Reflections on the concepts of model's calibration and validation

D. Andrey, P. Giordano, A. Vancheri

Several models for complex systems usually requires the evaluation of a great number of parameters to work correctly. It is the classical problem of model's calibration, faced by several authors using optimization techniques, statistical analysis, fuzzy logic methods, etc. Usually the problem can be schematized in the following way in case of urban systems:

- suppose to have empirical configurations of the system at different times;
- start the dynamical model taking as initial configuration the oldest one;
- use "some method to compare" the simulated configurations (outputs of the model) with the real ones;
- "adjust the parameters" so as to minimize the differences between this two types of configurations.

The problem of model validation ("what is the precise meaning we give to the sentence: the model describes the real system?") is frequently faced in a similar way: instead of "adjust the parameters" we consider different empirical configurations and, running the model as above, we compare again simulated outputs and real situations.

But these solutions to calibration and validation problems **seem to be incorrect** if faced using the above mentioned ideas, based only on the comparison of configurations. In fact they reflect exactly the following idealized situation:

- suppose to have a loaded die (the urban system);
- the only knowledge about this die is that it started in the past with the face '1' up, and after a throw it ended with the face '6' up (the empirical configurations);
- suppose to have a detailed physical model of a generic loaded die, with several parameters;
- the problems are: how can I calibrate model's parameters without any knowledge about the probability of the die's path going from '1' to '6'? How can I say that my model describe the given real loaded die?

It is important to consider that our urban models (generally speaking models of complex systems) usually are necessarily stochastic models so the problem is even greater: what simulated configuration do I have to compare with the real ones? The mean configuration? But what does it happen if the path of my urban system (the empirical configurations) doesn't reflect a mean behaviour? The risk is hence to calibrate our models to work with a mean dynamics corresponding to a real situation which is not "mean". It seems indeed really difficult to evaluate the probability associated to the path covered by the real city.

These problems generally affect every model of a complex system to which is not possible to apply the old paradigm of Physics about the repeatability of a physical experiment.

Calibration based on agents' behaviour pool

In our urban model all the parameters are introduced for the definition of the intensities of the (Poisson distributed) stochastic processes and for the potentials associated to each cell, see [Van et al 04]. These quantities depends only on the behaviour of populations of agents: even if each fuzzy membership function reflects the different evaluation of each agents, and their random behaviour is formalized using stochastic processes, a different set of parameters for these membership functions can theoretically cause a different behaviour of the model. Hence these parameters cannot be considered in the same way as physical constant (like e.g. the speed of light) but as quantities liable to a certain variability.

For this reasons the calibration phase of our model has been performed constructing a list of questions submitted to a chosen set of agents (architects, builders and people resident in Ticino, where the model will be concretely applied). The aim of this question is to restrict the best one can the possible values of the parameters. So we have question of type: "how do you evaluate the possibility to live in this (simple) urban configuration?", or "how do you evaluate the possibility to build a new residence in this urban configuration?".
In this way we really obtain for each parameter a (usually small) interval of possible values, depending on the different answers of the chosen set of agent for the poll. But any possible n-tuple of parameters chosen from these intervals represents a possible behaviour of agents and hence a possible model. This seems a meaningful way to define in a precise way the notion of "possible future sceneries". In fact a classification of the behaviour of the different n-tuple can be made using fuzzy methods to compare and to classify final configurations (see below). This idea will be performed in a future work.

Validation using experts' knowledge

We outlined above that the basic problem is that we only know some non-repeatable past configurations of the urban system. For the sake of simplicity we can call them "past configuration" and "present configuration": they are the analogous of the mentioned face '1' and face '6' of the loaded die. Following this example the questions are: is the die loaded or not? If yes, what is the probability of the path from '1' to '6'? What is the probability associated to different paths of the die?
Instinctively it seems an impossible problem, but may be this is not really so because the essential point here is that "the micro-dynamics producing the random movement of the die is known enough"; moreover I can make some questions to the users of the die…
Facing a so poor knowledge of the complex system, and the impossibility to have repeatable experiments (I cannot throw the die again) we think it cannot affirmed, like in classical Physics, that "our model describes, with a given precision, the reality", but rather that **we have a list of criteria on the basis of which I can say that the model is highly plausible**[1]:

- first of all the model has to include, between all its possible outputs, exactly the path $1 \to 6$. It may seems trivial, but if the (stochastic) model whatever it starts from the past configuration, it doesn't reach the present one, then the model itself is not valid.
- Together with the model one has to give a (either numerical or analytical) way to evaluate the probability that the model starts from the configuration '1' and ends to a final configuration "similar" to '6'. Here "similar" obviously has to be formalized using some scientific criterion. In the next section we will see our use of fuzzy logic for this aim. Obviously nobody knows exactly the probability of the path $1 \to 6$, but it is likely to say if this probability is *plausible*: if the evaluation of the model says that e.g. it is 0.2 and the users of the die say it is too low, then the model is not valid. The same criteria has to be applied if the evaluation of the probability is too high. For urban system the probability associated to the path 'past configuration $\to$ present configuration' has to be plausible with respect to experts' opinion.
- One can try to find indicators that, in spite of the randomness of the model, remain substantially unchanged. At the moment this is an idea that has to be investigated further, but some possible candidate for these type of indicators is described in our computer simulations and in the working paper 'computer_simulations.pdf'. The idea is to define these indicators, to prove by numerical simulations that they have a very low changing, and finally to find in some statistical data a possible confirmation of this model's forecast.
- Another possible criterion of validation can surely be seen if the model can be used in "different" urban systems without the need to be re-calibrated. Here with "different" we mean e.g. "different cities, but in the same cultural context, so that the behaviour of the agents remains unchanged".

[1] These criteria have to be applied **after** the calibration phase.

- The model has to be able to behave in a plausible way in idealized situations where the experts have a clear understanding and judgement. So it has to behave in "the most natural way" if it starts with an initial configuration e.g. given by a completely void building area, or a building area with few streets, or a completely protected area, etc.
- The model has to foreseen with a low probability idealized situations having a low plausibility: for example it has to evaluate with a low probability situations where it starts from an industrial zone and a high residential growth is foreseen, or where a great number of offices in town centre are transformed in private residences, etc.

These are possible criteria on the basis of which one can affirms that the model plausibly behaves like the real system.
In a certain sense this validation criteria permit to return to the concept of repeatable experiment: one propose a validation criterion (= experiment on model's behaviour) to the scientific community; the community can answers saying: "let's try this modification of your criterion (= repetition of the experiment), what is the behaviour of the model?". One experiments on the model instead of reality, but every results obtained has to be highly plausible.

The use of fuzzy logic to compare configurations

As we seen above, for several reasons there is the need to compare two configurations of a given urban system. In our context (continuum state CA) this problem can be faced using some kind of distance on the configuration space, e.g. the relative percentage deviation, but methods of fuzzy programming ([Rom 94]) seem to be more meaningful and efficient. In fact the problem can be seen as a goal attaining problem under constraints, where objectives and constraints are treated as fuzzy sentences connected by AND operators: the result of this fuzzy evaluation will not be "the goal has been attained or not" like in classical logic, but "the goal has been attained in the measure $\mu \in [0,1]$". More specifically one construct a list of fuzzy sets $G_k(x) \quad j = 1, \ldots, k$ which represent either goals or constraints on a universe of discourse E which in our case will be chosen to be the configuration space of the CA. Each goal (we do not actually use constraints) is constructed using one or more indicators of the configuration capturing one feature of the system at a chosen space scale. The goal consists in the (fuzzy) equality between the values of the indicators for the empirical and simulated state. A goal could be for example stated qualitatively through a sentence like: $G_1(e)$ = "the simulated configuration e shows the same density of the settlement inside a range of 1 km from the (given) cell c". After a set of goals for different indicators and space scales has been chosen one has to build the fuzzy set $D(x) := G_1(x) \wedge G_2(x) \wedge \ldots G_k(x)$ which measure the degree at which the goals have been attained. For the AND operators the whole set of t-norm is available and furthermore techniques are available in order to weight the different goals. This approach seems to have two main advantages with respect to the use of standard metrics on E to evaluate the distance between configurations:

- the goals can be chosen and weighted in such a way as to reflect the opinion of experienced peoples; in this way the distance between configurations will be a distance with respect to a given set of explicit criteria, disregarding not important features and enhancing the more relevant ones;
- increasing the number of indicators employed to construct the goals and reducing the space scale at which the urban space is regarded, a more refined description of distance is obtained.