaluate.

When modelling a system, the objective is to approximate the real system, while maintaining a tradeoff between precision and costs. The system components and their characteristics, behaviors and interactions are thus reduced to a set of parameters, equations and algorithms. The definition of the equations and algorithms is specified during the model design. Then, parameter values are estimated from observations of the real world, and adjusted during the calibration. The conclusions based on uncalibrated or inappropriately calibrated models could be misleading and even erroneous [1].

Little effort has been made toward a standardization of LUTI models calibration, even though it is still a crucial issue. Instead, many LUTI experts build on the experience acquired from other disciplines, such as econometry, demography or transport engineering, to tune their own application ad-hoc. Some aspects of the resulting process incidentally lack in both precision and meaning. The authors aim to help LUTI experts to tackle this issue by contributing to a standard Calibration process, with an efficient methodology and the associated tools.

More precisely, the authors propose in this paper an application of Iterative Local Search (ILS) optimization for producing a tool to help LUTI experts in the calibration of their models. An important part of LUTI simulations relies on the design of choice models to reproduce the decision process of the various actors of the system. For example, a company's location choice influences the number of available jobs in an area, which in turn may attract people to relocate closer to this area, as well as inducing more trips to and from it. With the relocation of the people, building construction is increased, thus influencing companies relocation. Following the work of McFadden [2], choice models in economic applications often depend upon the Utility theory. Utility represents the behavior of an individual facing a decision : he determines several criteria and tries to weigh up their importance in order to evaluate his alternatives.

Formally, Utility can be defined for any individual i and alternative j as a function $U_{ij} = f(x_{ij}\ ,\ s_i)$ where s is a vector of individual characteristics influencing tastes, and x is a vector of observed attributes of a choice [2]. A simple expression of f, commonly used in practice, is

$$f_{ij}(x, s) = A.x + B.s + \varepsilon_{ij} \tag{1}$$

with A and B vectors of real numbers, with the same sizes as x and s. The elements of A and B represent the relative importance an individual associates with each criterion. Not everything can be modelled, there exists unobserved factors which influence the utility as well. The random term ε is added to account for these factors. Most of the time, a probability

distribution is assumed for ε [3]. In this paper, we will use this expression of the utility function to picture the whereabouts of calibrating LUTI models.

Following Equation 1 and given a particular distribution for ε, two means of action appear for calibration : the tuning of the weights given by A and B, and the choice of the representative attributes in x and s. The latter is referred to as "variable selection". It lies at the border between model design and model calibration. Most of the time, it is considered as part of the former, and performed with a try-and-error procedure, as in [4]. However, this kind of approach usually gives poor and unreliable results. Instead, it is strongly advised to determine a consistent variable selection method. Current research trends have shown interest in that direction, for example with techniques of sensitivity analysis [5]. The authors' approach in this paper adopts a different point of view on the variable selection problem, which helps to solve it in an efficient and reliable manner, and broaden the range of techniques that may be applied.

The tendency for the tuning of the weights in utility-based models is heavily oriented on statistical methods such as the maximum likelihood [2, 3]. This process is termed model estimation. Likelihood-based techniques have the advantage of numerically expressing the model quality, and thus allow an easy and quick comparison of any two models. Several criteria have been derived from the maximum-likelihood estimator to make this comparison as precise as possible. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), for example, are commonly used in practice to compare utility-based models. Both penalize the value of the log-likelihood by a factor of the number of variables in the model. These criteria can easily be used to perform variable selection.

If the statistics field provides competent tools for estimating and comparing models, they are still not ideal. These methods are based solely on one particular state of the system, while the resulting model is supposed to capture its evolution in time. The more the simulation advances in time, the more the results become inaccurate. The estimated model is an approximation of the real system in a given state, with an error that is supposed to be contained. When executed multiple times – from one year to the next – the errors add and the control over it is lost. Model estimation is therefore not sufficient, it needs to be integrated in a broader framework of model calibration. Estimation can be understood as a preliminary step which outputs a good candidate model for calibration.

A common way to calibrate a model is to compare its results with observations of the real system. A frequent and efficient practice for calibrating simulation models is to optimize an objective function [6, 7]. It expresses the quality of the model by comparing the simulated results to reference data. Using different reference dates for initial estimation and calibration, it is possible to capture the time-related dynamics of the system, and thus increase the accuracy in predictive applications. In a sense, the objective function measures the distance of the model to the real system as a function of the parameters. Calibration amounts to finding the optimum of that function.

In the context of utility-based choice models, two means of action were identified. Both must be considered in the calibration process : depending on the outcome of the simulation, either the parameters (the weights) have to be tuned, or the variable selection will be modified. Three issues arise. First, given the complexity and computational cost of the simulation, parameter tuning in itself becomes time-consuming. Second, when the variable selection is modified, it has to be done all over again. Moreover, variable selection taken independently is a very complicated task, since the total number of combinations in the set $\mathcal{S}$ of all possible selections of n available variables is the cardinal of the powerset of size n : $card(\mathcal{S}) = card(\mathcal{P}(n)) = 2^n$. The shape of the search space is also very chaotic with many local optima, which makes its exploration hard. The second issue, though, can be mitigated by re-estimating the model produced from the new selection.

Some methods from the medical domain exist for performing variable selection from the whole set of available variables, such as the LASSO [8] and the Elastic Net [9]. To the best of the authors' knowledge, they have never been applied in the LUTI domain, though they might produce interesting results [10]. They would allow a rigorous estimation of the optimal variable selection, but like the pure maximum-likelihood estimation, they are inefficient for calibration. Indeed, in order to finely tune the selection and the parameters value, we need to have control over the process and be able to modify the selection. These methods do not allow it, they are static on a given state of the system.

Instead, the authors propose a simpler way of performing variable selection, based on Iterative Local Search (ILS). ILS relies on a local search procedure to produce locally optimal solutions, and a diversification one to widely explore the search space by starting from different solutions.

In order to perform local improvements, it is necessary to determine the neighborhood of a solution. The neighborhood of a variable selection $s_i \in \mathcal{S}$ is defined as the subset $\mathcal{N}$ of selections from $\mathcal{S}$ such that each $n_j \in \mathcal{N}$ is either s_i plus one variable, or s_i minus one variable. The local search procedure then consists in testing all variable removals and all variable additions from s_i, until none can be found.

The procedure may employ the statistical criteria of model quality or an objective function determined from the simulation results. In that respect, it can be used either in estimation to determine an initial solution, or in calibration to modify and fix a candidate solution. For estimation, two applications are possible : to find one decent initial solution from scratch, or to test and compare different promising directions determined by the expert.

However, it is not at all guaranteed that the resulting selection is a global optimum. Due to the size and complexity of the search space, there exist no method to do that in a reasonable time. LUTI experts might desire a better exploration the search space. Most optimisation metaheuristics include diversification aspects to various degrees, but the modeller needs to be careful to the potential running time increase. ILS has the secondary benefit that can be very easily distributed on many computational units so to linearly decrease the running time.

To illustrate the proposed technique, a LUTI model of the Paris area has been created based on previous work from the SIMAURIF project [11], which includes 12 million people among 5 million households, spread over 12 000 km^2. A framework [1] has also been designed in order to enable the comparison and testing of optimisation strategies. These research works are funded by the ANR [2] and part of the CITiES [3] project.

In a first step, only a household location choice model is estimated. The proposed procedure could also be applied to other submodels, or even extended toward global calibration. The input data comes from the General Population Census [4] at year 1990. A database of 50000 households was constructed from the data, and fed to the model for estimation. The households are spread among 572 zones. The variable selection operates on a set $\mathcal{S}$ of 42 variables, with 13 corresponding to the zone and 29 to the household characteristics.

The estimated variable selection is performed using Iterative Local Search as described in the previous section, and the value of the AIC as its objective function. 10 starting points have been determined with 9 variables each, so that the computational time was kept reasonable. Each starting model contained 3 variables from the zones dataset, and 6 from the households datasets. In order to widely cover the search space and possible exploration, the similarities were kept minimal. The starting models were constructed so that the maximum number of common variables remains minimal. They were then refined using test statistics. Each variable was tested on its significance towards the model, and a p-value was produced. Variables were removed from the model until each one met a satisfaction criterion of p-value < 0.1. Throughout the optimisation, it was made sure that the variable selections produced models which met a harsher criterion of p-value < 0.05. This ensured the produced models were consistent and meaningful. Each selection was estimated by maximum-likelihood two times, and another 10 times if a potential improvement was found. This way, the sampling bias in the selection of relocating households wasn't significantly influencing the results.

Figure 1 shows the AIC evolution of each initial solution throughout the local search. As expected, each one eventually gives very different results in terms of model efficiency, and it can be seen that the variations in AIC are very significant. There is a very strong evidence towards the local optima as opposed to the other intermediary solutions. The local search converged rapidly, after testing 80 to 250 models in the most extreme cases. The various local optima, whether good or not, outline interesting facts. First, if the initial selections contain more household characteristics than zone characteristics, the reverse effect tends to happen in the local optima. Also, when comparing the composition of the optima, only 33 variables appear, and 6 in more than half of the models. This information leads to the definition of a new indicator Q of the quality of a variable v. Given a set $\mathcal{O}$ of n locally optimal variable selections $o_i,\ i \in [1 \dots n]$, $Q(v) = \frac{card(\{o_i \ / \ v \in o_i\})}{n}$. $Q(v)$ is the ratio of the presence of v in the local optima. It reaches 1 if v is present in each one, and 0 if it is never present. This indicator adds to the information the expert can extract from the simulation, and will help him to more efficiently calibrate the model by facilitating the identification of most relevant variables.

Figure 2(a) pictures every solution tested by the algorithm during two local searches. As can be seen, a small variation in the variable selection often leads to a major change in the model quality, and emphasizes the need to accomplish it carefully and rigorously. There is a gap around run 49 which indicates a probable bad choice of variable since it happened when removing a variable. On an addition (run 70 in figure 2(b)), it brings evidence towards the essential role of the variable in the model.

The proposed procedure constitutes a first step towards a more rigorous variable selection in LUTI models, and serves as a new tool for a competent calibration. It was shown on a practical application that variable selection is a critical task, even though it is most of the time overlooked by LUTI modellers. The procedure is a really simple yet powerful algorithm, which may also serve as a reference when testing more elaborated methods. Inside a general simulation calibration framework, it has a threefold benefit : the determination of a decent initial solution, the fine tuning of the current solution, and an increased significance when comparing models. This last point is further improved by the determination of a new variable quality criterion. This is especially useful when considering interactive optimisation algorithms which allow the LUTI experts to guide the calibration. Future works will notably focus on the comparison of the ILS procedure against LASSO-based techniques. This procedure will then be integrated inside a full calibration methodology.

1. Using two existing softwares : OPUS (Open Platform for Urban Simulation : `www.urbansim.org`) for the land-use and PTV Visum (`http://vision-traffic.ptvgroup.com/en-us/products/ptv-visum/`) for the transport.
2. Agence Nationale pour la Recherche (*National Research Agency*, France) : `http://www.agence-nationale-recherche.fr/en/`
3. ANR CITiES - Calibration and validation of Transport – land-usE modelS : `http://www.multiagent.fr/CITiES`
4. *RGP – Recensement Général de la Population*, `http://www.insee.fr/en`

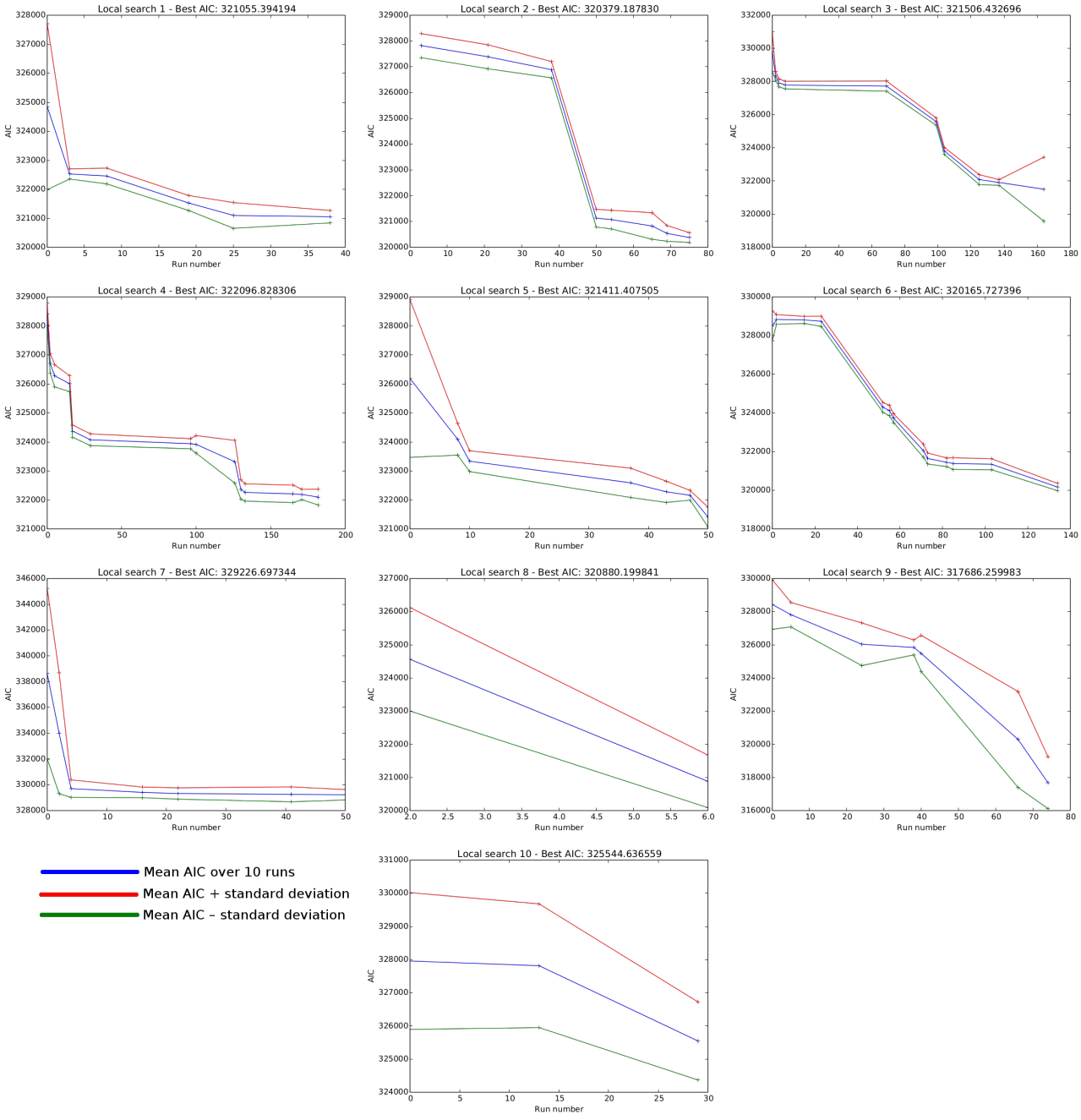
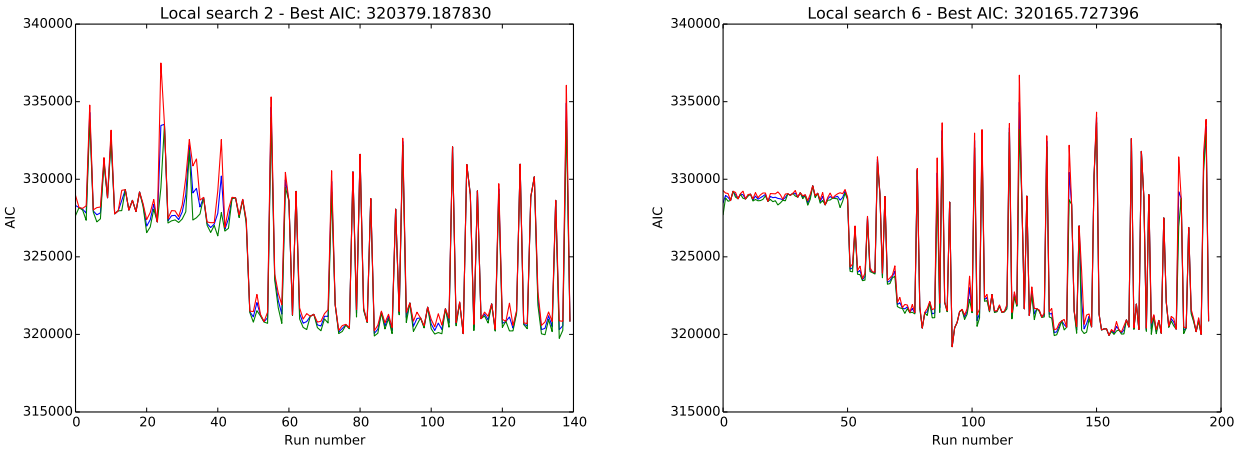


FIGURE 1 – *Local search runs*



(a) (b)

FIGURE 2 – *All tested selections – Runs 2(a) and 6(b)*

Références

[1] Park, B. and Qi, H. *Development and evaluation of a procedure for the calibration of simulation models*. Transportation Research Record : Journal of the Transportation Research Board, volume 1 :pages 208–217 (2005).

[2] McFadden, D., Talvitie, A., Cosslett, S., Hasan, I., Johnson, M., Reid, F., and Train, K. *Demand model estimation and validation*, volume 5. Institute of Transportation Studies (1977).

[3] Train, K. *Discrete choice methods with simulation*. Cambridge university press (2009).

[4] Haller, R., Emberger, G., and Mayerthaler, A. *A system dynamics approach to model Land-Use/Transport interactions on the national level* (2008).

[5] Dutta, P., Saujot, M., Arnaud, E., Lefevre, B., and Prados, E. *Uncertainty propagation and sensitivity analysis during calibration of an integrated land use and transport model*. In *International Conference on Urban Regional Planning and Transportation* (2012).

[6] Valente, P., Pereira, A., and Reis, L. P. *Calibration agent for ecological simulations : a metaheuristic approach*. In *Calibration Agent for Ecological Simulations : a Metaheuristic Approach* (2008).

[7] Fehler, M., Klugl, F., and Puppe, F. *Techniques for analysis and calibration of multi-agent simulations*. In *Proceedings of the 2004 ESAW*, pages 131–136 (2004).

[8] Tibshirani, R. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288 (1996).

[9] Zou, H. and Hastie, T. *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society : Series B (Statistical Methodology), volume 67(2) :page 301–320 (2005).

[10] Tutz, G., Pößnecker, W., and Uhlmann, L. *Variable selection in general multinomial logit models* (2012).

[11] Nguyen-Luong, D. *Simaurif : Rapport final de la 2ème phase*. *Technical report*, IAU-idF (2007).