

A Selection Criterion for a Class of Agent-Based Spatial Decision Models

Tei Laine
Indiana University
telaine@cs.indiana.edu

Abstract

A model selection method inspired by the idea of universal models is proposed for choosing between spatially explicit multi-agent decision models. Since this kind of models are often used as support in real-world decision making, and usually there are not much data available to validate the models, it is important that the selection of the best model is based on sound criteria, not merely on model's face value. In this research the best model is defined as one with an adequate fit to the current data and good generalizability to the future data. The criterion's ability to choose the best model is achieved by properly combining measures for goodness of fit and complexity in order to avoid over-fitting, i.e., the models ability to fit noise in the data, induced by its superfluous complexity. The proposed model selection criterion is applied and its performance is analyzed in the context of simple spatial agent-based decision models, which are made to generate artificial data. The preliminary results indicate that the criterion adequately balances the two measures by preferring the simpler model when there are not much data available, even when the more complex model generated the data.

Contact:

Tei Laine
Computer Science Department
Indiana University
Lindley Hall 215
150 S. Woodlawn Ave.
Bloomington, IN 47405-7104

Tel: +1 812 855-8702
Fax: +1 812 855-4829
Email: telaine@cs.indiana.edu

Keywords: Agent-based modeling, model selection, universal models, agent-based modeling, decision making

A Selection Criterion for a Class of Agent-Based Spatial Decision Models

Tei Laine

Introduction

The fact that many application domains of agent-based models are inherently complex tends to lead to complex models. Cognitive scientists and the machine learning community have mostly been concerned with over-fitting induced by the complexity, in other words, the model's ability to fit random variability in the data instead of capturing useful regularities in it. In other fields model validity, particularly, how well the model adheres to reality, is a central issue. Supposed realism, achieved by replicating real world processes and structures in great detail may introduce complexity that makes the model incomprehensible and undermines its ability to answer the scientific question it was built to answer (Burton & Obel, 1995). It is suggested that the more complex models are not necessarily more realistic than simple ones, but only more complicated.

The goal of a model selection method is to choose the best model among the candidates, or defer the selection if none of the models is supported enough by the available evidence. The best model is often determined by goodness of fit to the observed data, which are usually samples of a larger population. Using fit as a single criterion has a danger of over-fitting; a complex model can fit a data sample perfectly, but its true explanatory power, or generalizability, may be compromised. On the other hand, a model that is complex enough to fit a wide variety of data is not easily falsifiable (Roberts & Pashler, 2000). Therefore, the goal of the model selection criterion is also to choose an appropriate level of complexity required to explain the phenomenon (Kearns, Mansour, Ng, & Roi, 1997). This goal adhered to the principle of *parsimony*, known also as *Ockham's Razor*, which states that "entities should not be multiplied beyond necessity".

Unlike mathematical models, whose complexity can be measured in relatively simple and well-defined entities, such as the number of free parameters (Forster, 2000; Myung, 2000; Pitt, Myung, & Zhang, 2002), in complex adaptive systems and agent-based models the sources of complexity may be hard to discover and quantify. Without an adequate method of measuring the complexity models may be chosen on unsound grounds. Complex models may be incomprehensible and consequently not able to enhance understanding. Since some of the models may be used in real-world decision making, for instance in natural resource management, it is important to be clear in what the model predicts and what it does not.

In this research we propose and analyze a model selection criterion, based on *universal models* (Rissanen, 1999), that makes explicit the relationship between model's goodness-of-fit and its complexity. The criterion is applied to the class of agent-based models in which a group of foraging agents move around a two-dimensional landscape and consume food to replenish their energy supplies. The agents use different methods when deciding where to move next, and they incur a relatively high cost for every move they make.

Spatially Explicit Agent-based Models

Agent-based models consist of autonomous actors, which operate individually or co-operate by communication and coordination. They are designed to address the question of how large-scale phenomena emerge from actor heterogeneity — their individual characteristics — and small-scale phenomena, for instance local behaviors and interaction between the agents.

In this research we focus on a special class of agent-based decision models, namely models in which the spatial landscape is explicitly represented. Spatially explicit agent-based models have been used to model land-use and land-cover change, migration and eventual disappearance

of ancient civilizations, urban migration and segregation and urban sprawl. Hoffman, Kelley and Evans (2002) and Laine and Busemeyer (2004) use the agent-based approach to model private agricultural land-use decisions. Axtell *et al.* (2002) propose a multi-agent model of growth and collapse of Anasazi settlement in North-eastern Arizona from 1800 B.C. to 1300 A.D. Benenson (1998) uses a multi-agent model in studying population dynamics in the city. He simulates spatial segregation as a function of the agents’ cultural identity and economic status. Brown *et al.* (in press) model the effectiveness of greenbelts in preventing urban sprawl. Bodin & Nordberg (under review) study the effect of information networks in local adaptive management of natural resources with a spatial agent-based framework.

Model Selection Based on Universal Models

Rissanen (1999) has proposed a general principle for inductive inference, that is based on the concept of *universal models*¹. The idea behind the principle is to choose a model (class) that can extract the most useful regularities in the data. This is achieved by choosing the class \mathcal{M} that minimizes the worst case *regret*, i.e., the maximum discrepancy between the fit² of \mathcal{M} and the best fitting, or optimal model in hindsight, for each data sample. A noteworthy property of the criterion is that it does not assume that the data generating model, or “the true state of the world”, is among the candidate models or that such a model exists. Rissanen has proven that the universal model gives the same regret with respect to all data samples (Grunwald, Manuscript). This is uniquely achieved for model class \mathcal{M}_i using the *normalized maximum likelihood (NML)* distribution (Rissanen, 1999):

$\bar{P}_{nml}^{\mathcal{M}_i}(d) = \frac{P(d|\hat{\theta}_{\mathcal{M}_i}(d))}{\sum_{d' \in \mathcal{D}} P(d'|\hat{\theta}_{\mathcal{M}_i}(d))}$, where $P(d|\hat{\theta}_{\mathcal{M}_i}(d))$ is the probability that the maximum likelihood model in the class \mathcal{M}_i gives to data d .

Since probabilities are often difficult to calculate — remember that we do not know or assume anything about the true distribution, particularly, that it exists— we propose a simplified method of using error measures in place of probabilities, called *normalized minimum error*:

$\bar{E}_{nml}^{\mathcal{M}_i}(d) = \frac{E(d|\hat{\theta}_{\mathcal{M}_i}(d))}{\sum_{d' \in \mathcal{D}} E(d'|\hat{\theta}_{\mathcal{M}_i}(d))}$, where $E(d|\hat{\theta}_{\mathcal{M}_i}(d))$ is the error that the minimum error model in the class \mathcal{M}_i makes with respect to data d . The best model is the one that minimizes this score.

When only few data samples are available, the NML principle tends to choose overly simple models. This does not mean that simpler models are more likely to be true (Grunwald, Manuscript). Instead, a simpler model predicts future data more reliably than a complex one, or more specifically, a simpler model predicts more reliably how well it will do with the future data. And that is what we are interested in; finding the “best” model that can give us insight of the interesting regularities in the data, which in turn may be used to predict and understand future data.

Experiments

The goal of this research is eventually to apply the models selection criterion to the class of agent-based spatially explicit learning models of land-use and land-cover change. A class of simple foraging models was chosen for the first set of experiments because of its relative simplicity in implementation and analysis. This class of models has been used to study group dynamics, and

¹What Rissanen calls model, we call model class, and the model selection criterion proposed here actually chooses between model classes.

²In the original framework of data compression and the *minimum description length* principle the regret of \mathcal{M} was defined as the number of additional bits required to encode data d , if using \mathcal{M} instead of the best fitting model for d .

complex behavioral patterns emerging from simple behaviors and local interactions. Although these models do not simulate changes in the actual landscape, they still have to do with a domain in which the spatial organization — what and where — is important. In this research the model class is not used to understand the emergent behavior, but rather to demonstrate the qualitative and quantitative properties of the proposed model selection criterion. This is done by comparing criterion’s performance using several error measures applied to the same data.

The Model Class

The model class consists of a toroidal landscape grid of 10×10 cells. The cells may contain consumable energy, either distributed randomly or organized in special spatial patterns. The total number of initial energy items is varied between 5 and 75, and the value of each is originally 1000. A small number of autonomous agents (currently 10) also inhabits the landscape. Their task is to move from cell to cell and consume energy in order to compensate the energy loss resulting from moving. Each agent can move zero or more times at each time step. When an agent ends up in a cell containing energy it consumes some of it. When an agent’s energy level drops below zero, it deceases.

The only decision the agents make is to which direction to move. Several decision strategies were designed to represent different degrees of computational complexity and to use different types and amount of information, and supposedly exhibit qualitatively different behavioral patterns. The behavioral features of interest are agent locations on the landscape and their life-span, i.e., how they move and how long they live. Brief descriptions of the decision strategies, from the simplest to the most elaborated, follows.

Random agent chooses its next location randomly from the cell it is currently located in and eight cells directly adjacent to it (these nine cells form the neighborhood of a cell).

Hill-climber agent keeps track of its moving trajectory; if its previous decision took it into a cell with energy, it keeps moving to the same direction. Otherwise, it chooses randomly from its neighborhood.

Locally greedy agent checks all the cells in its neighborhood for energy and moves to one with most energy in it. If none of the neighboring cells contains energy, it chooses randomly.

Globally greedy agent has knowledge of all the cells containing energy. It calculates the distances to them, and moves zero or more steps to the direction of the closest one.

Social learner agent does not rely only on its own knowledge of energy locations but also that of the other agents. It asks them for the cell in their neighborhood that has the most energy, and moves to the direction of the closest such cell.

These five decision strategies are embedded into two model classes each; one of the classes assumes a homogeneous group of agents, i.e., all agents have the same parameter values, and the other one assumes a heterogeneous group, i.e., each agent has its individual parameter values. Ten models, one from each class, are used both as *source models*, i.e., the data generating models, and *candidate models*, among which the best model is to be chosen.

Method

Source models are made to generate data with fixed parameter values from ten initial spatial configurations of energy. The data consists of sequences of matrices that record the number of agents present in each cell of the landscape. Each data sequence is 25 time steps long.

After data generation each candidate model is fitted to the data. The free parameters are listed in the Table 1. The fitting is done by minimizing the sum of square error (or equivalently the Euclidean distance) between the data generated by the source models and the outcomes of the candidate models. The following error functions are used: ϵ_1 : *Absolute number* of agents in a cell at each of 25 time points, ϵ_2 : *Binary indicator* of agents presence or absence in a cell at each of 25 time points, 1, if there are one or more agents in a cell, 0 otherwise, ϵ_3 : *Average number* of agents in a cell during each consecutive five time-step period, and finally ϵ_4 : *Fraction*

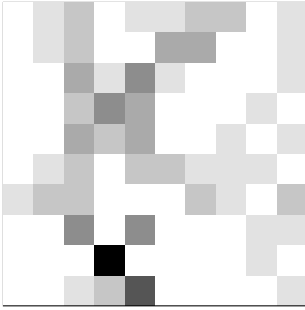


Figure 1: Absolute number of agents (ϵ_1)

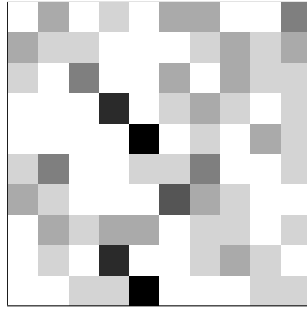


Figure 2: Binary indicator for agent absence/presence (ϵ_2)

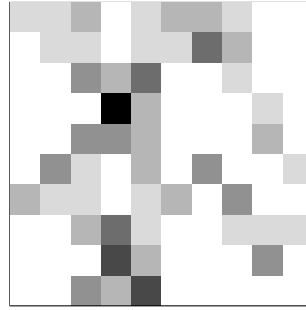


Figure 3: Average number of agents (ϵ_3)

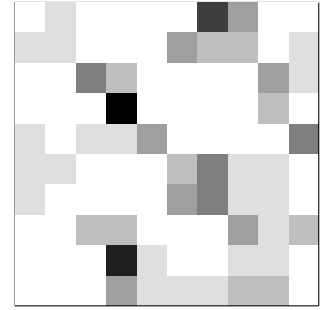


Figure 4: Fraction of time agent present (ϵ_4)

of time of five consecutive time steps that one or more agents are present in a cell. After model fitting the NME scores are calculated for each candidate model class over all data sets generated by each source model class. The model class with the minimum score is considered the selected model.

1.	Initial energy level
2.	Max number of moves per time step
3.	Intake proportion
4.	Size of social network (Social learner only)

Table 1. Free parameters calibrated to the data.

Preliminary Results

The results of the first set of experiments using the different error functions are presented in the Figures 1 - 4. The rows represent the source models, and the columns represent the candidate models in an increasing order of computational complexity. The first five, both row-wise and column-wise, are the model classes with homogeneous agents (referred to as simpler model classes later) in the order: random, hill-climber, greedy, locally greedy, globally greedy and social learner. The last five model classes, in the same order of computational complexity, have heterogeneous agents (later referred to as more complex model classes). The shades of grey represent the number of times the respective candidate model class was chosen by the NME criterion as the best model (class) for the data generated by each source model class: the darker the shade, the larger the number.

The proportion of darker cells on the left side of the Figures 1 and 3 indicates that the simpler model classes got chosen more frequently than more complex ones. The reason why the distinction is not equally obvious in the Figures 2 and 4 may be that in these experiments far fewer data points were used. In all four cases, especially for one or two of the computationally most complex model classes, the simpler one (with homogeneous agents) got selected even if the source model belonged to the more complex class (with heterogeneous agents). This implies that the extra complexity introduced by the individual parameters is not necessary to produce all the interesting regularities in the behavior.

With none of the four error functions the most complex model classes were prominently chosen. It is hypothesized that using more time points and a larger group of agents would either accentuate the trend found in the first set of experiments, i.e., the selection criterion’s tendency to prefer simpler model classes, or make the criterion able to select the “true” generating model, since it more likely captures the relevant regularities in the data.

Discussion and Future Work

The preliminary results with the proposed model selection criterion are promising, although a lot of theoretical work is required in order to establish its adequacy. Experimental work with more complex and larger data sets is also planned for the near future. The advantage of using artificial data is that the model selection criterion can be tested both by including the source model in or excluding it from the candidate models. Furthermore, as more data samples are made available, the more regularities can be extracted from them, and consequently, the criterion's choice should converge towards the real generating model. This can also be rigorously tested with artificial data. Finally, the selection criterion will be applied to real-world data of land-cover and land-use change in order to compare its performance with other model selection methods, such as cross-validation, when choosing between agent-based individual or social learning models.

References

- Axtell, R., Epstein, J., Dean, J., Gumerman, G., Swedlund, A., Harburger, J., Chakravarty, S., Hammond, R., Parker, J., & Parker, M. (2002). Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. In *Proceedings of the National Academy of Sciences of the U.S.A.* (Vol. 99: Suppl. 3, p. 7275-7279).
- Benenson, I. (1998). Multi-agent simulations of residential dynamics in the city. *Computation, Environments and Urban Systems*, 22(1), 25-42.
- Bodin, Ö., & Nordberg, J. (under review). Information network topologies for enhanced local adaptive management. *Environmental Management*.
- Brown, D. G., Page, S. E., Riolo, R., & Rand, W. (in press). Agent-based and analytical modeling to evaluate the effectiveness of greenbelts. *Environmental Modelling & Software*.
- Burton, R. M., & Obel, B. (1995). The validity of computational models in organization science: From model realism to purpose of the model. *Computational and Mathematical Organization Theory*, 1(1), 57-71.
- Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, 44, 205-231.
- Grunwald, P. (Manuscript). *Understanding the minimum description length principle*.
- Hoffman, M., Kelley, H., & Evans, T. (2002). Simulating land cover change in South-central Indiana: Complexity and ecosystem management. In M. E. Janssen (Ed.), *The theory and practice of multi-agent systems*. Edward Elgar, Cheltenham, U.K.
- Kearns, M., Mansour, Y., Ng, A. Y., & Roi, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1), 7-50.
- Laine, T., & Busemeyer, J. (2004). Comparing agent-based learning models of land-use decision making. In C. L. Marsha Lovett, Christian Schunn & P. Munro (Eds.), *Proceedings of the sixth international conference on cognitive modeling* (p. 142-147). Pittsburgh, PA, USA: Lawrence Erlbaum Associates.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44, 190-204.
- Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3), 472-491.
- Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal*, 42(4), 260-269.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358-367.