AGENT-BASED MODEL SELECTION FRAMEWORK FOR COMPLEX ADAPTIVE SYSTEMS

Tei Laine

Submitted to the faculty of the Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in Computer Science and Cognitive Science
Indiana University

August 2006

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

Doctoral Committee

Filippo Menczer, Ph.D.
(Principal Advisor)

Michael Gasser, Ph.D.

Jerome Busemeyer, Ph.D.

July 13, 2006

Marco Janssen, Ph.D.

Copyright © 2006

Tei Laine

ALL RIGHTS RESERVED

Acknowledgements

I want to thank my advisor, Filippo Menczer, and the members of my doctoral committee, Michael Gasser, Jerome Busemeyer and Marco A. Janssen, for support and guidance, and first of all for introducing me to research communities that integrate computing with other disciplines, such as decision making, learning, language and evolution, and ecology.

I am grateful to ASLA Fulbright Program for giving me the opportunity to pursue my academic interests to fulfillment, together with a possibility to educate myself both culturally and geographically.

Besides the Fulbright Foundation, my doctoral studies were funded by Computer Science Department and the Biocomplexity grant (NSF SES0083511) for the Center for the Study of Institutions, Population, and Environmental Change (CIPEC) at Indiana University. I am grateful for getting an opportunity to work in this truly multidisciplinary group of scientists led by Elinor Ostrom and Tom Evans, and share their enthusiasm in solving hard real-world problems. The collaboration allowed me to gain great insight to and appreciate the importance of environmental modeling. I would also like to take my opportunity to thank CIPEC's GIS/Remote Sensing Specialist, Sean Sweeney, and graduate students Shanon Donnelly, Wenjie Sun and David Welch for providing me with the data I used in my modeling studies.

Of course, none of this work would have been possible without the Computer Science department's superb system support group. They solved my problems in a timely manner and provided me with an outstanding environment to work in.

My friend Marion deserves to be acknowledged for her meticulous effort in proof-reading the text and making useful suggestions to improve its readability. Thanks also go to students in GLM and NaN groups — Brian, Fulya, Jacob, Josh, Mark, and Thomas — for attending my practice defense and giving me plenty of insightful suggestions to improve slides and the oral presentation.

I also like to express my appreciation of Bloomington community and the numerous friends I made here during my stay. The welcoming atmosphere of this town made it really easy to mingle in and get to know local people in private or business contexts.

Finally, I thank Tomi, my colleague, long time partner and best friend, not only for fixing me breakfast every morning and laundering my running gear, but for endless encouragement, and most importantly, great companionship in our numerous adventures in the US. We have a whole lot more miles to cover!

Abstract

Human-initiated land-use and land-cover change is the most significant single factor behind global climate change. Since climate change affects human, animal and plant populations alike, and the effects are potentially disastrous and irreversible, it is equally important to understand the reasons behind land-use decisions as it is to understand their consequences. Empirical observations and controlled experimentation are not usually feasible methods for studying this change. Therefore, scientists have resorted to computer modeling, and use other complementary approaches, such as household surveys and field experiments, to add depth to their models.

The computer models are not only used in the design and evaluation of environmental programs and policies, but they can be used to educate land-owners about sustainable land management practices. Therefore, it is critical which model the decision maker trusts. Computer models can generate seemingly plausible outcomes even if the generating mechanism is quite arbitrary. On the other hand, with excess complexity the model may become incomprehensible, and proper tweaking of the parameter values may make it produce any results the decision maker would like to see. The lack of adequate tools has made it difficult to compare and choose between alternative models of land-use and land-cover change on a fair basis. Especially if the candidate models do not share a single dimension, e.g., a functional

form, a criterion for selecting an appropriate model, other than its face value, i.e., how well the model behavior confirms to the decision maker's ideals, may be hard to find. Due to the nature of the class of models, existing model selection methods are not applicable either.

In this dissertation I propose a pragmatic method, based on algorithmic coding theory, for selecting among alternative models of land-use and land-cover change. I demonstrate the method's adequacy using both artificial and real land-cover data in multiple experimental conditions with varying error functions and initial conditions.

Filippo Menczer	Michael Gasser
Jerome Busemeyer	Marco A. Janssen

Contents

List of Tables

List of Figures

1

Introduction

Agent-based models are used in ecology, not only to understand global environmental change and human role in bioecological systems, but to inform decision makers in the process of designing environmental programs and policies. Changes are due to human actions, and they can occur in different time scales and spatial resolutions and extent — from choosing annuals to grow on one's yard to changing pristine natural resorts to urban development. Decisions are always somewhat local, even if they may have more far reaching consequences such as global climate change. Since the direct or indirect consequences of these decisions may be disastrous and at worst irreversible, it is important that the choice of the model that decision makers put their confidence on, is based on sound principles.

Computer modeling is a common research practice and theory testing method within disciplines in which the structures or processes underlying a real-world system of interest are difficult to observe and measure directly, or controlled experimentation is impossible. The theoretical assumptions of these structures and processes are implemented in a computer model, whose performance is compared to the observed data. The task left to the scientist is to choose a performance measure for the comparison, and a criterion for determining if the model adequately

explains the empirical system.

Two methods, used in testing models and choosing between them, are *null hypothesis testing*, which is commonly used in behavioral sciences such as psychology, but also in biology and ecology, and *model selection*, which is more or less an emerging approach in many fields. In null hypothesis testing one model, namely the "null hypothesis", is considered favorite a priori and is rejected in favor of the alternate hypothesis only if it fails to statistically explain the data. In model selection several candidate models are considered at the same time, and they are usually, but not always assumed equiprobable *a priori*. A model that is best supported by the observed data is chosen. If none of the models gains significantly more support than others, the selection can be deferred until there is enough evidence to choose one model over the others (Golden, 2000; Johnson & Omland, 2004).

The question of model selection has been addressed in several fields, for instance in cognitive science (Pitt, Myung, & Zhang, 2002), ecology and biology (Boyce, 2002; Ellison, 2004; Johnson & Omland, 2004; Stephens, Buskirk, Hayward, & Rio, 2005; Strong, Whipple, Child, & Dennis, 1999), genetics (Sillanpää & Corander, 2002), organizational science (Burton & Obel, 1995), sociology (Weakliem, 2004) and maybe most prominently in machine learning (Kearns, Mansour, Ng, & Roi, 1997). Cognitive scientists and the machine learning community have mostly been concerned with model complexity and overfitting. In other fields model validity, particularly, how well the model adheres to reality, is a central issue (Burton & Obel, 1995). Supposed realism, achieved by replicating real world processes and components in great detail, may introduce complexity that makes the model incomprehensible and undermines its ability to answer the scientific question it was build to answer. It is suggested that more complex models are not

necessarily more realistic than simple ones, but only more complicated.

The best model is often determined by goodness of fit to the observed data that usually consists of samples from a larger population. Using the fit as a single criterion has a danger of compromising a model's generalizability and undermining its true explanatory or predictive power. An overly complex model may fit a data sample perfectly, but it is not clear if it captures interesting regularities in the data or just random variability in the sample. On the other hand, a model that is flexible enough to fit a wide variety of data is not easily falsifiable. The goal of a model selection method is to choose the model that best explains a phenomenon of interest, and also to choose an appropriate degrees of freedom required to explain the phenomenon (Kearns et al., 1997).

If real-world data exists, the quality of performance is relatively easy to measure, but individual sources of complexity may be much harder to identify. Several approaches have been proposed to address the trade-off between goodness-of-fit and model complexity. Most of them combine a maximum likelihood term that measures fit and a penalty term that measures complexity. Traditionally, the most common factors included in the complexity term are the number of free parameters, the functional form, the value range for free parameters and the number of independent data samples (Forster, 2000; Myung & Pitt, 1997; Myung, 2000; Pitt et al., 2002).

Science favors simple explanations, since they are both more likely and more comprehensible, and thus more capable of increasing common understanding and knowledge. Modeling practice tends to follow this scientific ideal by preferring models that are simplifications, abstractions and idealizations of the system they are designed to mimic (Vicsek, 2002). This goal adhered to the principle of *parsimony*, known also as *Ockham's Razor*, which states that "entities should not be

multiplied beyond necessity."

However, many application domains of agent-based models are *complex adaptive systems (CAS)* (Bradbury, 2002) in which the large-scale behavior emerges from small-scale behavior and local interactions. The class of land-use and land-cover change models naturally falls into this category. An inherent characteristic of these systems is that the behavior of the whole cannot be understood by simply observing the behavior of individual components, so it seems apparent that modeling of these kinds of systems cannot be reduced into an analysis of the simple systems that constitute them. Particularly, the validation of the simple systems and their behavior is in most cases impossible, because no data about them exist. Neither can a complex adaptive system be abstracted into straightforward statistical or probabilistic models so that the inherent emergent properties of the original system will be preserved (Bradbury, 2002). As it turns out, models of complex adaptive systems are often complex adaptive systems themselves.

Most of the existing model selection methods have been designed with 'simple' statistical models, sets of probability distributions, in mind, with which the model selection problem reduces to an inference about the model's structure, i.e., how many parameters to include, and a search for their values. These methods barely scale up to handle models belonging to the class of complex adaptive systems, since their behavior seldom can be formulated as a deterministic function of parameter values in the application domains of any practical interest. Or at least, such a function would be extremely complicated. This in turn defies the whole purpose of modeling, which is to understand the data with the help of the model.

I adopt a very pragmatic approach to studying model selection methods for complex adaptive systems, and propose a criterion based on the practical, also called crude, version of the *Minimum Description Length (MDL)* principle, coined

for model selection purposes by Rissanen (1978, 1999). Rather than an algorithm for model selection the MDL principle is a general method of inductive inference based on the idea that any regularity in data can be used to compress them, and the model that compresses the data most is able to extract most regularities in it. The principle has several desirable properties; first, it does not assume that a 'true model' exists that generated the data, then go ahead looking for it; secondly, in the form the principle is adopted here, it does not make any subjective judgements of the structure of the model, but bases its preference for a model (over others) solely on the model's performance; and thirdly, it has a neat communicative interpretation, applicable in many practical contexts. This will be elaborated in Chapter 4.

1.1 Research Questions

In this dissertation I study model selection method for agent based land-use and land-cover change models. The research is framed by the following questions:

Question 1. What are good measures to be used to distinguish the performance of different adaptive spatially explicit agent-based models?

Question 2. What is an appropriate selection criterion to choose a model that best explains the available data?

Question 3. How does the choice of the performance measure influence the behavior of the model selection criterion?

1.2 Overview of Dissertation

The study consists of formulating the model class of land-use and land-cover change, followed by the design and implementation of a practical framework for comparing models belonging to this class, and incorporation of the proposed model selection criterion into the framework. The last phase is to conduct several empirical tests using artificial and real data, to assess adequacy and usefulness of the criterion.

I finish this introductory section with definition of terms and concepts used in the rest of the dissertation. Chapter 2 focuses on two main topics: first, the current state of the art in agent-based modeling, particularly in land-use and land-cover change, and secondly, basics of model selection. Since model validation is a essential part of the modeling process and prerequisite to model selection, issues related to validation of agent-based models are also addressed.

In the Chapter 3 I describe the agent-based land-use and land-cover change framework in which the model selection criterion, proposed in Chapter 4 is tested. I also introduce classes of learning algorithms between which the selection is done, and error functions that are used to assess the models' performance.

Chapter 4 outlines the basics of the MDL principle. The crude version of the principle is applied to the class of land-use and land-cover change model through an extended example. Finally, an enhanced version of the principle is introduced, and its tied to another, theoretically sound version of the MDL principle based on universal models (Rissanen, 1999).

The experimental evaluation of the proposed model selection criterion is conducted in three phases in Chapter 5. Experiment I, presented in Section 5.2, functions as a proof of concept; with a simple and abstract agent-based land-use and

land-cover change class the criterion's ability to identify the 'true' generating model class is challenged. Experiment II, discussed in Section 5.3, consists of a series of tests to analyze criterion's sensitivity to error functions and factors external to the model class. Finally, in Experiment III evaluates the criterion's performance with real world data. This phase is presented in Section 5.4.

Final Chapter 6 is dedicated to general discussion and outlines the direction of future work.

1.3 Terminology

One obstacle for fluent scientific discourse in multi-disciplinary research is that every participating discipline brings to the party not only their knowledge and expertise together with research practices and methodology, but also their own concepts and terminology. Some disciplines have also adopted a practice of exploiting or overriding terminology from other fields, which makes the communication between even close disciplines susceptible for misunderstandings and unnecessary disputes. Finally, different disciplines just define the terms differently.

Agent-based modeling of land-use or land-cover change is an endeavor that brings together scientists from computer science, ecology, economics, biology, geography, and even anthropology, political science and psychology. Introducing model selection, which mostly derives from artificial intelligence and statistical learning, to the set, just adds another degree of potential confusion.

Modeling as Explanation vs. Prediction

To start with, scientists coming from different disciplines use the term "model" to refer to different entities; for statisticians it equals a distribution or (point) hypothesis in a parametric family of probability distributions (Myung, 2000), while for computational economists or psychologists it may mean an abstraction of a real-world system, or theoretical assumptions assumed underlying the system, formulated as a computer program, and used to understand the system. Even more general and intuitive interpretation among sciences is "a system that behaves in a similar way as the 'real system"', 'real system' meaning the data generating process.

Rissanen in his seminal paper (1978) gives the following description for a model:

> ... 'model' is used for any hypothesis that one uses for the purpose of trying to explain or describe the hidden laws that are supposed to govern or constrain the data.

Despite being an adequate depiction of what a model actually means to many scientists, this definition is still relatively vague. Later Rissanen (1989) makes the distinction between "model as a realization of a theory" and "model as depiction of reality." Also in the former case, he argues, the theory tells us not only how the model works, but also how the real world works. This is the fundamental theme running throughout this dissertation; modeling pertains to an attempt to figure out what is going on in the real world, and the data is used to infer the processes and structures underlying the observed behavior. Here models are not used to predict the future because of the unpredictable character —- introduced by sensitivity to initial conditions, path dependency, and agent adaptation — of the models of interest, namely complex adaptive systems (Bradbury, 2002). In some

marginal sense the CAS models are also applied in evaluation of scenarios, but again, the objective is to understand behavior not to predict it. For instance, one can run a CAS model to generate a distribution of histories and then use them to understand the general underlying process (Rand et al., 2003). So, what is the difference between explanation and prediction then?

Explanation means an attempt to understand how structures and mechanisms underlying a system contribute to the observed behavior of the system. Prediction in turn is inference about what is going to happen in the future, not accessible to us yet, based on the knowledge of the current state of affairs. An explanation answers questions like How? and Why?, while prediction answers questions like What?

The purpose of this chapter is to make precise the central concepts and terms frequently used in the rest of the dissertation. First I describe the basic terms regarding the modeling enterprise in general, such as model and model class, data, goodness-of-fit and generalizability, then more advanced concepts pertaining to model comparison and selection. The chapter is closed by an introduction to the terminology used in the application domain, namely in modeling of land-use and land-cover change. However, some of these specific terms may later be used also in their everyday meaning; in such a case, an attempt is made to accompany them with a note of the intended reading.

Model

In general terms a *model* in this dissertation means either a running computational algorithm or procedure implemented in any general purpose programming language or a verbal or mathematical theory formulated precisely enough to be instantiated as a computer program. In either case, the model is a collection of

structures and processes assumed to underlie the behavior of the system of interest. Given this basic presumption, the following characterize different depictions of models in the process of modeling and model selection.

Model class contains models with the same functional or algorithmic form, that have the same number and type of parameters.

Model is an instantiation of a model class after the parameters are fixed; either set by the experimenter or estimated from data.

Generating model is the assumed data generating process, either nature or the model devised by a scientist.

Candidate model is a model among a set of competing models we consider in the comparison, and among which we want to select one.

True model is the assumed data generating process. This is a feasible concept only when using artificially manufactured data. Only in such a case — when the generating model is devised by a scientist — it is known that a true model exists. When we are working with real-world data, we do not consider a 'true model,' since assuming its existence may be dangerous, and its verification close to impossible.

Optimal model is the best-fitting model in the *maximum likelihood (ML)* sense.

Best model is the model that is 'best supported by the data' (Johnson & Omland, 2004), a requirement that is ultimately determined by the scientist. In the current research the best model is the one that captures the most useful regularities in the data with an appropriate level of flexibility. In other words, the best model is one that teaches us something interesting about the data, which can be used to understand the system or process that generated it.

Null model is baseline model used in the model comparison process, a close equivalent to a null hypothesis in traditional statistics.

Data

Data denotes whatever numerical and unprocessed (i.e., not summary statistic) information either output by a model when run (*artificial data*), or acquired from real world by using other media and methods (*observed data*). Data may represent quantitative information of a single event or a series of events.

Metrics refer to various numerical measures calculated from the data that characterize it either quantitatively or qualitatively.

Sample usually denotes to a sequence of observations, where each observation is an outcome of some process or system. In land-use and land-cover change context a sample is a sequence of land-cover changes observed over time.

Model Selection

Model selection in general terms is form of statistical inference the goal of which is to identify among the candidates the best model of behavior after observing samples of that behavior. More specifically model selection is a process whose outcome is the model, which outperforms other candidate models according to a predefined selection criterion, or in case the criterion is not conclusive, i.e., it is not able to distinguish between the candidate models, a decision to defer the selection until more evidence is gathered. The inputs to the process are a set of candidate models, the method of measuring fit of the

models (error or loss function), and the model selection criterion for determining the best model among the candidates. The decision how to come up with a representative set of candidate models is entirely different issue, not dealt with here, that reflects general goals of the study, the research paradigm and its history.

In statistics model selection is used to estimate parameter values for a known parametric form, not the structure of the model. Presupposing a certain structure or functional form for an adaptive agent-based model is a simplifying if not preposterous assumption, as if saying that we know which one is the 'true model.' Therefore, this research is about selecting between model classes. For the sake of fluency, and in accordance with common practice, the term 'model selection' is used in this dissertation instead of 'model class selection.'

Model selection criterion is a numerical measure for determining which of the candidate models is the best with respect to a modeling objective. The selection criterion does not say anything about a model's adequacy for its purpose, other than how well the model outcome complies with the observed data. More importantly, it does not validate the model's structure, functionality or other assumptions built into it. The consistency and plausibility of model assumptions pertaining to the modeled system need to be considered when choosing the candidate models. The selection criterion is not able to distinguish a plausible model from an implausible one with equal performance. A substantial amount of subjective judgement is left to human scientists to decide if the model complies with well established theories in the field, and the empirical observations or common knowledge of the modeled system.

Goodness of Fit (GOF) is the deviation between the model outcome and the data

measured after the model parameters have been calibrated to the data so that the deviation is minimized.

Maximum likelihood parameters are the parameter values that maximize model fit or alternatively minimize the lack of fit.

Generalizability is a model's explanatory or predictive accuracy on yet unseen data.

Flexibility is a measure for a model's ability to fit a variety of different data patterns.

Complexity in turn is used to refer to the intricacy of the model, which in general terms, refers to the amount of detail built in the model of the real-world domain. More specifically, model's complexity is a function of both, the number of interacting components, and the extent and refinement of computation. Rather than making the model too flexible to fit wide variety of different data patterns, complexity makes it perfectly replicate a single or very few samples.

Land-use and Land-cover Change

Land-use decision making is a complex, multi-asset, real world decision task. The land-owner has to consider which activities she wants to implement on her land and decide where on that land to implement them. The decision maker's task is to find an effective way of using her assets — size and quality of land, technology, education and experience — in allocation of available resources — labor and land — to different uses. The number of factors to be considered range from the suitability of the land, dictated by various biophysical variables, to the expected

monetary or non-pecuniary returns from the activities. The optimal or good decision does not depend solely on the careful consideration of the afore-mentioned factors, but also on the decisions of neighboring owners and the activities they implement on their land.

The outcome of the decision process is a land-use or a change in the land-use. In this research, I am primarily interested in changes that are human-initiated, although various natural phenomena have at least a partial role in all change. Clarifications to some of the frequently occurring terms follow.

Land-Cover and Land-Use Land-cover is any of the biophysical attributes used to characterize the condition of the landscape (Brown, Pijanowski, & Duh, 2000). Land-use in turn refers to the human activity on the landscape that is influenced by various economic, social, cultural, political, ad historical factors (Brown et al., 2000). Land-use and land-cover are intercorrelated, but not identical, and in some context they are treated equivalent. Most of the time, but not always, land-use has visible effects on the land-cover. However, the land-cover can change without the land-use changing. In order to retain a reasonable level of abstraction and simplicity, in this dissertation both land-use and land-cover, used interchangeably, refer both to the biophysical condition of the landscape and the agent activity that results in the condition.

The encoding of the land-cover can be qualitative characterization, such as old growth forest, secondary succession, wetland, or pasture, or binary classification to, for instance, forest and non-forest or urban and rural. Agent activities corresponding to these cover types could be recreational activities, such as hunting or hiking, timber harvesting or development.

Land-cover Change is a process in which the biophysical properties of the landscape changes as a result of a natural phenomenon (e.g., wildfire or forest

growth) or human land-use activities (e.g., development, logging). Land-use and land-cover change is often abbreviated LUCC.

Parcel is region of land owned by a single economic agent, for instance individual, organization or company.

Spatial or Spatially Explicit (also called 'spatially referenced' in literature) means that together with the quality or type of a biophysical variable, its location on the landscape is explicitly encoded, as opposed to just recording the quantitative or aggregate measures of the variable. Baker (1989) makes this more explicit in the context of land-use change models. He distinguishes between *whole landscape models*, *distributional models* and *spatial landscape models*. While the whole landscape models relate to the change of a single variable or a set of environmental variables associated to the whole landscape, distributional models track the changes in the distributions of variable on the landscape. Spatial landscape models focus on both configuration and physical locations of the changes in the variable values on the landscape.

Externality in general terms means a benefit or cost resulting from a decision that is enjoyed or born by others than the decision maker herself. Spatial externality means an effect caused by land-uses on the adjacent parcels. The effect can be positive or negative, and it can be between the same or different land-uses.

A positive externality means increase in revenue or some other valuable asset induced by the neighbor's land-use decisions. A negative externality is the cost incurred by a land-owner resulting from the neighbor's decisions, when the neighbor does not account for all of the costs herself.

Suitability is an indicator of land's goodness for various purposes.

Heterogeneity and homogeneity refer to how some property is distributed over entities or entity. For example, landscape heterogeneity indicates that the landscape varies in some characteristic, e.g., slope or soil, from location to location, while a homogeneous landscape means that the characteristic is equal in every location. Likewise, agent heterogeneity means that the agents vary by one or several attributes, while homogeneous agent are equal in this aspect, i.e., they have the same attribute value(s), e.g., age or education.

2

Background

2.1 Agent-Based Models of Land-use and Land-cover Change

The fundamental idea behind *agent-based models (ABMs)* is that decision making is distributed among autonomous actors, which either operate individually or may communicate and cooperate. The focus is on the macro-level patterns in collective behavior emerging from agents' individual characteristics and micro-level phenomena, such as local behavior and interaction between the agents.

ABMs come in multiple disguises but here I am particularly interested in models in which agents inhabit a simulated environment, so that they are 'physically' tied to a specific location and have a fixed neighborhood. Alternatively, if the spatial aspect is not important, an agent's environment and neighborhood can be defined by other agents it interacts with.

The agents perceive the state of the environment, and then act according to the information they possess. They may change either some objects in the environment or themselves, for instance by moving relative to the other agents. Agents may be

intentional and have goals and actively change the state of the world in order to achieve their goals by following an internalized decision strategy. Besides goals, agents may have cognitive properties such as emotions, needs and memory, and they may learn from their own or other agents' actions. They may also choose to interact with other agents in order to seek information, or communicate their intentions or some properties of the environment to them.

The agent-based approach has been applied to studying, for instance, social dynamics and communication and collaboration under environmental risk (Andras, Roberts, & Lazarus, 2003; Axelrod, 1984; Schelling, 1978), ecological economics, e.g., commons dilemmas (Jager, Janssen, Vries, Greef, & Vlek, 2000), military conflicts (Cioffi-Revilla & Gotts, 2003), types of complexity in artificial life applications (Menczer & Belew, 1996), language evolution (Bartlett & Kazakov, 2004) and language change (Laine & Gasser, 2003), people-environment interaction for recreation management (Deadman & Gimblett, 1994; Itami & Gimblett, 2001), and agricultural economics, e.g., land-use and land-cover change (Berger, 2001; Cioffi-Revilla & Gotts, 2003; Deadman, Robinson, Moran, & Brondizio, 2004; Evans & Kelley, 2004; Laine & Busemeyer, 2004b; Parker, Manson, Janssen, Hoffman, & Deadman, 2003). Janssen (2004) lists other applications of agent-based models in ecological economics: innovation diffusion, learning in natural resource management, and participatory approaches. Grimm (1999) reviews what he calls *individual-based models*; these models simulate animal population dynamics emerging from individual characteristics and behaviors. Tesfatsion (2002) lists potential application domains of agent-based modeling in computational economics, for instance learning and embodied cognition, design of agents for automated markets, study of organizations, and experiments with human subjects and computational agents, just to mention few.

The agent-based approach has also been used to model various land-use and land cover change related processes in several areas of the world: for instance agricultural land-use decision making by colonist households in Brazilian Amazon (Deadman et al., 2004), migration and deforestation in Philippines (Huigen, 2004), agricultural household land-use decision making in the US Midwest (Hoffman, Kelley, & Evans, 2002; Evans & Kelley, 2004; Laine & Busemeyer, 2004b, 2004a), reforestation in the Yucatan peninsula of Mexico (Manson, 2000), ex-urban development in Maryland, US (Irwin & Bockstael, 2002), spatial planning in Netherlands (Ligtenberg, Bregt, & van Lammeren, 2001), and technology diffusion and resource utilization related to agricultural land-use in Chile (Berger, 2001). In Janssen (2002) several other application domains of agent-based simulation and modeling studies have been presented, for instance the effect of policy switches in several farm-related variables, innovation diffusion and adoption of organic farming, interaction of social, economic and ecological variables in household's agricultural decision making, and finally management of grazing on rangelands. The study areas extend from Western Europe and the Midwestern United States to Africa and Australia.

Models of LUCC

These days land-use change is one of the most prominent forces affecting the planet we live on. Besides its local effects, such as potential animal and plant habitat destruction and contamination of ground water supplies, land-cover change also has irreversible effects on global climate (Agarwal, Green, Grove, Evans, & Schweik, 2002). The general objective of modeling land-use and land-cover change (LUCC) is to understand this global environmental change and the human impact

on bio-ecological systems. More specifically, the goal is to explain how various social, economic and ecological factors influence land-use and resulting land-cover patterns in multiple spatial and temporal scales.

Empirical measurements are not necessarily sufficient to understand the combination of the forces driving the change (Parker et al., 2003). On the other hand, experimental manipulation of landscapes is often impractical if not impossible or unethical (Baker, 1989). Therefore, computer models may be used to study the social, psychological, and bio-ecological processes that are assumed driving the land-use. By testing possible explanations for a phenomenon one can explore implications of theories and formulate new hypotheses. For instance, in order to understand the observed patterns of land-cover change within a time period, one needs to explain how people make decisions, and what role individual preferences or learning and communication play in decision making, or alternatively how decision-maker and landscape heterogeneity evidence themselves in land-use outcomes (Kelley & Evans, under review; Laine & Busemeyer, 2004b; Deadman et al., 2004; Schneider & Pontius, 2001). Computer models may also have a descriptive role in the evaluation of policies, or they can be used in decision support systems to inform decision makers in natural resource management about the potential consequences of their decisions (Agarwal et al., 2002; Baker, 1989; Berger, 2001; Casti, 1997; van Daalen, Dresen, & Janssen, 2002; Itami & Gimblett, 2001).

A number of different techniques have been used in modeling land-use and land-cover change, for instance equation-based models, logistic regression models based on suitability maps (Schneider & Pontius, 2001), system dynamic models, statistical methods, symbolic or rule-based systems combined with qualitative expert knowledge, evolutionary models, such as genetic algorithms, and perhaps

most commonly cellular automata (CA) and Markov chain (MC) models or combinations of them (Brown, Riolo, Robinson, North, & Rand, 2005; Jenerette & Wu, 2001; Parker et al., 2003). Cellular models are most suitable to study spatial interactions, while MC models lend themselves to modeling of state transitions. In landscape modeling a CA is a two-dimensional grid of cells that can be in one of a finite set of states at a time. The spatial dynamics are implemented by changing cells' state according to fixed rules, so that the new state depends a cell's current state and the state of its neighboring cells within specified temporal margin. Time is modeled in discrete steps and state transitions may be either synchronous or asynchronous.

Land-cover changes are often initiated by human decisions. Computer simulations have been used in ecological modeling for a long time. However, most of the modeling efforts have concentrated on the biophysical processes rather than human actions (Itami & Gimblett, 2001). Even if there have been attempts to incorporate social processes into a cellular framework by translating them in terms of forces applying to physical systems (Rand et al., 2003), intentional cognitive states and adaptive behavior is difficult to implement in a system whose dynamics are based on the immediate neighborhood and the finite number of fixed transitions rules (Berger, 2001; Ligtenberg et al., 2001).

On the other hand, many mathematical and statistical models ignore the spatial aspect of the land-cover change (Manson, 2000), as do the majority of models that incorporate socio-economic drivers of LUCC (Parker & Meretsky, 2004). A relatively recent development in LUCC modeling is a hybrid approach that combines a cellular automaton, representing the biophysical landscape with an agent-based component, which consists of decision makers, institutional or individual. Land-use then represents the link between the agent and the landscape (Parker et al.,

2003; Evans, Sun, & Kelley, in press).

Since the main motivation behind this dissertation is choosing among learning and decision models, a quick review of learning and decision making in both in laboratory experiment and contemporary LUCC modeling is given next.

Learning and Decision Making in Agent-based Models of LUCC

Learning in decision tasks is common subject in studies that combine laboratory experiments with model selection. Busemeyer & Myung (1987) address learning in a resource allocation experiment in which participants allocate land between three crops in an artificial setting. The subject repeat the decision until they meet the learning criterion. After each decision the payoff is displayed. Subjects' goal is to maximize their total payoff and finish the experiment as soon as possible, i.e., minimize the number of rounds they need to reach the criterion. Two learning models are compared to the human performance: functional learning model and hill-climbing model. The results indicate that a hybrid of these may be used by the subjects; functional learning strategy for exploration and hill-climbing for exploitation.

Rieskamp, Busemeyer & Laine (2003) use an abstract resource allocation task in comparing two learning algorithms to experimental data. In the experiment 20 participants repeatedly allocate a fixed amount of money between three assets, of which one produces a constant return and the payoff from the other two depends on the proportions allocated to them. The payoff function is constructed such that it contains a single global optimum and in addition to it, one local optimum. The goal of the study is to find out, if the participants are able to learn the location of the global optimum from the feedback they receive of their allocations.

The two algorithms are the *local adaptation (LOCAD)* model, that makes small adjustments to the allocations based on the feedback received from the previous allocation, and the *global search (GLOS)* model, that keeps track of the all the allocations tried so far and probabilistically samples the allocation space to find areas that have proved profitable before. The authors conclude that the LOCAD models gives a little better account for the participants' behavior, although neither of the models is able to produce accurate predictions. With one exception, the models are usually too conservative in their predictions compared to the experimental results.

Evans, Sun & Kelley (in press) study land-use decision making with an ingenious computer-aided laboratory experiment in which participants make abstract land-use decisions on a two-dimensional landscape. The participant each control a portion of the landscape, and they can see the outcomes of the decisions of other participants. After each decision they receive a reward that depends on the land-use structure they chose and potentially the landscape suitability. Evans *et al.* compare the experimental results to the decisions of an expected utility maximizing model.

The comparisons reveal two clear trends: there was more variation in the participants' payoffs than in the model's, and the participant produce much more heterogeneous land-use patterns. While the expected utility maximizer found the optimal land-use pretty quickly, the participants deviated from it significant number of times even if they could observe the price trends of the land-uses after each decision. Also, even if the price trends were predictable and the suitability pattern regular, the participants produced highly irregular land-use patterns compared to the model.

Model selection has also been discussed in context of other kind of decision tasks, most commonly in strategy learning and games. Salmon (2001) compares

models from two classes, reinforcement learning and belief update models in identification of learning rules used by subjects in normal form games. He generates artificial data from different sources — different versions of the Camerer & Ho's experience-weighted attraction model (EWA) (Camerer & Ho, 1999) – and evaluates the learning models in their ability to identify the data generating process. Salmon attributes the model's poor performance partly to the experimental design rather than solely on the models' inherent disability to recognize the decision and learning processes.

Since the introduction of a human component into LUCC models various decision making and learning techniques have been employed, for instance decision trees, constrained expected utility maximization, genetic algorithms, and rule or search based heuristic strategies. While some of the decision algorithms are relatively general in character, some of them are complicated and highly specialized to the task.

Deadman *et al.* (2004) apply a domain specific heuristic decision tree to agent decision making when studying if heterogeneity in household composition, household wealth, soil quality and burn qualities lead to quantitative land-use outcomes comparable to the trends observed in real world study area of Amazon rain forest near Altamira, Brazil. Agent learning has not been implemented yet.

Huigen (2004) proposes an agent-based framework, called *MameLuke*, for studying human-environment interaction and land-use change. In this framework agents are classified into categories, which are user definable and determined by the study objective, so that each agent can belong to multiple non-conflicting categories. Decision making is rule-based and implemented in so called *potential option paths (POPs)*, actions available to an agent, which depend on the agent's category. The framework can be applied from migration and settlement of agents to consequent

land-use decisions and activities. Huigen applies MameLuke's settlement model in the San Mariano watershed in Philippines, and simulates demographic variables, such as population age and household sizes and their effect on the spatial distribution of new settlements.

Hoffman, Kelley & Evans (2002), Evans & Kelley (2004) and Laine & Busemeyer (2004b, 2004a) use different approaches to model households' annual agricultural land-use decision making in rural South-central Indiana, and evaluate their models against real land-cover data. Hoffman *et al.* and Evans & Kelley's agents follow myopic constrained expected utility maximizing strategy; after observing a set of exogenous variables, such as prices and biophysical properties of their land, agents opt for optimal labor allocation over the available land-use activities, and the optimal locations for these activities.

While Hoffman *et al.* and Evans & Kelley's agents look ahead one year at the time, Laine & Busemeyer's models look back either one year or a longer period of time, and adjust their current decisions based on the payoffs earned in the past from different activities. In other words, the agents learn. Two algorithms, based on reinforcement learning (Sutton & Barto, 1998), are compared; the *local-adjustment agent* examines the payoffs received from different activities in last two decision rounds, and adjusts the corresponding land-use allocations according to the sign of the difference between the payoffs. The *experience-based agent* stochastically samples all the previous land-use allocations and chooses the best proportional to the total payoff it produced.

Manson (2000) incorporates both individual decision making agents and institutions in the framework that combines an agent-based component to a generalized cellular automaton[1] to simulate biophysical dynamics in the Yucatan Peninsula, Mexico. Three different decision models has been implemented: in the first one agents follow heuristic rules when choosing what to do on their land and where; in the second one they use multi-criteria evaluation based on agent and landscape variables; and in the third one agent strategies are evolved using genetic programming.

Berger (2001) uses an agent-based system to model technological innovation diffusion among agricultural decision makers in Chile. The adoption of new water-saving technologies is driven by the incentive for higher income generated by production of export goods, which, in order to succeed, necessitates new technological advances. The agent households make several decisions, e.g., innovation adoption, tenure, production, and land-use decisions, implemented as simple linear programming problems, which each agent solves separately in order to maximize household income without over-utilizing the available resources. The modeling framework includes both economic and hydrological processes. Although currently used to study implications of the policy change, the framework can also be extended to model land-cover change.

Irwin & Bockstael (2002) study the fragmented pattern of ex-urban development in Maryland. They model the timing of private landowners' decisions to subdivide their land for development. Agents opt for optimal timing taking the net returns from developing the parcel, and the foregone agricultural returns into account, modified by the interaction effect which is a function of the number of neighboring parcels that are already developed. The authors postulate a negative

[1] 'Generalized' implies that the adjacency requirement in the cells' state change rules is relaxed.

interaction effect between neighboring agents to explain why neighboring parcels are developed at different times, even if the observed positive spatial externalities would support early development. The positive externalities pertain to land's value once it gets developed if the neighboring land also will be developed, together with potential sense of community resulting from a compactly developed neighborhood. The negative interaction effect comes into play if agents are concerned with congestion or potential loss of aesthetic environment.

While the above discussed modeling studies are interested in agents making decisions about their individual parcels, Ligtenberg *et al.* (2001) proposes a model of urban planning in which agents make decisions about common land. The model combines a multi-actor approach with a cellular automaton in SWARM. They apply the model with three types of agents, with different preferences and voting powers in a single-use framework to predict the spread of urban areas in the eastern Netherlands. While all the agent types have a right to vote for the land use, only one of the agents, called the planning agent, can actually implement the land-use change if it is agreed upon by all the agents.

Validation of LUCC Models

While Quadrat-Ullah (2005) argues that structural validation against real-world domain knowledge is strict enough a test for model's validity, i.e., whether the model generates 'right behavior for right reasons,' other scientists strive for more objective and robust behavioral measures in order to build confidence in their models. By structural validation it is meant that the components and algorithms in the model replicate the systems and processes of the real-world phenomenon accurately enough the model to be considered a plausible, and a model of the phenomenon. Bayarri (2002) points out that without validating model behavior

against real data it may be hard for model's designer to convince the scientific community about the correctness and adequacy of the model structure.

LUCC models are most commonly evaluated by their spatial outcomes using several collective or individual metrics to characterize the landscape composition and pattern. Some researchers also validate their models against well-established theories, household surveys, census data, expert knowledge, laboratory or field experiments or other computational models (Carpenter, Harrison, & List, 2005; Parker et al., 2003; Manson, 2002; Tesfatsion, 2002).

A common agreement in the field is that models should be validated both qualitatively by the type of changes and spatially by the location of changes (Brown, Page, Riolo, Zellner, & Rand, 2005; Pontius, Huffaker, & Denman, 2004; Parker et al., 2003). Spatial comparison to observed data can be carried out in several ways. One possibility is *separation through time* (Pontius et al., 2004); if time series of landscape data for rounds $1, \ldots, N$ is available, the model is fitted — its free parameters calibrated — to the first M landscapes, where $M < N$, and thereafter made to predict the rest of the series, i.e., the landscapes $M + 1, \ldots, N$. Another method is *separation through space*; if several data series of the same geographic region are available, the model is fitted to one of them, and then validated with the other(s).

Pontius *et al.* (2004) claim that not a single study exists in which the model's predictions of exact location have been more accurate than the Null model's, the model that predicts no change, when using the resolution in which the data is available. On the other hand, Jenerette & Wu (2001) argue that prediction of the pattern of ecological processes is much more important than prediction of exact location. Neither Pontius *et al.* or Jenerette & Wu use an agent-based model, though.[2]

[2]Since validation of spatial outcomes is not specific to agent-based models, in this chapter and

Pontius *et al.* use suitability maps derived by logistic regression from real land-use changes between the first two time points to predict the location and magnitude of changes between latter two time points (Schneider & Pontius, 2001). Their study area is the Ipswich watershed in Massachusetts, USA. Jenerette & Wu use a cellular Markov chain model to study the effects of urbanization on the desert landscape in the Phoenix area in Arizona, USA. They use a version of genetic algorithm to learn model parameters.

Manson (2002) presses the importance of spatio-temporal validation of agent-based land-use change models, and raises some concerns both in using aggregate measures and pattern indices, but does not explicate what the apparent problems are, other than that they are usually related to scale and resolution.

Scale, Resolution and Spatial Metrics

Scientists working with spatial real-world data are often not fortunate enough to have several data sets from the same phenomenon or the same geographical area in order to conduct extensive validation or generalizability tests with their models. Furthermore, seldom do they have actual data of the decisions that lead to different land-use outcomes, but instead they have data on the outcomes themselves and the observable bio-physical processes occurring on the landscape. With a single or few data sets at hand, the only option for a scientist is to be careful when choosing the spatial or aggregate metrics, and applying them in an appropriate level of temporal scale and spatial resolution when validating her models.

Evans & Kelley (2004) test their Indiana model in various spatial resolutions, and obtain the best fit with the finest resolution, with 60m$\times$60m cells, compared

the next I will also discuss a couple of studies that do not use agent-based LUCC models, but some other type of an architecture.

to coarser resolutions of 90, 120, 150, 240, 300 and 480 meters. The authors suggest that the decreased accuracy with lower resolutions is due to the lost agent and land-cover heterogeneity resulting from cell aggregation.

Brown *et al.* (2005) also address the problem related to aggregate landscape measures. They emphasize the distinction between models that predict a certain phenomenon right most of the time and models that predict different things right at different times. They argue that aggregate spatial metrics do not necessarily enable the distinction, since they ignore the initial conditions or path-dependent spatial processes, which may play a crucial role in producing model outcomes. In other words, totally irrelevant or spurious processes may lead to accurate predictions of aggregate metrics. Spatial metrics, while potentially leading to less accurate predictions of exact locations, enable prediction of outcomes that are generated by the assumed underlying processes, not the artifacts due to uncertainty in behavior.

Parker & Meretsky (2004) demonstrate with a simple stylized agent-based model, how various spatial metrics can identify socio-economic implications and landscape patterns that result from spatial processes, such as edge-effect externalities and transportation cost. Just to name few, some of the metrics they consider are landscape composition, i.e., the proportion of the landscape in different economic land-uses, number of patches/mean patch size, mean nearest-neighbor distance, total contrasting edge and area-weighted mean shape index. Jenerette & Wu also (2001) use a relatively extensive set of spatial indices in validating their cellular urbanization model: landscape composition, largest patch size, number of patches, edge density and mean nearest neighbor. Evans & Kelley (Evans & Kelley, 2004) and Laine & Busemeyer (2004b) use landscape composition, basically forest cover percentage, and total edge length to validate their Indiana models. Some

researchers just use a single metric to track landscape changes throughout time; trends in different uses (Deadman et al., 2004), changes in composition (Ligtenberg et al., 2001), location of changes (Irwin & Bockstael, 2002), or demographic instead of spatial variables (Huigen, 2004).

Summary

In general, the choice of the model validation methodology should be exclusively driven by the purpose of the model (Casti, 1997; Burton & Obel, 1995; Pontius et al., 2004), i.e., what is the scientific question one wants to answer with the help of the model, and how accurate one wants the answer to be. Furthermore confusing calibration with validation confounds the practice of choosing the model for sound reasons, since it suggests that the model that best fits the data, should be trusted most. The rest of the dissertation is devoted to discussion on the current state of the art in model selection, addressing some of its shortcomings, and eventually proposing a practical method for selecting between agent-based models of land-use and land-cover change.

2.2 Model Selection

Many scientific disciplines that have experimentation in their methodological repertoire, use *null hypothesis testing* in evaluating theories. In this method the confidence in the *null hypothesis* — i.e., the no-effect or no-difference hypothesis — is statistically tested against the alternative hypothesis, i.e., empirical or theoretical claim that the scientist wants to make about the effect or the difference. For the sake of simplicity, I assume that hypotheses are explanations for a phenomenon.

Depending on the outcome of the test the null hypothesis is either rejected in favor of the alternative hypothesis or not rejected. This method has at least two apparent shortcomings. First, the null hypothesis is considered favorite a priori even if there is no apparent reason to believe in its higher likelihood. Secondly, if the null hypothesis is rejected in the lack of evidence to support it, the goodness of the alternative hypothesis, particularly with respect to other possible explanations, remains unexplained. The method itself does not guarantee that the alternative hypothesis is an adequate explanation of the phenomenon nor does it indicate how good it is. Weakliem (2004) lists still another well known objections for null hypothesis testing, namely the influence of a sample size; with large enough samples the nearly all hypothesis are rejected. Stephens (2005) argues that in ecological sciences null hypothesis testing often leads to trivial hypotheses, and the testing aims for statistical significance instead of biological significance.

For a scientist who wants to test several competing theories or explanations, the null hypothesis testing is not a viable option. First of all, not always there is a good reason to select one of the alternatives as a favorite. Moreover, the outcome of pairwise comparisons of theories is influenced by the order the comparisons are conducted. Therefore, in order to select among several competing theories such a method is needed that, among other requirements, gives the alternative theories an equal footing, and whose outcome is not biased by the way the comparisons are made. Furthermore, a scientist may be interested in relative ranking of alternative models, rather than choosing the single best model, especially if there is not much difference in their performance. This is where *model selection* and *model selection criteria* come into play.

The question of working with several possibly competing or contradicting hypotheses was addressed as early as in the end of 1800's (Chamberlin, 1890). Several methods for choosing among multiple alternative models have been proposed since, but model selection has been a prominent approach mostly in computer science and machine learning. However, in the last few years the approach has gained in popularity also within behavioral and social sciences. This trend is illustrated for instance by special issues in model selection both in mathematical psychology and sociology: *Journal of Mathematical Psychology* published a special issue in March 2000 and then again in April 2006, and *Sociological Methods & Research* in November 2004.

Model selection is also an ardent topic in disciplines in which modeling has not been a prominent approach, such as ecology and biology. The increased computing power has changed both the method of making science in these fields and the analysis of results (Boyce, 2002; Ellison, 2004; Johnson & Omland, 2004; Sillanpää & Corander, 2002; Stephens et al., 2005; Strong et al., 1999).

Objectives of Model Selection

As often as the modeling goals vary, the model selection goals vary as well. What do we want the model selection criterion to achieve?

There are both pragmatic and philosophical goals. Some of the pragmatic goals Grünwald (2005) lists are introduced to decide between general theories, to gain insight for future experimentation, to determine functional dependencies between variables in order to select the pivotal ones, and finally to guide prediction. While the pragmatic reasons are fundamentally about why to use model selection as a part of scientific practice, the philosophical considerations are about what should

be selected when selecting between models, e.g., should we select a best fitting model or use some other criterion to find what we are looking for.

If models were just simple and non-parameterized distributions, the task of choosing among them would reduce to finding the best fitting one. But this seldom is the case; models vary in their functional form and they usually have one or more free parameters that are estimated from the data. The model fit is conditional on the parameter values. Thus, the model selection essentially becomes a task of finding values for the (unknown) parameters. More formally, the goal is to select a particular density (point hypothesis) from a set of competing models (Forster, 2000)[3], where density is a particular assignment of the parameter values, namely the maximum likelihood values. Because the effect of the functional form relative to the number of free parameters is imperative when sample sizes are small (Pitt et al., 2002), as they often are in ecological models, the model selection criterion should be sensitive to it. In the CAS framework the functional form may be conceived as one potentially infinite dimensional parameter.

From the Cognitive Science perspective, Myung (2000), Myung & Pitt (1997), Pitt *et al.* (2002) and Pitt & Myung (2002) argue that the goal is to select a model that best captures the underlying mechanism of the mental process, or choose a model that is the best approximation of the mental process that generated the data. Kearns *et al.* (1997), Busemeyer & Wang (2000) and Lendasse *et al.* (2003) state explicitly that the goal of model selection is to minimize the generalization error, i.e., the error the model makes with respect to data not used to calibrate its parameters. While Kearns *et al.* compare different methods to find out an appropriate hypothesis complexity, the Busemeyer & Wang propose a methodology for testing model's ability to generalize to a new experimental design when it has been calibrated to

[3]In statistics, 'model' is what I call model class, and densities within a model are instantiations of my model class, i.e., models (cf. Chapter 1.3).

another. This latter approach guards the modelers from arbitrary experimental artifacts that may influence the model's behavior. Lendasse *et al.* compare several cross-validation and bootstraps methods applied to time series prediction models.

The classical frequentist or Bayesian model selection approaches assume that a 'true' process (or for statisticians a distribution) exists that generates the observed data, and the ultimate goal of the model selection is to find the model that gets closest to the 'truth.' Grünwald (2000, 2005) and Rissanen (1978, 1999) argue that the assumption of the 'true' model is unfounded, if not even preposterous, since the existence of such can never be verified, and suggest that the goal of model selection is to find the model that compresses the data efficiently by extracting most regularities. In other words, the objective is to find a model that can teach us something interesting and/or useful about the data. In my dissertation I adopt this point of view for reasons explained in Chapter 4.

Simplicity vs. Complexity vs. Flexibility

I open this section with a disclaimer about the terminology; what is generally termed 'complexity' in modeling literature, i.e., the characteristic of a model that makes it fit well a wide variety of data patterns, I call 'flexibility', and dedicate the term 'complexity' exclusively to characterize systems that are complicated by structure or underlying processes, such as models in the family of complex adaptive systems.

Why do I adopt this twist of terminology? I consider the concept of 'flexibility' more appropriate than 'complexity' for several reasons. First, complexity as a term has been so burdened with multiple meanings and contexts of use, that is almost impossible to be clear enough about the intended reading. Secondly, complexity

has become more or less a culprit that one wants to get rid of; however, in the class of models of interest in the study, complexity arises from the real-world domain and consequently is an inherent part of the model. Finally, complexity, per se, is not a problem, the problem is the number of dimensions along which the model can be easily made to fit to variety of data, i.e., over-specified models that have more parameters or explanatory variables than the hypothetical 'true model.'

Good scientific practice prefers simple models, theories or explanations, since they are likely both more probable[4] and more comprehensible than their complicated counterparts. Motivation for simplicity may come from the modeled domain as well. For instance, Chater & Vitanyi (2003) argue that simplicity is also a driving principle in human cognitive system, but admit that this assumption is difficult to test empirically. However, simplicity is always relative to the chosen representation, and the simplest patterns or interpretations may not be interesting after all. Chater (2005) suggests that the human perceptual system, being a system that makes inferences about the structure of the environment from sensory input, may use something like the Minimum Description Length principle (Rissanen, 1978) in the process of choosing among several competing interpretations. In other words, it would prefer interpretations that require shorter descriptions.

The simplicity is not only for convenience but more importantly it safeguards us against the illusion that we know more than we actually do; excess complexity, and also flexibility, may make the model look better than it actually is and for spurious reasons. All real-world data, often samples of a larger population of behaviors, contains random variation due to the errors in the collection procedure — measurements or observations — or uncertainty in the process that generates

[4]Simple hypotheses are more probable in the sense that the joint probability of multiple factors is always lower than the probability of a single factor. However, this does not mean that simple hypotheses or simple theories are more likely true.

the behavior. If fit is the sole evaluation criterion for the model's goodness, overly flexible models, for instance ones with many free parameters, can be easily made to fit all these anomalies, byproducts of errors and noise, without capturing the regularities underlying the behavior. A model like this does not really inform us about the interesting patterns that may exist in the population, but just reflects the idiosyncrasy present in each individual sample. This is called overfitting.

While excessively flexible models are prone to overfit, very simple models equally likely will underfit. As Peter Grünwald (2005) points out:

> If you *overfit*, you think you know more than you really know. If you *underfit*, you do not know much but you know you do not know much. In this sense, *underfitting* is relatively harmless, but *overfitting* is dangerous.

Grünwald continues that a simple model's predictions are relatively reliable indication of the models performance with the future data, while a flexible model's are not. Thus, it is always safer to choose a simpler model; it can be gradually made more flexible, and also more complex, as more evidence is obtained to justify it.

Roberts & Pashler (2000) suggest, not only the fit should be taken into account but also the non-fit. In other words, if a model is seen as a collection of constraints and restrictions present in the data (Rissanen, 1989), in order to be useful, it should limit the number and type of data patterns it can fit well. And the scientist, the designer of the model, should be able to predict which data patterns her model is unable to fit.

Realism

A central question in modeling is how much detail one wants to build into the model. Exact replication of real-world components and processes introduces another kind of complexity that, rather than resulting in overly flexible models, produces highly specific ones that apply to very limited cases. For instance, unlike in many experimental fields, such as cognitive psychology, in LUCC modeling data is often acquired first, and the model is built to reflect the idiosyncrasies of the data. In other words, the theoretical assumptions behind the model are derived from the observations.

However, the complexity arising from the modeled domain itself is not necessarily its cornerstone, when the goal is not to predict future data or to formulate and evaluate general theories, but to understand and highlight the internal workings of a single system. As said above, modeling objectives come in many forms, and not everyone designs models with generalizability in mind. In ecology models are often written for policy evaluation purposes for a specific area or group, in which case sufficient detail is required to convince decision makers of the appropriateness of the model.

A rather impressive example of this kind of model is the Albuquerque's traffic model, a complete replication of the city's street system with households, travelers and vehicles (Casti, 1997). The model can be used to study traffic patterns throughout the city area for 24-hour time periods, and it enables zooming in and tracking of single individual travelers. The model's ultimate goal is to help measure environmental impact, manifested in air-pollution levels, induced by changes in traffic patterns. Similar simulation models have been designed for traffic forecasting in Helsinki (Karasmaa, 2003) and other Finnish and European cities.

Of course, like this example demonstrates, the recent developments in computer technology, expansion of available storage space and memory together with the drastic increase in execution speed, enables implementation and running of this kind of huge and complicated systems. However, even if the exploitation of this power is possible it does not make it warranted, especially if there are no means to analyse and understand the system's behavior; how the model outcome is interpreted and used to inform decision making is a human decision after all. Finally, despite the model's faithful precision with respect to the real world, the question remains: how much confidence should be put in it?

Model Selection Algorithms

There are several model selection algorithms (or methods or criteria, used interchangeably here) that are commonly used and applied, and all of them take the flexibility and fit into account into various degrees. Here I will only cover few of them, which seem to be the most commonly applied, to highlight their basic characteristics.

Most of these algorithms can be considered as maximum likelihood methods, and they can be roughly classified into *penalty-based methods* (Kearns et al., 1997) and *generalization test* methods. The basic difference is that in the former the term for complexity punishment is explicitly represented. *Root mean square deviation (RMSD)*, *Akaike's information criterion (AIC)* (Akaike, 1973), the *Bayesian information criterion (BIC)* (Schwarz, 1978) and the *Minimum description length principle (MDL)* (Rissanen, 1978) belong to the first class and *Cross-validation (CV)*, the *Prequential approach* (Dawid, 1984), and *Bootstrap methods* (Lendasse, Simon, Wertz, & Verleysen, 2005; Zucchini, 2000) into the latter.

RMSD is the simplest one of the criteria, and it ought to be minimized as:

$$RMSD_i = \sqrt{\frac{ER_{sq}(\mathcal{M}_i)}{N - p_i}},$$

where $ER_{sq}(\mathcal{M}_i)$ is the sum of squares' error the model in class i makes with respect to the data, n is the sample size and p_i the number of free parameters in the model class $\mathcal{M}_i$.

The AIC and BIC are, superficially, similar in form even if they have been derived differently. While the BIC was derived in the Bayesian framework, and designed to find the 'true model' among the candidates, AIC was derived from the Kullback-Leibler distance (Kullback & Leibler, 1951). It relies on the assumption that a 'true model' exists, and tries to approximate this 'truth.' The criteria, to be minimized, are defined by the following formulae:

$$\begin{aligned} AIC_i &= -\ln f(D|\hat{\theta}_i(D)) + 2p_i \\ BIC_i &= -\ln f(D|\hat{\theta}_i(D)) + p_i \ln n, \end{aligned}$$

where $f(\cdot)$ is the likelihood function that gives the probability to the data D using the maximum likelihood parameters $\hat{\theta}_i(D)$ (cf. Chapter 1.3), p_i again the number of free parameters in class $\mathcal{M}_i$ and n the size of the data sample. In both equations the term to the right of the plus sign constitutes the complexity penalty.

The *Minimum Description Length* principle (Rissanen, 1978; Barron, Rissanen, & Yu, 1998; Hansen & Yu, 2001), inspired by algorithmic coding theory and later introduced as a method for model selection by Rissanen (1978), is commonly used to infer structure in the data. The *two-part code* or crude version of the MDL principle is the following:

$$MDL_i = L(D|\hat{\theta}_i(D) + L(\hat{\theta}_i(D)),$$

where $L(D|\hat{\theta})$ is the code length function that returns the number of bits needed to describe the data D with help of the model (parameters) $\hat{\theta}$, and $L(\hat{\theta})$ the length of

the encoding (in bits) of the model parameters that minimize the sum. The number of bits required to describe the model (parameters) itself constitute the flexibility term (Grünwald, 2005). The best model according to this principle is the one that minimizes the number of bits required to encode the model itself and the data with help of the model. In other words, a very flexible model that requires a large number of bits to be encoded is selected only if it substantially reduces the number of bits required to encode the data. Some other formulations of the MDL principle, not discussed here, explicitly take the functional form into account (Grünwald, 2000; Myung, 2000; Pitt et al., 2002).

Cross-validation (CV) techniques do not explicitly account for potential sources of complexity, but test a model's generalizability. In CV data are partitioned into K sets $(D_1, \ldots, D_K)$ and each set, in turn is used as validation data, while the other sets are used as calibration data. The criterion to be maximized is formalized as:

$$CV_i = \sum_{k \in K} \sum_{d \in D_k} f_i(d | D \setminus D_k),$$

which gives the model performance for class $\mathcal{M}_i$ with the validation data d using the parameters estimated from the calibration data $D \setminus D_k$. In leave-one-out cross-validation $K = N$. For $K \ll N$, a CV criterion is calculated several times for different partitions and then averaged. The partitioning can be done either temporally, or in the case of a laboratory experiment by participating subject or experimental design (Busemeyer & Wang, 2000).

As a side note, Dawid's (1984) *Prequential (sequential prediction)* approach can be introduced since it is also based on generalization. If the data D^n are a sequence of instances $(d_1, d_2, \ldots, d_n)$, the following entity measures how well $\mathcal{M}_i$ is able to predict the immediate future:

$$Prq_i = \sum_{t=0}^{n} \ln f_i(d_{t+1} | D^t).$$

The entity Prq_i represents the sum of probabilities the model class $\mathcal{M}_i$ gives to the each instance d_{t+1} after seeing the instances $D^t = (d_1 \ldots d_t)$.

Finally, bootstrapping is another method for estimating a model's generalization performance by resampling data with a replacement to obtain the calibration data, and using the rest of the data as validation data. The process is repeated multiple times until the estimate converges (Lendasse et al., 2005; Zucchini, 2000).

Summary

The paradigm shift from the traditional approach of null hypothesis testing to model selection is gaining ground in several disciplines, partly because of the obvious shortcomings of the former methods and the benefits offered by the latter. However, model selection does not safeguard scientists from all the problems for which null hypothesis testing is criticized. Regardless of the data analysis tools, a scientist can still formulate trivial hypotheses or include excess parameters in her models to improve their fit. For model selection purposes, she needs extra care in choosing a plausible set of candidate models. Finally, she must not confuse the best model, chosen by the criterion, with a good model (Stephens et al., 2005).

3

Model Selection Framework

In this chapter I present the conceptual framework within which the model selection criterion, proposed in Chapter 4 is evaluated. I choose the model class proposed by Cioffi-Revilla & Gotts (2003), that is the class of agent-based models of *Territorial Resource Allocation Processes* in two-dimensional space (TRAP^2), and more specifically spatially explicit agent-based learning models of *Land-Use and Land-Cover Change (LUCC)* (Parker et al., 2003).

The choice of domain is inspired by on-going research in individual agricultural land-use decision making based on real land-cover, land ownership and landscape suitability data covering two townships in South-central Indiana (Evans & Kelley, 2004; Hoffman et al., 2002; Laine & Busemeyer, 2004b). The research project integrates computer modeling with other approaches, such as household surveys and institutional analysis. The goal of the modeling component is to compare various decision, learning, and non-learning strategies embedded in an agent-based modeling framework. The above mentioned modeling efforts have raised several well-founded questions: first, how much real-world complexity to include in the model and where to place it; secondly, how to validate the model against the available data; thirdly, what measurements to use when comparing the models; and

finally, which criterion to use when deciding which model is the best one. This dissertation primarily intends to provide an answer to the last two questions, but the second one is addressed as well.

3.1 Objective

Besides providing a platform for the proposed model selection criterion, one of the main functions of the framework is to enable the distinction between the model and what is modeled. In some cases this distinction is obscured by the tendency to introduce the processes and structures of the modeled domain in great detail in the model for the sake of 'realism'. Keeping the complexity coming from the field distinct from the abstract mechanisms of the framework is achieved by introducing domain-specific assumptions as exogenous forces or input data to the model.

3.2 TRAP^2 Assumptions

The following are the conceptual assumptions of the TRAP^2 model class adapted from Cioffi-Revilla & Gotts (2003):

1. The *landscape* is an abstract rectangular area divided into *cells* of equal size, which serve as the decision-making units. Each cell has a fixed neighborhood, and it is identified by the x and y coordinates in the landscape grid.

2. Each cell has various biophysical properties that remain constant over time.

3. The main actors in the model are autonomous *agents*. They have a potentially infinite existence, although they can perish or decide to exit. All agents are of the same type, but their individual characteristics may vary.

4. Agents control a region, called a *parcel*, which is a set of adjacent cells on the two-dimensional landscape. Agents have exclusive access to this region, unless they yield control to a new agent.

5. Agents make resource allocation decisions on their parcel in order to satisfy their goals. Agents have a limited set of available actions, i.e., options to which to allocate their resources. Agent actions change the use of the cells on their parcel.

6. Agents may decide to interact with their neighbors. The neighborhood structure may be defined as based on the physical proximity of the parcels, or some other criterion (e.g., social network or global neighborhood).

7. The points (1.) and (6.) define two levels of neighborhood relations: one that connects the agents and the other that connects the cells on the physical landscape.

8. Decision are made synchronously in discrete (abstract) time-steps.

9. All agents have the same decision strategy. For example they use the same type of reinforcement learning, but each individual adjusts to changes in environment by changing the type of decisions they make. They use the feedback from the past to modify their future decisions.

10. The global environment consists of external conditions that are common to all parcels. These conditions may change over time.

There are two deviations from the original framework proposed Cioffi-Revilla & Gotts worth mentioning. First, property transfer between existing agents is not possible in the current framework. Parcels change ownership only when an agent decides to leave or perishes, and then a new agent is introduced. No parcelization

is implemented either, i.e., the existing parcels are not divided among current or new agents. Secondly, agent actions are assumed successful in the sense that their parcels always produce the maximum yield. The biophysical properties of the land affect the payoff the agents receive from the yield, not the yield itself. Although well worthy of a research agenda of their own, these two aspects are excluded for simplicity.

3.3 Other Assumptions

The current implementation supports only two land-uses or land-covers. First of all, this restriction makes the analysis of model outcomes easier. The binary land-use classification can be interpreted as 'use X' vs. 'non-X use', for instance urban vs. rural, or forest vs. non-forest. Even if a variety of numbers of land-uses/covers is present in the literature, for instance two in Parker & Meretsky (2004), three in Jenerette & Wu (2001), four in Deadman *et al.* (2004) and Schneider & Pontius (2001), five in Irwin & Bockstael (2002), and as many as eleven in Ligtenberg *et al.* (2001), in pursuit of a general framework there is no real justification for any other particular number of uses besides two.

In general, the framework is kept simple to highlight the model selection procedure, but flexible enough to enable testing the basic assumptions of the following factors and their effect on the performance of the proposed selection criterion:

- Agent Learning and decision making
- Spatial metrics (cf. Chapter 1.3)
- Exogenous factors

- Biophysical processes

There are several straightforward ways in which the framework can be extended so that it is able to model a wider variety of processes operating in land-use and land-cover change. These include:

- Introducing endogenous forces that change as a result of agent actions. For instance, prices could be made to fluctuate to reflect supply and demand, and land suitability could be contingent of agents' land use.
- Making biophysical processes have direct effect on agents' decisions and their outcomes, instead of indirectly through the revenues.
- Adding other types of decision-making agents, e.g., institutions.
- Making it possible for the agents to change the type of their learning and decision-making algorithm.
- Introducing either endogenous or exogenous land markets, for instance by parcelization, or allowing the agents to acquire land from their neighbors.

3.4 Architecture

The main components of the framework equal the environment and the agents. The environment consists of the physical landscape and various exogenous variables. Agents represent autonomous decision makers, such as individuals, families, households or other groups that bear the consequences of their decisions.

The framework architecture combines the *object oriented view* and the *field view* of landscape change modeling (Brown et al., 2005). The discrete entities in the

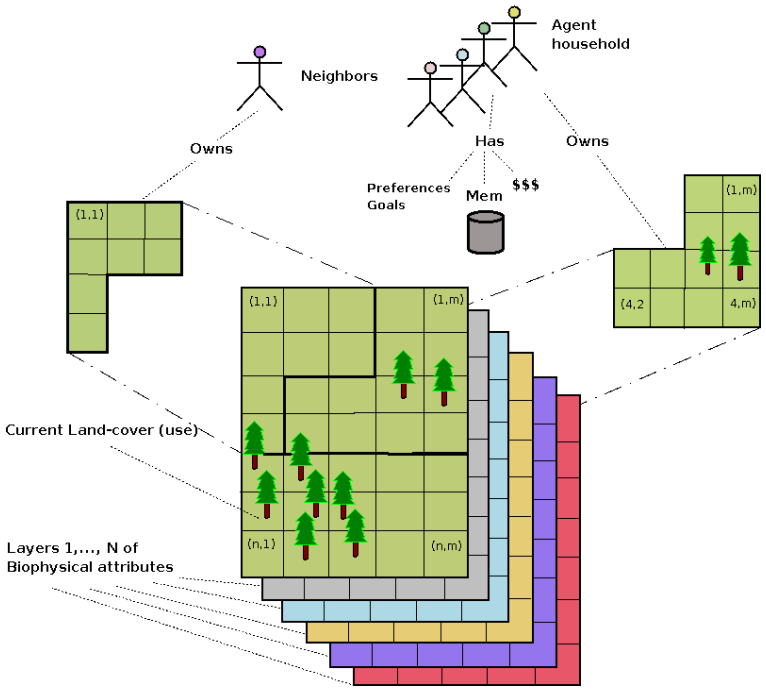


Figure 3.1: Main components of the TRAPP^2 modeling framework.

model are the agents and their attributes. The landscape is represented by spatially distributed geographic variables, such as current land-cover and suitability of land-uses. Agents are spatially linked to the landscape through their parcel. The parcel objects and their attributes are mapped to the spatial landscape by their location attributes. The architecture is informally presented in Figure 3.1.

The land-use decision-making agents are assumed to be social actors that have different personal and demographic characteristics, for instance subjective preferences, risk attitude, age, social status, occupation, wealth, skills and knowledge, and goals. This heterogeneity is captured in three kinds of attributes, summarized in Table 3.1: *input parameters*, *individual (free) parameters* and *free parameters associated to learning algorithms*. While the first type of parameters are controlled by the

Parameter type	Parameter name	Value	Range
Input parameters	Initial wealth	w	$w \in \mathcal{N}(\mu, \sigma^2), \mu, \sigma \in \mathbb{N}$
	Household Size	h	$h \in \mathcal{N}(\mu, \sigma^2), \mu, \sigma \in \mathbb{N}$
	Size of social network	n	$n \in \mathcal{N}(\mu, \sigma^2), \mu, \sigma \in \mathbb{N}$
	Parcel cover	c	$c \in \{0, 1\}$
	Parcel suitability	s	$s \in \{0, 1 \ldots 100\}$, $s \in \mathbb{N}$
Individual parameters	Subjective preferences	α	$\alpha \in \mathbb{R}$
	Other free parameters	β	$\beta \in \mathbb{R}$
Learner parameters	Learning & decay rate, weight,etc.	(cf. section 3.5	(cf. section 3.5)

Table 3.1: Attributes associated with the agents. Parameters associated with the learning strategies are introduced together with the strategies.

experimenter, the latter two parameter types are estimated from the data.

Wealth controls the agent's chances of survival, the household size determines the extent of labor the agent has available for different activities, and the size of the social network constrains the number of other agents which with it may communicate or which it may use as sources of information. The parcel is a two-dimensional representation of the biophysical attributes related to the region of the landscape controlled by the agent, and is also the domain of the agent's decision making.

The *individual* parameters are known as subjective preferences (α) for alternative actions, and any number of other parameters the user would like to define (β). The parameters associated with learning algorithms, together with the additional attributes required by them, are described together with the algorithms.

3.5 Learning and Decision Making

Decision Algorithm

At each decision round agents observe the state of their land, and make a decision about its use in the next round. They make the decision for each cell separately; they either decide to keep the old use or select another use from the given alternatives. After making the decision for each cell[1], they observe the payoff earned from the decision.

The payoff structure is adopted from the design developed for the laboratory experiments conducted by Evans *et al.* (in press). It combines both monetary and non-pecuniary returns, and depends on the number and location of cells allocated to the activities, and unit returns from these activities. The location factor includes the suitability and externality effects. At each decision round the agent i's wealth is modified by the total payoff received from all the activities by the following:

$$\Delta\omega_i = \sum_j \sum_k [I(j,k)\alpha_k(s_{jk}w_s + e_{jk}w_e + p_k)] - C(i),$$

where j enumerates the cells of agent's parcel and k possible land-uses, and $I(j,k)$ indicates whether the cell j is in use k. α_k constitutes the pecuniary return, and quantifies the agent i's general preference for the land-use k.

s_{jk} is the cell j's suitability for the use k (cf. Chapter 1.3). e_{jk} the externality effect (as explained in Chapter 1.3) of use k on cell j, calculated as $e_{jk} = \sum_i \delta(i,j)$, where i enumerates the immediate neighbors of the cell j; $\delta(i,j) = 1$, if cells i and j are in the same use, 0 otherwise.

The suitability and externality weights, w_s and w_e respectively, are included in the payoff in order to test the relative impact different components have on the

[1]The land use is 'automatically' realized and the revenues calculated.

payoff structure. For instance, by varying the suitability weight one can determine the level at which the agent starts to pay more attention to suitability as opposed to the monetary return p_k, which is unit price earned from the use k.

The cost factor $C(i)$ is calculated as follows:

$$C(i) = \frac{\beta\delta_i}{\eta_i},$$

where β is interpreted as a cost of change, δ is the number of cells changed by the agent, and η is the agent's household size.

Learning Algorithms

Even if there are some general purpose learning algorithms, not many of them directly apply to specific real-world domains, such as land-use change. The choice of the learning and decision algorithms in the current study was driven by the desire, first, to keep them very general and abstract without incorporating *ad hoc* assumptions or details tailored to the task beyond necessity, and secondly, to have representative set of algorithms that are (supposedly) able to exhibit different — and interesting — patterns of behavior. These algorithms are described next.

Greedy agent prefers continuity in the use of neighboring cells. For each cell it selects a use from the uses of the neighboring cell and the cell itself that generated the highest payoff in the previous round. There are no free parameters.

Q-learning agent (Kaelbling & Littman, 1996; Watkins & Dayan, 1992) follows a form of reinforcement learning. It maintains and updates so called *Q values* associated to state-action pairs (s, a), where s is current land-use and the action a the new use. After performing the action a in the state s, which results

in a new state s', it updates $Q(s,a)$ by the following rule:

$$Q(s,a) = Q(s,a) + \phi[P_{s,a} + \gamma \max_b Q(s',b) - Q(s,a)],$$

where P is the payoff received from the action a in the state s, ϕ is the learning rate and γ is the future discount factor for the Q value of the best action b in the next state s'. A Q-learner selects the action with the highest Q-value in the current state. There are two free parameters, ϕ and γ. For more information, see Watkins (1992).

Social experience-weighted attraction (sEWA) is a version of the EWA algorithm proposed by Camerer and Ho (1999) for normal-form games. The EWA learner i maintains attraction values $A_i^j(t)$ for actions s^j, and updates them using the following equation:

$$A_i^j(t) = \frac{\phi N(t-1) A_i^j(t-1) + [\delta + (1-\delta) I(s^j, s_i(t))] u_i(s^j, s_{-i}(t))}{N(t)},$$

where $N(t) = \gamma N(t-1) + 1$ is the experience weight (γ is the discount factor), ϕ is the discount rate for previous attractions and δ is the weight for the payoffs of unchosen actions. $s_i(t)$ is the action chosen by the agent at time the t, while $s_{-i}(t)$ is the actions chosen by the agent's neighbors at the time t. $I(x,y)$ is an indicator function returning 1 if $x = y$, and 0 otherwise. The crucial feature of EWA learning is that it also reinforces unchosen strategies by discounting the payoffs that they would have generated if chosen. The utility for agent i from the action j, using its payoff P_i and the average payoff received by others $\bar{P}_{-i}$, is calculated as follows:

$$u_j = \begin{cases} \rho \bar{P}_{-i} + (1-\rho) P_i, & \text{if } P_i \geq \bar{P}_{-i} \\ \chi \bar{P}_{-i} + (1-\chi) P_i, & \text{otherwise .} \end{cases}$$

An agent following EWA selects the action with the highest attraction value.

ϕ, δ, γ, ρ and χ are the free parameters. For more details see Camerer & Ho (1999).

Individual EWA (iEWA) is a version of an experience-weighted attraction learning model that only considers its personal payoff when updating attraction values. There are three free parameters: ϕ, δ, and γ.

Two *Null* models are used as baseline models in the comparison, a *model of pure persistence* and a *random model*:

Model of pure persistence does not make any changes in the environment.

Random decision model chooses the use of a cell randomly from a uniform distribution of the available alternatives.

3.6 Spatial Metrics and Error Functions

The question remaining is how to measure the model's fit or lack of fit. Usually, the performance of land-use and land-cover change models is assessed by calculating a set of spatial metrics from a series of landscapes generated by the models and compared to the same metrics calculated from real landscape data. Deviation from the data is often measured by the sum of squared error over time but any other error functions can be used as well.

In this dissertation it is hypothesized that the choice of metrics used to quantify the fit is equally important as the choice of the model selection criterion, since the criterion's performance will depend on the metrics. It is assumed that different learning and decision-making strategies exhibit different behavioral regularities and produce varying land-use and land-cover patterns, and the spatial metrics

differ in their capacity of identifying these patterns (Parker & Meretsky, 2004). Some of the metrics measure spatial composition, and some of them measure configuration, and they may do this with or without regard for their exact locations. The ultimate goal is to choose a collection of metrics that together are able to reveal diverse patterns in the model outcomes with a relatively light computational burden.

The four spatial metrics chosen for this study are described next. All of them can be calculated both on the landscape level and the individual decision maker's parcel level. A couple of minor restrictions are in order, though; first, in the two-land-use scheme the metrics are calculated only for one land-use (land-cover) at a time; and secondly, because of the general nature of the framework, the metrics are calculated in abstract units of the landscape grid cells instead of the real units of length or area, say square meters or acres.

The metrics are a function of either one or two landscapes, and defined as follows:

Mean absolute difference (m.a.d.) is a cell by cell absolute difference between two landscapes or parcels L and L' divided by the total area:

$$m.a.d.(L, L') = \frac{\sum_{i \in L, L'} |c_i - c'_i|}{TA}$$

where c_i is the cover of the cell i on the landscape L and c'_i is the cover of the cell i on the landscape L'. TA is the total area of the landscape or the parcel.

Composition (c) measures the percentage of the landscape or parcel L that is in the land-use c. It is calculated by the following:

$$c_c(L) = \frac{N_c(L)}{TA},$$

where N_c is the number of cells in the use c, and TA is the total number of cells in the landscape or the parcel L.

Edge density (e.d.) measures the relative complexity of patterns in which the land-use c occurs on the landscape or the parcel L.

$$e.d._c(L) = \frac{E_c(L)}{N_c(L)},$$

where E_c is the length of the border between the use c and other uses. This is the number of cells with a different use neighboring c. N_c is as above.

Mean patch size (m.p.s.) is the total area in patches divided by the number of patches, where a patch is a continuous area of a single land-use c completely surrounded by other land-uses.

$$m.p.s._c(L) = \frac{TAP_c(L)}{NP(L)},$$

where TAP_c is the number of cells in patches of use c, and NP is the number of patches.

Given data $x^T \in X^T$ over time period T, the error model $H \in \mathcal{M}_i$ makes with respect to the data is defined by either of the two error functions: one for the metric $diff(x^T, H^T)$ and one for the remaining three metrics.

$$\begin{aligned} ER_{sum}(x^T, H^T) &= \sum_{t \in T} m.a.d.(x^T, H^T) \\ ER_{sq}(x^T, H^T) &= \sum_{t \in T} (y(x^t) - y(H^t))^2, \end{aligned}$$

where y is in $\{c_c, e.d._c, m.p.s._c\}$ and H^t is the landscape generated by the model H at the time t.

3.7 Summary

The four spatial metrics and the error functions are integral parts in the model selection process as will be evidenced in Chapters 4 and 5.3. Therefore, care needs to be taken when choosing which metrics are to be used to compare models with models or models with data; different metrics extract qualitatively and quantitatively different patterns in the landscapes and some of them are easier to fit than others.

In the next Chapter the model selection criterion based on the algorithmic coding theory is introduced, and the discussion on the spatial metrics and model selection is resumed in Chapter 5.3.

4

Model Selection Based on the Minimum Description Length Principle

4.1 Background

Which model selection algorithm to use for agent-based models of land-use and land-cover change? Several issues make this a compelling question.

In selecting a good model we need to consider at lest two questions: how well the model performs relative to the data we have, i.e., the model's goodness of fit, and how well it performs relative to all other data, i.e., the model's flexibility, or in other words, its propensity to overfit. Most commonly used model selection methods balance goodness of fit and flexibility by taking into account factors such as the number of free parameters, number of data samples, number of data points fitted, or using some geometric measures to quantify the complexity of the functional form (Pitt et al., 2002).

These variables are relatively easy to identify and measure for simple polynomial or probabilistic model classes. However, a typical LUCC model is as much unlike a physical law, such as law of gravity or speed of light, whose accuracy can be easily verified by repeated measurements, as it is unlike a probability distribution to which penalized ML methods, such as AIC (Akaike, 1973) or BIC (Schwarz, 1978), can be applied. A LUCC model can be best characterized as a *Complex Adaptive System (CAS)*, a system consisting of multiple autonomous components and processes that interact at multiple spatial levels and temporal scales. Because of these interactions, the behavior of a CAS is not always predictable and the errors a model generates may not be a deterministic function of its parameter values. Furthermore, any qualitative or quantitative judgment of structural factors assumed underlying the model's flexibility is subjective and susceptible to various biases due to ontological and implementational considerations.

The data available for validation of LUCC models are not plenty and often not random samples. As noted above, with small sample sizes, the effect of the functional form overshadows that of the free parameters. Classical hypothesis testing does not take the functional form into account. Therefore, if it used for model selection, the method best applies to nested models, i.e., to models one of which is a special case of another. Consequently, the selection amounts to choosing a value for some parameter θ, i.e., testing statistical significance of the null hypothesis $h_0 : \theta = 0$ against the alternative $h_a : \theta \neq 0$. Both AIC and BIC restrict their consideration of model structure to the number of free parameters. The RMSD method (Myung, 2000) cannot handle cases with small sample sizes and potentially large number of parameters, since it may yield negative square roots. Furthermore, unlike for the other methods discussed in the current section, there is no statistical justification for RMSD.

With the exception of some simple cases, it could be dangerous to assume that the data is generated by a certain 'true' model class, and base further inferences about the state of the world on this fact. Simple cases that lend themselves to this assumption are, for instance, the model of the average height among six-graders in the Midwestern United States, or the preliminary polling results in the second round of the Finland's presidential elections in 2006. In the former case, it is relatively safe to assume that the average height (and the standard deviation) exists even if it is unknown, and in the latter case, that the distribution of votes exists although it changes all the time (as more data is gathered in exit polls). In any more complicated case, such as a mental model of a decision maker, the 'true' model is almost impossible to recognize and verify.

Moreover, model parameters and functions are not inherent properties of the system we want to model, but theoretical constructs used by us to describe the system. We impose these properties on the system. There are always multiple models that equally well replicate the behavior of a system. Any of them can be refined indefinitely to resemble real system more and more in detail until it closely matches the 'truth.' Again, there is no way to verify that such a 'true model' exists, and consequently the task of estimating something that does not exist becomes quite impossible.

The above considerations make it relatively clear that most of the existing methods, such as AIC, BIC or RMSD are inapplicable for our purposes. Cross-validation (CV) and bootstrap methods make no assumptions of the model itself, but they only use data. The effects of free parameters and the functional form are implicitly manifested in the selection process. However, in order for these techniques to be reliable, one needs to have quite a bit of data: first of all, to be able to partition the data into the calibration and validation sets, and secondly, in order to

repeat the test multiple times with different partitionings. Real world data is not always readily available in quantities that warrant usage of CV or bootstrap methods. Furthermore, both these methods still use the same data for calibration and validation; even if the data is partitioned into separate calibration and validation sets, the process is usually repeated with different partitionings. Eventually, when the process finishes, all data has been used both in validation and calibration.

Where are we now? Penalized maximum likelihood and Bayesian methods are inapplicable because of the nature of the model class and the assumptions we make about the real state of the world, and the scarcity of data renders cross-validation infeasible. How about using the *Minimum Description Length (MDL)* principle? Despite having some neat theoretical properties, for many practically interesting model classes the MDL criterion and especially its refined formulation, *Normalized Maximum Likelihood (NML)* distribution (Rissanen, 1999), cannot be calculated. By this I mean that the worst-case optimal universal code, on which the 'ideal' NML principle is based, does not exist, which undermines the usage of the principle. Furthermore, the component, called *parametric complexity* of the NML criterion is usually impossible to compute analytically, because it may be infinite. For details, see for instance Rissanen (1999), Myung *et al.* (2006) and de Rooij & Grünwald (2006). Several alternative solutions have been suggested. These rare practical successes have mainly been demonstrated with toy models on artificial domains or with relatively simple probabilistic models on real data, for instance with psychological data (de Rooij & Grünwald, 2006; Myung et al., 2006; Pitt et al., 2002).

As discussed above, a CAS does not lend itself easily to this sort of theoretical analysis on which recent derivations of the MDL principle are based. Fortunately, we have not run out of options, since the MDL principle has a convenient and

practical interpretation in the context of communication. The main objective of the dissertation is to develop a method that is applicable in practice — if not ideal, at least reliable. This is the direction I pursue next.

4.2 Minimum Description Length Principle and Model Selection

The *Minimum Description Length (MDL)* principle is a general method of inductive inference. The principle is based on the idea that regularities in the data can be used to compress it (Rissanen, 1978; Grünwald, 1998). Applied to model selection, MDL suggests that the best model to explain the data is one that compresses the data most efficiently. In other words, the model using the least number of bits in describing the data most likely captures its underlying regularities[1], which can be used to gain insight to the underlying system or processes that generated the data. This principle of parsimony is not new in science; it has been attributed to Medieval English philosopher William of Ockham and is these days commonly called *Ockham's razor* and frequently applied in the modeling literature.

The MDL principle only uses data in selecting among the candidate models. After all, the goal is to find a good model for the data without relying on the assumption that the data was generated by some model. According to the MDL philosophy, the model's task is to describe the properties of the data, not to pretend to be the system that generated the data. Neither does MDL define a specific algorithm for selecting among models; it lays out a general principle of using compression to detect interesting patterns in the data. There may be several equally

[1]What do I mean by regularity? I take it to purport to any property of data that enables it to be described or encoded in fewer symbols than is required to list it literally (Grünwald, 2005). The basic idea behind the compression is to code frequently occurring symbols with shorter codewords.

optimal ways of implementing the principle.

Notation

For the time being, let us assume that the data samples are observations coming from all possible data samples of size n: $(x_1, \ldots, x_n) \in \mathcal{X}^n$, ($x_1, \ldots, x_n$) is often abbreviated x^n. Each x_j comes from a finite, fixed and countable alphabet $\mathbf{A}$. If the data size is irrelevant or clear from the context I denote a data sample with D instead of x^n, and the set of all possible data samples with $\mathcal{X}$ instead of $\mathcal{X}^n$.

A *code* for $\mathcal{X}$ is a mapping from $\mathcal{X}$ to the set of binary strings $\bigcup_{m>1}\{0,1\}^m$. I only consider binary *coding alphabets*, and consequently all the logarithms used are in base two. The encoding of data D, also called a *codeword* for D, is denoted $C(D)$, and the length of encoding of D is denoted $L_C(D)$, where L_C is the code length function $L_C : \mathcal{X} \to \mathbb{N}$, and C is a code or a coding method. I also assume a prefix (also called prefix free) coding method throughout the discussion, so no delimiter symbols are required. In prefix code no codeword for an observation is a prefix of a codeword for another observation, which assures unique and instantaneous decodability.

Each code is associated to a candidate model class $\mathcal{M}_i$. Models belonging to the class $\mathcal{M}_i$ are denoted H[2].

Preliminaries of Principle

Let us first consider a simple example.

[2]H stands for (point) hypothesis as these models are often called in the statistical literature.

Example 4.1. We have two data sequences s_1 and s_2 of length $n = 30$, consisting of symbols from the alphabet $\mathbf{A} = \{a, b\}$ that we want to describe as efficiently as possible. The sequences are:

$$\begin{aligned} s_1 &= abaaaaabaaabaaabaaaaaaabaaaaab \\ s_2 &= bbaabbaabbbaabbbbbbbabbaabaabb \end{aligned}$$

Seemingly similar, both of these sequences exhibit some kind of regularity, but for one of them it is totally incidental. For the sake of illustration, let us assume that we know the generating processes behind these sequences: s_1 consists of 'ab's followed by an even number of a's, and s_2 is produced by 30 tosses of a fair coin so that an 'a' marks a head and a 'b' marks a tail, or vice versa.

We can construct a simple code so that $C_{simple}(a) = 0$ and $C_{simple}(b) = 1$, but unfortunately, with this code we do not achieve any compression. Alternatively, we can enumerate all 2^{30} sequences of length 30 consisting of symbols from A, and encode each sequence with the number corresponding its rank in the enumeration. This code allows significant compression for some sequences, but produces relatively long codes for most of them.

Finally, we can use some insight and utilize the repeating patterns in s_1 to express the sequence in significantly fewer bits than n. For instance, if we characterize the regularity in s_1 as "'ab' followed by m times 'aa', where $m > 0$ can be even or odd", we can construct code C_1 so that $C_1(ab) = 1$, $C_1(aa) = 0$, and achieve the code length $L_{C_1}(s_1) = 15$. Thus, C_1 has a compression rate of $.5$. Or alternatively, we can exploit some other regularity and design code C_2 so that $C_2(abaa) = 0$, $C_2(aa) = 10$, and $C_2(ab) = 11$, which also achieves the code length $L_{C_2}(s_1) = 15$. The binary code tree for this code is shown in the Figure 4.1. These two codes achieve a unique encoding for the sequence s_1; however, several other encodings

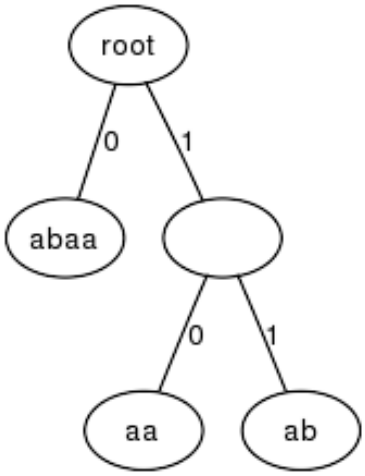


Figure 4.1: Binary code tree for the alphabet $A = \{a, b\}$ and code $C_2(abaa) = 0$, $C_2(aa) = 10$, $C_2(ab) = 11$.

are also possible, but it is unlikely they can minimize the description length much more than C_1 and C_2 do.

Regarding s_2, several seemingly regular patterns, such as repeating sequences 'aa' and 'bb', can be detected *post hoc*, but given the random underlying process, their occurrence is not predictable. Consequently, there is no effective way of describing s_2 in fewer symbols than is required to state the sequence literally. To see why this is the case, consider the number of different substrings that s_2 can be divided into. The number of bits required to describe them increases by one for every 2^i additional strings, where i is the length of the currently longest codeword. For instance, even if we design code C_3 so that the longest repeating substrings, such as the four instances of 'bbaa', obtain short codewords, i.e., we use $C_3(bbaa) = 0$, the remaining symbols or substrings necessarily have code length of at least two bits, and consequently we are not able to compress the sequence at all.

It is worth noticing that these examples only define a partial code. However,

what we really want is a coding system with which we can describe any sequence of size n.

Two-part Code

In the process of selecting among agent-based land-use and land-cover change models we only have data, candidate model classes and the errors that models from these classes make on the data. The goal is to select the class containing a model that best explains the data. Particularly, no assumptions are made on the existence of a 'true' model, nor any subjective, *a priori* assessment of the model's structure, independent of its performance, is made. This enables a fair and unbiased selection between candidate models. As opposed to Quadrat-Ullah's view (2005), who suggests that structural validation is sufficient to be convinced about the model's adequacy to the task, I argue that both in validation and selection the behavioral accuracy should be under strict scrutiny. Otherwise the model's adequacy may be judged purely on its face value. Likewise, the measure of flexibility should be based on the data and model's performance with it, not any *a priori* knowledge about the model structure. After all, the model's flexibility is contingent on the data.

Let us return to our simple example. The encoding schemes for sequences s_1 and s_2 described above may seem quite *ad hoc* and trivial; coming up with a sufficiently efficient code for a data sample once it is observed is relatively easy. There are (at least) two interrelated problems with this approach. First, the efficient coding depends on the data. The code designed for one data sample may prove extremely uneconomical when used to encode another sample. Second, the MDL principle is often presented in the context of a communication channel between two individuals, a sender A and a receiver B (Grünwald, 2005). A sends B a data

sample that she has encoded using some coding method. In order for B to be able to interpret the encoded message A sent her, she needs to know which code was used to encode it. However, the best code A supposedly used depends on the individual sample. So, B is faced with a dilemma; to interpret the data she needs the code to decode it, but in order to decide which code to use, she needs to know the data.

A and B could potentially decide beforehand which code to use, but while any particular code may lead to efficient encodings for few samples, it may produce disastrously bad encodings for many of them. Another alternative is that together with the data, A transmits information about the code that was used to describe the data.

How does this relate to model selection? Remember, we are interested in selecting a model class that best explains the data. The less bits a model needs to describe the data, the more interesting patterns it can extract from it. On the other hand, the more intricate the class the model belongs to, the more bits are required to describe it. The model selection objective is then to balance these two factors, so that the extra intricacy or flexibility is justified only if it allows for substantially shorter description of the data, i.e., a better detection of regularities.

The *two-part code*, also called a *crude version* of the MDL principle, selects the model class $\mathcal{M}_i$ that trades off the flexibility of the class to the superior fit of the best-fitting model H in it (Grünwald, 1998, 2005):

$$L(D|\mathcal{M}_i) = \min_{H \in \mathcal{M}_i} L(D|H, \mathcal{M}_i) + L(H|\mathcal{M}_i),$$

where $L(H|\mathcal{M}_i)$ is the length of the description of the model $H \in \mathcal{M}_i$, and $L(D|H, \mathcal{M}_i)$ is the length of the description of the data sample D using the best-fitting model H. In the case of simple polynomial model classes, the number of bits required to encode the model is directly related to the number of free parameters, which in turn

defines the complexity of the model class (Grünwald, 2005), and the description of the data amounts to encoding of the errors the model makes when describing the data.

Two-part code for LUCC Models

The next step is to design a two-part code for the class of land-use and land-cover change models. For model selection purposes in general we do not need the actual encodings but can work with code length functions (Grünwald, 2005). However, as noted before, since a LUCC model class cannot easily be translated into a probability distribution in the traditional (frequentist) sense, we need to adopt a different approach. Not only does the design of a code make the abstract notion of the MDL principle more concrete, but it gives us the code lengths which can be translated into probabilities. The probabilities can then be used in model selection as we will see later.

Let us turn to another example.

Example 4.2. Here I adopt the communication interpretation of the MDL principle again. I also assume that both A and B agree on the message structure and the candidate model classes — a set $\mathcal{M}$ of discrete classes $\mathcal{M}_i$ that can be either nested or totally unrelated, i.e., have distinct functional forms.

If the number of candidate model classes is $m = |\mathcal{M}|$, and the number of free parameters in each $\mathcal{M}_i$ is p_i, the description of the model can be constructed in the following manner.

- First we need to identify the model class that is used the encode the data. For this we use a uniform code with $\lceil \log m \rceil$ bits. Note that we do not need

to communicate the number of bits required to encode the model class, since both the sender and receiver have agreed on the set of candidates beforehand. Neither do we need to transmit the number of parameters since it depends on the class.

- The second step in describing the model is to transmit the parameter values. Since we are using binary encoding, we first need a method of encoding numbers in a prefix free manner; note that binary numbers are not prefix free. I assume that the parameter values are numbers $0 < k \leq n, k \in \mathbb{R}$. First, in order to describe the values themselves the sender needs to decide how many bits to use to encode each value. She may choose a finite precision $d = \lceil \log n \rceil$, and discretize the parameter values so that the description of each 2^d parameter value takes d bits to encode. In order to communicate the precision in a prefix free manner, the sender uses unary code; she first transmits $d-1$ ones followed by a zero. She then encodes p parameter values in some non-prefix binary representation with pd bits.

  Alternatively, the sender may use a non-fixed precision, and use different number of bits for each parameter value. In this scheme she first uses unary code to encode the precision $d_j = max(\lceil \log k_j \rceil, 1)$ of jth parameter to inform the receiver how many bits the binary representation of parameter value k_j requires. So, for each parameter the sender first transmits $d_j - 1$ ones followed by zero, and then uses d_j bits to transmit the parameter value k_j. For example, in order to transmit the parameter value 5 she needs $\lceil \log 5 \rceil = 3$ bits for its binary representation 101, and 3 bits to encode the length of the binary representation in unary. The resulting code will be $C(5) = 110101$, and its length $L_C(5) = 6$.

  The sender may use a more advanced code C' (Grünwald, 1998), and first

transmit the length of the binary representation of d_j in unary manner, then send d_j using its the non-prefix binary representation, and finally send the parameter value k_j as above. Using the same example as above, the binary code for parameter 5 remains the same, but only two bits are needed to encode its length in binary (i.e., $3_{dec} = 11_{bin}$). Additional two bits are required to encode the length of this length in unary. The resulting code for the parameter 5 will then be 1011101. While the more advanced code uses more bits than the simpler code C to describe values as small as in this example, as soon as the required precision d_j exceeds 8 bits, the advanced code becomes more efficient.

If the advanced code for precision is used the description length for the model H in class $\mathcal{M}_i$ becomes:

$$L(H|\mathcal{M}_i) = \lceil \log m \rceil + \sum_{j=1}^{p_i} (2\lceil \log d_j \rceil + d_j).$$

where d_j is the precision used to encode the parameter value k_j. The first term gives the length of the description of the model class, and the second term, for each parameter value — the number of which depends on the class — combines the length of the code for the precision and the precision itself, followed by the parameter value described with this precision.

Once the model has been described in the prefix free manner, we need to encode the data. In order to accomplish that, it is sufficient to list the errors. In order to communicate the errors in the land-cover the sender first encodes the number of time points and then for each time point lists the number of errors followed by the locations of errors[3]. I utilize the same idea as above for parameter values to transmit the precision of error locations.

[3]The description of error locations suffices since the errors in binary landscapes are either zero or one. This is equivalent to using the spatial metric 'mean absolute difference' as the error measure.

- If the number of time points is denoted T, the unary code for encoding the length of the description length of T's takes $\log\log T$ bits and the description of the length itself takes another $\log\log T$ bits. Finally T can be described in $\log T$ bits.

- A similar logic can be used to describe the number and locations of errors. Let us assume m_t is the number of locations that are incorrect at time t. The number of bits required to encode the length of the description length of m_t at each time point t is $\log\log m_t$. The description of the length then takes another $\log\log m_t$ bits, and finally the encoding of m_t takes $\log m_t$ bits. At this point the receiver knows how many errors there are and how many bits are used to encode their locations, so she can decode the rest of the message.

The description length for the data D will be:

$$L(D|H,\mathcal{M}_i) = 2\lceil\log\log T\rceil + \lceil\log T\rceil + \sum_{t=1}^{T}(2\lceil\log\log m_t\rceil + m_t\lceil\log m_t\rceil).$$

The total description length of the data D described with the help of the model H in class $\mathcal{M}_i$ is:

$$L(D|\mathcal{M}_i) = L(D|H,\mathcal{M}_i) + L(H|\mathcal{M}_i).$$

Albeit not necessarily optimal in the sense that it gives the absolutely shortest description lengths, this coding method rewards models that make small errors and substantially punishes models that have more free parameters. In other words, the error range of a well fitting models is small, since the errors are small. Therefore, fewer bits are required to encode each individual error. Likewise, a model with a large parameter range requires more bits to encode the parameter values.

Summary

Even if we let the parameter and error precision be flexible, the just presented coding scheme does not say much about the model's absolute flexibility. Judging by the number of parameters is not sufficient for several reasons: first, the current coding method gives all parameters an equal weight which may not be realistic — some parameters may have more radical influence in determining the model's scope than the others; secondly, the major factor behind the potential flexibility may lie somewhere else than within the number of free parameters; thirdly, the model itself does not define how well it fits the data, but an error function is required; and finally, we are not interested in maximum likelihood parameter values after all, i.e., the best fitting model, but we would like to find a well fitting model in a class that is not overly flexible. However, the flexibility is unknown, since we do not necessarily know how the model performs with other data samples. Therefore, we need to improve this code. In the next section I propose an alternative model selection criterion based on normalized minimum errors, and present how it relates to the refined version of the MDL principle, namely the *Normalized Maximum Likelihood (NML)* principle (Rissanen, 1999).

4.3 Enhanced Code for LUCC Models

The fit of a probabilistic model class is measured in the probability it gives to data, whereas the fit of a non-probabilistic class is quantified by the error it makes on data, so that the higher the probability or smaller the error, the better the fit. Since LUCC models are non-probabilistic, their goodness-of-fit is based on errors.

For any finite or countably infinite set $\mathcal{X}$, $\sum_{x \in \mathcal{X}} P(x) = 1$, i.e., the probabilities sum up to one. This implies that we cannot assign very high probabilities to many

x. However, the errors do not have this property of summing up to constant; they can sum up to any arbitrarily small or large quantity. Next I will present how errors can be used in model selection in a principled way.

Normalized Minimum Error Criterion

How do we use errors so that the trade-off between fit and flexibility is adequately treated? If we want to explain the observed data sample x^n with the help of a model class $\mathcal{M}_i$, ideally we want $\mathcal{M}_i$

1. to contain a model H_{x^n} that makes a small error on x^n, and
2. to contain models H_{y^n} that do not make small errors on most y^n belonging to a set of all possible data samples $\mathcal{X}^n$.

This can be achieved by minimizing the ratio between the error the best fitting model in class $\mathcal{M}_i$ makes on sample x^n and the total error it makes on all samples in $\mathcal{X}^n$ using the respective error minimizing parameters. I call this ratio *Normalized Minimum Error (NME)*:

$$NME(x^n, \mathcal{M}_i) = \frac{ER(x^n|\hat{\theta}(x^n, \mathcal{M}_i)}{\sum_{y^n \in \mathcal{X}^n} ER(y^n|\hat{\theta}(y^n, \mathcal{M}_i))}, \tag{4.1}$$

where $ER(\cdot)$ is the error model class $\mathcal{M}_i$ makes on x^n using the parameter values $\hat{\theta}(x^n)$ that minimize the error, and y^n are 'all possible data samples'. By normalizing each error this way we obtain a relative measure for fit and flexibility, which we can use as a model selection criterion.

The MDL principle is a general method of doing inductive inference, and the NME criterion is one way of implementing it. In the next section I will relate the

criterion to another interpretation of the principle, namely to *Normalized Maximum Likelihood (NML)* distribution on $x^n \in \mathcal{X}^n$ (Rissanen, 1999):

$$NML(x^n, \mathcal{M}_i) = \frac{P(x^n|\hat{\theta}_i(x^n))}{\sum_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}_i(y^n))}, \tag{4.2}$$

where $P(\cdot)$ is the probability that model class $\mathcal{M}_i$ gives to x^n using the parameter values $\hat{\theta}_i(x^n)$ that maximize such probability, and y^n are as above.

The NML criterion selects a model class $\mathcal{M}_i$ whose *universal model* H, not necessarily in $\mathcal{M}_i$, minimizes the worst case regret; regret of model H with respect to class $\mathcal{M}_i$ is the extra number of bits that are required to describe the data sample x^n using H instead of using x^n's maximum likelihood model in $\mathcal{M}_i$. H is called a universal model, since it tries to mimic all models in the class $\mathcal{M}_i$. It has been proved (Rissanen, 1999) that the NML criterion defines a unique model that minimizes the maximum regret.

The term in the denominator is the most crucial aspect in both criteria, since it accounts for their ability to penalize for excess flexibility (see Chapter 6 for more discussion on this issue). It is important to realize that the denominator goes over 'all **possible** data', not just over 'all available data' or 'all observed data'. In theory, this means that the term contains errors made on (in case of the NME) or probabilities given to (in case of the NML) all possible permutations of data of certain size. If a model class contains models that either give high probabilities to or produce small errors on many such permutations, the model class is considered overly flexible. Neither criterion penalizes model classes for giving few data samples high probabilities or small errors, i.e., they do not impose a penalty on models that generalize well.

Associating Errors to Code Lengths and Probabilities

Several different model selection methods, such as AIC, BIC, or NML apply to probabilistic model classes. However, complex adaptive systems, particularly the class of land-use and land-cover change models do not easily lend themselves to probabilistic interpretation.

According to Grünwald (1998, 2005) all models are probabilistic, not in the traditional sense that all models can be considered distributions from which data are drawn, but in a more fundamental way that all models give a probability to data. In his doctoral thesis (1998) Grünwald presents a method called *entropification* that associates non-probabilistic model classes with probabilistic ones using squared errors the model makes with respect to data. Lee (2006) uses this method for choosing the best parameterization for the class of hierarchical generative models in the optimal stopping problem.

I do not consider agent-based LUCC models as probabilistic, but I make the connection to probabilistic models by replacing probabilities as measures of fit with errors. The insight is based on the fact that the code length can be associated to model fit; the better the fit, the shorter the resulting code for the data and vice versa (Grünwald, 1998). Specifically, a high probability given to data by a probabilistic model class implies a short code length. Likewise, a small error made by a non-probabilistic class means a short code length. The connection between probabilities, the domain of NML criterion[4], and errors, the domain of NME criterion, is made in the following steps:

[4]Here, 'probability' is strictly used as a measure of fit; high probability does not indicate that a particular data is very likely, but the model class fits the data well. The probabilities in the denominator in the definition of the NML criterion (cf. equation 4.2) are produced by different (maximum likelihood) parameter values, and consequently do not sum up to one. If the model class is sufficiently flexible, it is possible to find sets of parameter values, i.e., models within the class, so that the class gives high probabilities to all data samples.

1. The correspondence between probabilities and code lengths is established by the following theorem (Grünwald, 2005), derived from *Kraft inequality* (Cover & Thomas, 1991): If $\mathcal{X}^n$ is an finite and countable set and P a probability distribution on $\mathcal{X}^n$, then there exists a prefix code C for $\mathcal{X}^n$, such that for all $x^n \in \mathcal{X}^n$, $L_C(x^n) = -\log P(x^n)$. Similarly, if C' is a (prefix) code for $\mathcal{X}^n$, there exists a probability distribution P', such that for all $x^n \in \mathcal{X}^n$, $-\log P'(x^n) = L_{C'}(x^n)$.

2. The result of Rissanen (1989) establishes the correspondence between errors and code lengths: for each $H \in \mathcal{M}_i$ there exists a probability distribution $P(\cdot|H)$ such that for all x^n,

$$-\log P(x^n|H, \mathcal{M}_i) = ER(x^n|H, \mathcal{M}_i) + K, \tag{4.3}$$

   where K is a constant that does not depend on the sample x^n or the model H, but may depend on the data size n. The existence of the probability distribution implies the following equality:

$$L_C(x^n|H, \mathcal{M}_i) = ER(x^n|H, \mathcal{M}_i) + K. \tag{4.4}$$

3. The final step is to show that a code C can be constructed such that the equation 4.4 holds, namely that the coding of errors is prefix and the constraints on constant K exist. These are shown in the next section.

Using the established relationship between probabilities and errors, we can

associate maximization of probabilities with minimization of errors, and consequently maximization of the NML criterion to minimization of the NME criterion:

$$\begin{aligned}
\arg\max_{\mathcal{M}} NML(x^n, \mathcal{M}_i) &= \frac{P(x^n|\hat{\theta}_i(x^n))}{\sum_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}_i(y^n))} && (4.5)\\
&= \frac{2^{-L_C(x^n)}}{\sum_{y^n \in \mathcal{X}^n} 2^{-L_c(y^n)}} && (4.6)\\
&= \frac{2^{-(ER(x^n)+K)}}{\sum_{y^n \in \mathcal{X}^n} 2^{-(ER(y^n)+K)}} && (4.7)\\
&\sim && (4.8)\\
\arg\min_{\mathcal{M}} NME(x^n, \mathcal{M}_i) &= \frac{ER(x^n)+K}{\sum_{y^n \in \mathcal{X}^n} ER(y^n)+K} && (4.9)
\end{aligned}$$

Sketch of Prefix Code for Errors

In the previous section I showed that for any model class $\mathcal{M}_i$ such a code C exists that gives short code lengths to small errors, particularly to the minimum error, and how these relate to high probabilities. In this section I demonstrate with an example that the minimum error values meet the requirements so that we can construct such code C for LUCC models. The following two conditions need to be met:

Condition 1. The error space is finite and the values enumerable. This ensures unique decodability, i.e., the codewords for errors are prefix-free.

Condition 2. The total code length is a function of data size n, and does not depend on the model class or individual data samples. This assures K in equations 4.3 and 4.4 is a constant.

As mentioned above, a model is not enough to determine its fit to data, also an error function is required. The error depends on a particular spatial metric used

to characterize the landscape; different metrics unveil different characteristics in a model's behavior, and they vary in magnitude. However, we want the description method or code to be independent of specific values. The approach adopted here applies to any error function presuming the error space can be truncated and/or discretized to a finite and countable set. In the context of land-use models, when the error measures a landscape characteristic, this is always guaranteed.

Let me remind the reader that an individual error value is either a direct difference between two land-covers (mean absolute difference (m.a.d.)), or a sum over time of square error between two landscape metrics (composition (c), edge density (e.d.), and mean patch size (m.p.s.)), cf. Section 3.6.

Let us consider a finite and countable set of error values $e \in E$, so that $\varepsilon_{min} \leq e \leq \varepsilon_{max}$, where ε_{min} is trivially zero, and ε_{max} usually depends on landscape size or temporal span of the modeled period, or both. However, we only need codes for non-zero errors. Once we have determined the range $[\varepsilon_{min}, \varepsilon_{max}]$, we can discretize error values with feasible precision and rank the errors by their magnitude. Finally, we can construct a rank-based code so that smaller errors receive shorter codewords and larger errors longer codewords.

Using a fixed precision d for errors, meaning that we can describe $D = 2^d$ distinct error values, gives ranks $r \in (1, 2, \ldots, D)$. The rth error, e_r can then be associated with its rank r. Now, we need to encode the ranks r, and use a non-fixed precision for them. We can use the same method as was used when constructing prefix codes for numbers (cf. Section 4.2) by associating the smallest error with number one, the second smallest with number two, ..., and finally the largest error with D.

These considerations guarantee that the first condition is met. To ensure the other condition, the number of overhead bits K used to describe the errors is allowed to depend only on data size, not the model class or individual data samples. This is guaranteed by the fact that the number of these bits is determined by the precision use to describe the errors (or their ranks). However, the precision is a function of the error range; the smaller the range, the less bits are required to describe both the errors and the precision of errors.

Spatially explicit landscape data is usually represented in discrete resolution, for instance a grid of cells, instead of some continuous quantity. In addition to the spatially explicit aspect these models introduce a temporal dimension; they are usually run for several consecutive rounds. The error functions, i.e., the spatial metrics, often use the same spatial and temporal precision in which the data is available. Consequently, the error range is strictly determined by the landscape size and spatial data resolution — in addition to the number and size of individual parcels for heterogeneous agents — and time scale in which the model operates in. This guarantees that K depends on data size only.

Next I present a way of linking the NML principle to the NME criterion using the rank-based coding of errors and the earlier established relation between minimized errors and maximized probabilities.

Let us assume that with the chosen precision for errors, D distinct error values can be obtained with data samples $x^n \in \mathcal{X}^n$. If we replace these errors with their ranks, we have ranks $r \in R = (1, 2, \ldots, D)$. If the same advanced method as was used in Section 4.2 to describe numbers is chosen to encode ranks, we need $\lceil \log r \rceil$ bits to describe each rank r. In addition an extra $2 \log \log r$ bits are required to encode the length of the encoding of r, together with the number of bits needed to

describe this length. Thus, the total description length of a rank r becomes:

$$L_C(r) = 2 \log \log r + \log r. \tag{4.10}$$

In order to illustrate the relationship between the NML criterion and the NME criterion, we start by noticing that maximizing the original version of the NML criterion, NML_{org}, is the same as maximizing the 'average' version of it, NML_{avg}, since they only differ by a constant $\frac{1}{n}$, where n is the number of components in the sum (cf. equation 4.2). On the other hand, maximizing NML_{avg} is essentially the same as minimizing its inverse NML_{avg}^{-1}.

Similarly, the NME criterion can be formulated as minimizing the ratio of the numerator to the mean of denominator. It turns out that NML_{avg}^{-1} closely resembles this 'average' version of the NME criterion, NME_{avg} as explained below. The transformations go as follows:

$$\begin{array}{rcccl} \arg\max_{\mathcal{M}} NML_{org} & = & \arg\max_{\mathcal{M}} NML_{avg} & = & \arg\min_{\mathcal{M}} NML_{avg}^{-1} \\ & \sim & \arg\min_{\mathcal{M}} NME_{avg} & = & \arg\min_{\mathcal{M}} NME_{org}. \end{array} \tag{4.11}$$

Using the rank-based code length function defined in Equation 4.10, these equalities can be demonstrated with simple math. For the clarity of the argument, I only use the largest term $L_C(r) \approx \log r$. The 'average' formulation of the NML criterion is as follows:

$$NML_{avg}(r) \cong \frac{2^{-\log r}}{\frac{1}{n}\sum_{r' \in R^*} 2^{-\log r'}}, \tag{4.12}$$

$$\cong \frac{\frac{1}{r}}{\frac{1}{n}\sum_{r' \in R^*} \frac{1}{r'}}. \tag{4.13}$$

where r is the rank of the minimum error on data sample x^n, R^* contains the ranks associated to minimum errors $\hat{e}'$ made on all $y^n \in \mathcal{X}^n$, and $|R^*| = n$.

If we invert 4.13 we get the following entity to be minimized:

$$NML_{avg}^{-1}(r) = \frac{r}{\frac{n}{\sum_{r' \in R^*} \frac{1}{r'}}}, \tag{4.14}$$

where the term in the denominator is the harmonic mean of the ranks.

On the other hand, the 'average' version of the original NME criterion is given by:

$$NME_{avg}(r) = \frac{r}{\frac{1}{n} \sum_{r' \in R^*} r'}. \tag{4.15}$$

Now, NML_{avg}^{-1} has the harmonic mean of rank values as its denominator, whereas NME_{avg} has the arithmetic mean. The means are the same when the rank values are the same, and quite different for the rank values with large variation, so that the harmonic mean is never larger than the arithmetic mean. The harmonic mean extenuates the effect of larger ranks, i.e., it tends toward the lower ranks. A simple artificial example can be used to demonstrate that criteria using these two means still can select the same model class.

Example 4.3 We consider three candidate model classes $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$ and three data samples $D = (d_1, d_2, d_3)$. Let us assume that the model classes in $\mathcal{M}$ make errors $e \in [1, 10]$ on these data, so that the class $\mathcal{M}_1$ has high error variation, and the classes $\mathcal{M}_2$ and $\mathcal{M}_3$ have very low error variation.

Let us also assume that the errors can be truncated to 100 distinct values. Consequently, they can be encoded using their ranks $R = (1, 2, \ldots, 100)$. For instance, the errors the model class $\mathcal{M}_1$ makes rank 1., 50. and 100. on samples d_1, d_2, and d_3, respectively. These ranks together with their arithmetic and harmonic means are presented in Table 4.1 for the three model classes and data samples. The NME_{avg} and NML_{avg}^{-1} scores corresponding to these ranks and means are shown in Table 4.2. The scores of the selected model classes are shown in boldface.

Model classes:	Ranks			Arithmetic mean	Harmonic mean
	d_1	d_2	d_3		
$\mathcal{M}_1$	1	50	99	50	2.9123
$\mathcal{M}_2$	50	50	50	50	50
$\mathcal{M}_3$	49	50	51	50	49.9867

Table 4.1: Error ranks for three model classes and three data samples, used in Example 4.3, and the two different mean values associated to them.

Model classes:	NME_{avg}			NML^{-1}_{avg}		
	d_1	d_2	d_3	d_1	d_2	d_3
$\mathcal{M}_1$	**.02**	1	1.98	**.3433**	17.1667	33.99
$\mathcal{M}_2$	1	1	**1**	1	**1**	**1**
$\mathcal{M}_3$	.98	1	1.02	.9803	1.003	1.0203

Table 4.2: Average versions of the NME and NML^{-1} scores calculated for the Example 4.3.

Both criteria select the class $\mathcal{M}_1$ for d_1. While NME_{avg} criterion is indifferent between the candidate classes for d_2, NML^{-1}_{avg} slightly prefers the class $\mathcal{M}_2$ over the class $\mathcal{M}_3$. For data d_3 both criteria select the model class $\mathcal{M}_2$. However, the score differences between the classes $\mathcal{M}_2$ and $\mathcal{M}_3$ are negligible with all data samples and both selection criteria. As was insinuated above, the two means and the resulting scores are closest to each other when the error variation is low.

This simple example indicates that for large error variation the harmonic mean's inclination toward small values aggravates score differences between data samples, i.e., the lower ranks in the absolute scale receive relatively lower NML scores than the higher ranks. Consequently, model classes that generate errors with high variation likely fare well in comparison to other models, at least for data samples for which they generate small errors. This implies that the NML^{-1}_{avg} score can be

relatively lenient to flexible model classes. The arithmetic mean does not have a comparable impact. It is sensitive to small and large outliers, but not to the error range itself. The analysis of how the real error distributions affect the exact relationship of these criteria is subject of future research.

Error Range and Precision

For the sake of illustration, the actual minimum and maximum errors the four spatial metrics produce are presented here for homogeneous agents only. However, for heterogeneous agents it is somewhat more complicated to come up with a particular range, since the data has more dimensions to it, namely the number and the sizes of individual parcels, which may vary as a function of each other. Consequently, the calculation of particular values becomes a multi-dimensional optimization problem which resides outside of the scope of this dissertation.

Let T stand for total number of time points, and M ($= m_1 \times m_2$) the size of a rectangular landscape. Ideally the minimum error any model can make is zero, when it either generates a landscape with no cover of interest or perfectly reproduces the cover observed in the real landscape. This is very unusual in reality, though. This unrealistic case set aside, I am going to utilize a more plausible minimum error scheme that in practice can be interpreted as a model making a tiny error at one time point and generating the remaining landscapes perfectly. The maximum error is more straightforward to produce for most of the metrics. The sample values, given in parentheses right of equations, are calculated for the artificial landscape ($m_1 = m_2 = 15, T = 50$) that will be used in Experiment II, Section 5.3.

Mean absolute difference Minimum error is achieved when the land-cover of one

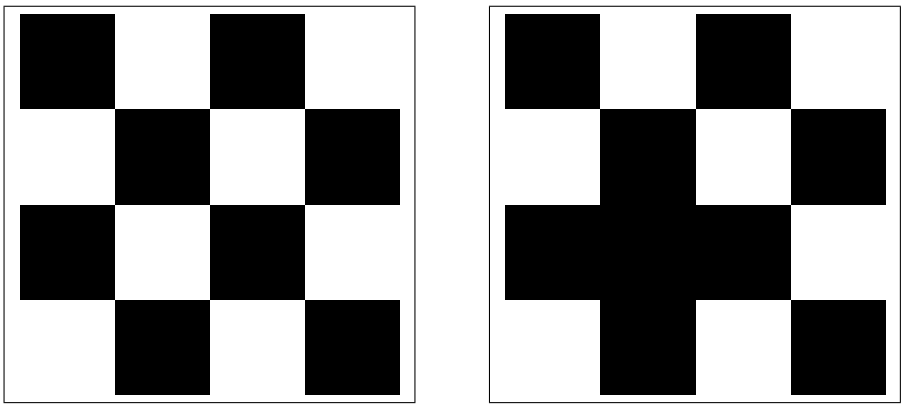


Figure 4.2: Landscapes between which the mean absolute difference produces the minimum error.

cell differs from data at one time point (see Figure 4.2 for example landscapes), whereas the maximum error is produced if the cover of every cell is incorrect at each time point (see Figure 4.3).

$$\begin{aligned} e_{min}(\textit{m.a.d.}) &= \tfrac{1}{M}, \qquad (.0044) \\ e_{max}(\textit{m.a.d}) &= \textstyle\sum_T \tfrac{1}{M} M. \qquad (50) \end{aligned}$$

Composition Again, the minimum error in composition is achieved when the land-cover of interest deviates from data by one cell at one time point. The maximum error is produced when the difference is $M - 1$ cells for each T time points.

$$\begin{aligned} e_{min}(c) &= \tfrac{1}{M^2}, \qquad (1.975 \times 10^{-5}) \\ e_{max}(c) &= T(1 - \tfrac{1}{M})^2. \qquad (49.6) \end{aligned}$$

Edge density The minimum error is produced by a landscape that differs from the data by one cell at one time point. The edges bordering the landscape or some boundary zone around it do not count toward edges. A checkerboard pattern produces the maximum edge, and the maximum error when compared to the

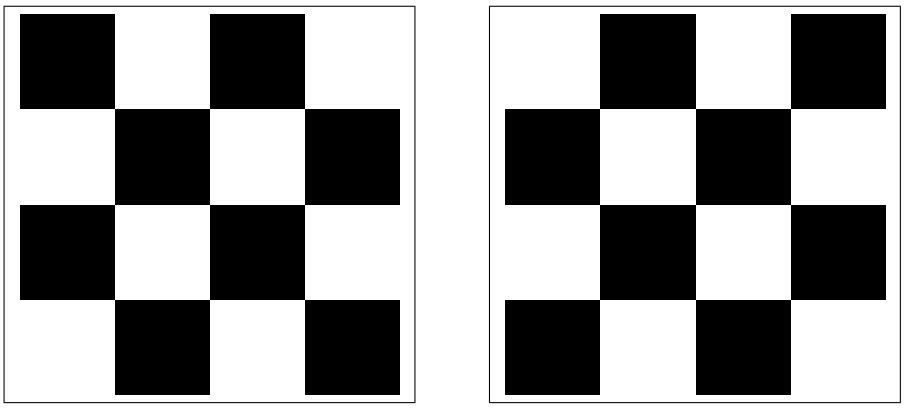


Figure 4.3: Landscapes between which mean absolute difference produces the maximum error.

landscape with minimum edge. Figure 4.4 illustrates the edges that count and edges that do not count towards the edge density.

$$\begin{aligned} e_{min}(\textit{e.d.}) &= < 1 && (< 1) \\ e_{max}(\textit{e.d.}) &\approx \begin{cases} T\left(4 - \frac{2(m_1+m_2)}{m_1 m_2}\right)^2, & \text{if } m_1, m_2 \text{ even} \\ T\left(4 - \frac{2(m_1+m_2+2)}{m_1 m_2+1}\right)^2, & \text{if } m_1, m_2 \text{ odd .} \end{cases} && (691) \end{aligned}$$

The approximation in the maximum error is explained by the fact that I only consider cases when m_1 and m_2 are both either even or odd, and the landscape is approximately square. The actual maximum values vary a little with different ratios of landscape dimensions.

Mean patch size is quite different from the other metrics, since its components vary as a function of each other; the more patches there are, the less area each of them potentially covers. On the other hand, a smaller number of patches do not necessarily take up more space. The minimum error is achieved when the same number of patches differ in their total size by one cell at one time

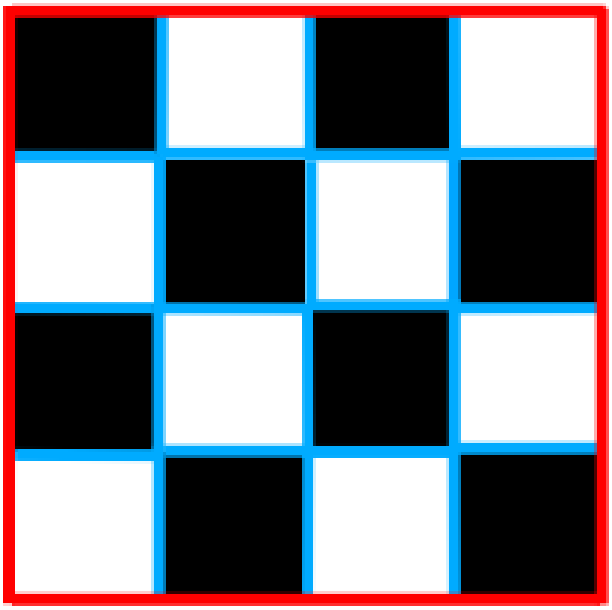


Figure 4.4: Checkerboard pattern produces the maximum error in edge. Blue lines mark the edges that do count towards the edges, and red lines those that do not.

point, whereas the maximum error in mean patch size is the difference between one single cell patch and one M cell patch squared and summed over T time points.

$$\begin{aligned} e_{min}(\textit{m.p.s.}) &\in \quad]0,1[, \qquad (.016) \\ e_{max}(\textit{m.p.s.}) &= \quad T(M-1)^2. \quad (2.5M) \end{aligned}$$

Since the mean patch size depends both on the number of patches and their sizes, the true minimum value is hard to calculate analytically. The value $.016$ was obtained relying on the fact that the maximum number of (single cell) patches on a $m_1 \times m_2$ landscape is $\lceil \frac{m_1}{2} \rceil^2$, with the simplifying assumption that $m_1 = m_2$. The example minimum value is given as the difference between the average patch size of a landscape with $\lceil \frac{m_1}{2} \rceil^2 - 1$ single cell patches and a

landscape with $\lceil \frac{m_1}{2} \rceil^2 - 2$ single cell patches and one two-cell patch:

$$e_{min}(m.p.s) = \left(1 - \frac{\lceil \frac{m_1}{2} \rceil^2 - 1}{\lceil \frac{m_1}{2} \rceil^2 - 2}\right)^2 .$$

These minimum and maximum errors are of course hypothetical, but they give a guideline in deciding which precision to use to describe the errors. What actually would be of interest are the typical error values. Like the *post hoc* analysis of error values produced by different spatial metrics indicates, the typical values are not arbitrary (see Section 5.3 and Appendix A, Section A.4 for error distributions and summary statistics for homogeneous and heterogeneous agents on artificial data). For mean absolute difference they tend to cluster around the median, whereas for the three other metrics (composition, edge density and mean patch size) they tend to gather around zero. Even for m.a.d. the error values predominantly populate the lower end of the scale. Note, that the three metrics are aggregate measures, whereas mean absolute difference takes exact locations into account; thus, it is supposedly harder to fit, and its error distribution is wider (see discussion in Chapter 6).

However, we cannot base the model selection criterion on typical values, since we want the code to be truly lossless; that is, we need to be able to construct a code for each possible error value we may encounter within the limits of the chosen precision.

Summary

The empirical error distributions (presented in Appendices A and B, Sections A.4 and B.1) strongly imply that it would be beneficial to use short codewords for smaller and more frequent errors and longer codewords for larger and more rare errors.

In this chapter I demonstrated the relationship between the normalized minimum error criterion (NME) and the normalized maximum likelihood (NML) criterion. Referring back to the communication interpretation of the MDL principle; when describing a particular data, ideally we want to use its maximum likelihood code. However, for the reasons discussed in section 4.2 this is not feasible. The NML principle has been proved optimal in the sense that it defines a unique model that minimizes so called maximum regret, i.e., the additional number of bits required to describe the data if using this model instead of the optimal maximum likelihood model.

I presented a method for relating errors to code lengths using a rank-based coding method. The rank-based method allows a prefix-free encoding of a finite and countable set of errors, and it assigns short codewords for small errors and longer codewords for larger error. The code lengths were used to associate the errors with probabilities, which are used by the NML principle as a measure of fit.

5

Experimental Evaluation of the Framework

The selection criterion will be evaluated in three sets of experiments using different agent-based learning models and varying domains of land-use and land-cover change with increasing complexity.

Experiment I The first phase functions as a proof of concept by testing the criterion's ability to tell the difference between different model behaviors. A model class consisting of a single agent making decisions between two abstract land-uses is used. The landscape consists of two cells with either homogeneous or heterogeneous suitability.

Experiment II The second phase is designed to analyze the criterion's sensitivity to both exogenous factors, such as initial landscape configuration and biophysical variables, and the spatial metrics used to measure and compare models' performance. This experiment, or rather a battery of experiments, is conducted with a multi-agent model class and a two-dimensional landscape with two abstract land-uses.

Experiment III The goal of the final experiment is to test the selection criterion's adequacy in penalizing for excess flexibility and to further analyze the components of the NME score in various experimental conditions. Real land-cover data from two townships in the Midwestern United States is used.

The first two experiments are conducted with artificial data generated by models belonging to multiple candidate classes, i.e., different learning and decision strategies. Usage of artificial data is purely for academic purposes. The MDL principle does not assume a 'true model' exists — "all models are wrong, some may be useful" (Box, 1979) — but in order to test the soundness of the proposed selection criterion, it is useful to study cases in which a true model exists, but it may or may not be among the candidate models. If the criterion behaves well in all or most of such cases, we can be confident that it behaves reasonably well with real cases.

5.1 Method

Data generation is conducted so that each decision strategy makes repeated land-use decisions and the resulting changes in the landscape are recorded for T time steps. Thus, the data consists of sequences of matrices that record the land-use for all landscape cells. N data samples are generated from each strategy with randomly selected parameter values and from either a random or predefined initial landscape configurations. Particular values of N and T depend on the experiment.

After the data generation, candidate models (usually the same as the generating models, but not necessarily) are fitted to all data samples. The fitting method, outlined next, is common to all three experiments. Currently the Nelder-Mead (Nelder & Mead, 1965) multidimensional minimization algorithm is used to find

the optimal parameter values to minimize the errors between generated data and the outcome of the candidate models, but any non-linear multidimensional optimization algorithm would do, such as genetic algorithms or gradient descent.

In the minimization process, the following steps are executed until the error converges or a preset time limit is met. Currently the limit is 1000 iterations. The index i enumerates minimization iterations. The algorithm is as follows:

0. Set $i \leftarrow 1$. Initialize parameters and the minimization algorithm.
1. Set the parameters p_i for the candidate model H.
2. Run the candidate model H for T time steps from start to end of the modeled period.
3. Calculate the spatial metrics y_H and y_{obs} from the outcome of the candidate model H and the observed data, respectively.
4. Calculate the sum of squares error between the fitted metrics and the observed metrics over time as explained in Chapter 3, section 3.6.
5. Adjust the model parameters according to the minimization algorithm to get p_{i+1}.
6. If converged exit, otherwise set $i \leftarrow i + 1$. Go to step 1.

These steps are repeated for each candidate model class and for all generated data samples. After fitting, the candidate model's NME score for each data sample is calculated. For a particular data sample the selected model class is the one with the lowest score.

5.2 Experiment I

The proposed model selection criterion's adequacy hinges on its ability to distinguish between qualitatively and quantitatively different data, i.e., patterns that result from different behaviors. This first phase of the experimental work is devised to test whether the criterion is capable of identifying the model that generated a specific set of data.

Model Class

For this experiment a spatially explicit agent-based model class with a small number of interacting components and few free parameters is implemented. In this class a single agent makes repeated land-use decisions between two abstract land-uses (called 0 and 1) on a two-cell landscape. The payoff for the land-uses $\vec{x} = \{0, 1\}$ at the time t is:[1]

$$P_t(\vec{x}) = \sum_j u(x_j) + \delta u(1 - x_j)$$

where $u(x_j)$ is the utility obtained from selecting use x on cell j. The latter term represents the discounted forgone utility from not chosen land-use $1 - x_j$. This guarantees that the agent can potentially switch to the land-use it has never opted for. The utilities directly reflect the price trends set for the two land-uses over time; the trend is decreasing in use 0 and increasing in use 1. In addition to the discount factor δ, the agent has two more free parameters α_0 and α_1, which reflect the subjective preferences for two land-uses. These individual preferences come into play in the decision phase.

[1] The payoffs are calculated separately for both land-uses, thus the vector.

Conditions:	Cell 1	Cell 2
1.	(.5, .5)	(.5, .5)
2.	(.9, .1)	(.5, .5)

Table 5.1: Suitability conditions for two landscape cells: condition 1 = homogeneous suitability, condition 2 = heterogeneous suitability.

The question of interest is whether the land-use outcomes vary qualitatively as a function of agent characteristics and landscape heterogeneity. Agent characteristics are defined by varying the decision mechanism and the preference parameters. Landscape heterogeneity is controlled by varying suitability of each cell to the land-uses in two conditions. In suitability condition 1 both cells are equally suitable for both uses, and in condition 2 one of the cells is highly suitable to one use, and practically unfit to the other, whereas the other cell is moderately suitable for both.

The exact values of the suitability conditions are presented in Table 5.1.

These two suitability schemes are expected to result in different land-use histories depending on whether the agent takes the suitability into account or not when making land-use decisions. Four decision mechanisms are implemented and they are named random, ignorant, uninformed and informed. The agent makes the decision for both cells separately. The final choice is stochastic; the probability $p_{t,j}(x)$ of choosing land-use x at the time t on cell j is determined by each decision strategy as follows:

Random The probability of selecting use x (or symmetrically $1-x$) is $.5$ regardless of the time point, cell or cell's suitability.

Ignorant The probability depends on the agent's subjective preference for land-use x, and nothing else, i.e., $p_{t,j}(x) \propto \alpha_x$.

Uninformed Besides the personal preference, the probability also depends on the past payoff received from use x on cell j: $p_{t,j}(x) \propto \alpha_x P_{t-1,j}(x)$.

Informed In addition to the preference and the past payoff, the probability is also a function of the cell's suitability for use x: $p_{t,j}(x) \propto \alpha_x(P_{t-1,j}(x) + s_j(x))$, where s is the suitability of cell j to use x.

The suitabilities do not affect the payoffs; their primary function is to enable the distinction of the informed strategy from three others. On the other hand, the suitabilities can be interpreted as factors that generate non-pecuniary benefit or personal satisfaction to the decision maker.

These decision strategies are used both as generating and candidate model classes.

Hypotheses

The first two strategies, random and ignorant, are assumed to exhibit random decisions or random preferences, whereas the uninformed strategy should be sensitive to price changes over time. The decisions of an informed agent are also expected to reflect the price trends, but suitability condition 2 should hold it back from switching from use 0 totally to use 1 in the first cell, whereas there should be no effect in the second cell.

Since all the decision models have an equal number of free parameters — two preference values and the discount factor in the payoff function — the difference in flexibility is solely due to the number of factors incorporated in the decision strategy. So, each model should be sufficiently flexible to fit the data generated by itself, but not have excess degrees of freedom to fit well to data generated by other model classes, assuming the models behave differently.

Method

Four decision strategies are made to generate 100 data samples with randomly selected parameter values and from random initial landscape configurations. Each strategy is run for $T = 100$ time steps. The initial values for the preference parameters are so that $\alpha_i \in \mathbb{N}(0, 1)$ and $\delta \in \mathbb{U}[0, 1]$.

The following error functions are used:

Error ϵ_1 *Absolute point by point difference* between generated landscape and fitted landscape calculated at each time point and then aggregated over time and over the cells.

Error ϵ_2 *Absolute difference between mean landscapes*. Both generated and fitted landscapes, i.e., the land-uses 0 and 1, are first averaged over each t consecutive time points, and the differences between averages are aggregated over time and the cells.

Results

The NME criterion's behavior is contrasted to a method that only uses the errors, the numerator values of the NME equation, as the selection criterion. The criterion, hereafter called the *ERR criterion*, is as defined in Chapter 3.6:

$$\begin{aligned} ER_{sum}(x^T, H^T) &= \sum_{t \in T} m.a.d.(x^T, H^T) \\ ER_{sq}(x^T, H^T) &= \sum_{t \in T} (y(x^t) - y(H^t))^2, \end{aligned}$$

where y is a value given by a spatial metrics and H^t is the landscape generated by the model H at the time t.

The confusion matrices for the 2×2 design (2 error functions $\times$ 2 suitability conditions) are presented in Figure 5.1 for the NME criterion and in Figure 5.2 for the ERR criterion.

The rows in these matrices present four generating model classes and the columns four candidate model classes in the order: random, ignorant, uninformed and informed. Each cell in the matrices represents the number of times the respective candidate model is selected when the respective generating model was the source of the data; the darker the color, the larger the number.

The matrices suggest a relatively strong effect of both suitability and error function on the NME criterion's tendency to choose the generating model — evidenced by the dark diagonal. With the error ϵ_1 the criterion identifies reliably the random class: 99 times out of 100 samples, and the uninformed class: 72 times out of 100 samples. Suitability condition 2 increases these numbers by one. The selection accuracy for two other classes remains below 50%. However, using suitability condition 2 instead of 1, increases the identification of the informed class from 22% to 39%.

The error ϵ_2 somewhat levels out the selection accuracy, but the criterion still identifies the random and uninformed classes over 50% of the time: 76 and 66 times out of 100, respectively, and the ignorant and informed classes are identified 48 and 39 times out of 100, respectively.

Suitability condition ϵ_2 increases the recognition accuracy for the random, uninformed and informed classes to 78, 68 and 56 times out of 100, respectively. With metric ϵ_2 the number of consecutive time points over which the averaging is done is 20, resulting in 5 error points.

The ERR criterion's accuracy in selecting the generating class is inferior to the NME criterion's when using the error ϵ_1; while it recognizes uninformed class

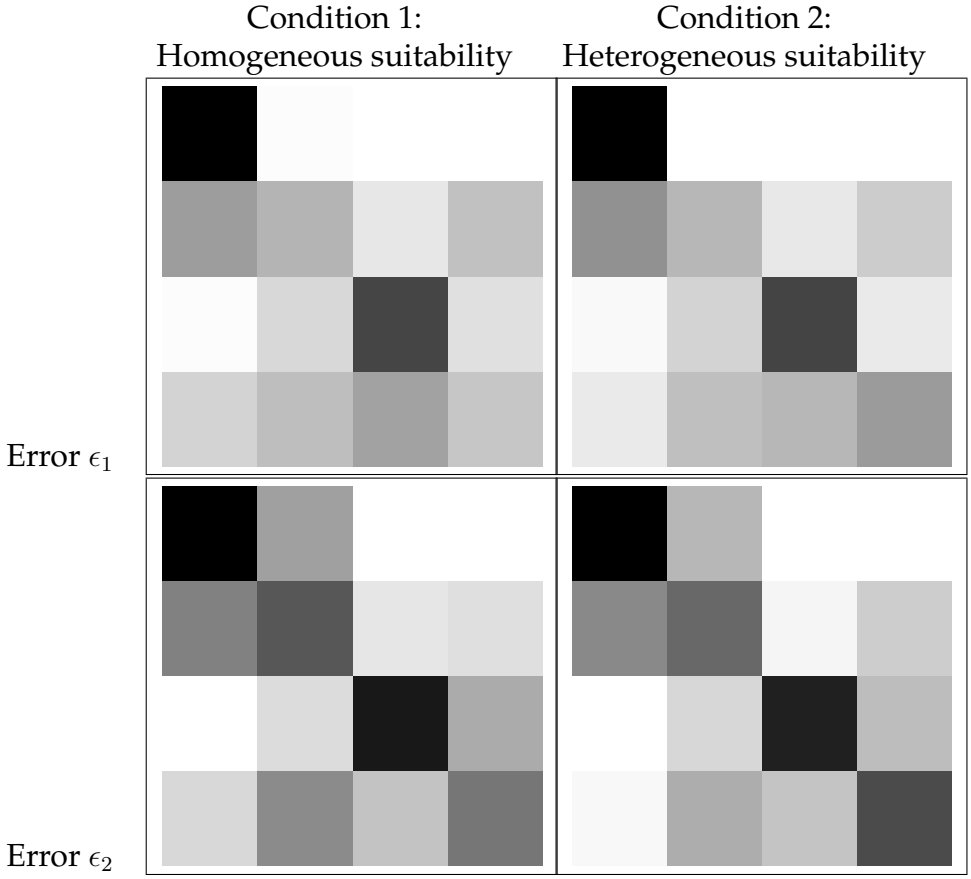


Figure 5.1: Confusion matrices, using the NME criterion, for two error functions and two suitability conditions. Generating and candidate classes are in rows and columns, respectively, in the following order: random, ignorant, uninformed and informed.

pretty reliably (74 times out of 100), as can be seen in Figure 5.2, it cannot tell the difference between four classes when a model from the random or ignorant classes generated the data. However, the criterion's behavior using the error ϵ_2 compares to NME criterion rather well. With the exception of the random class, which the ERR criterion recognizes with almost 100% accuracy, the number of times the other three classes are selected when generating are slightly lower than for the NME criterion. The suitability conditions do not have any observable effect on the ERR criterion's performance.

The NME criterion's sensitivity to the error function is further studied by varying the number of consecutive time points over which the landscapes is averaged. This number is varied from 1 to 100 ($t = 1, 2, 5, 10, 20, 25, 50, 100$), so the actual number of error points summed over are 100, 50, 20, 10, 5, 4, 2, and 1, respectively. Note, that that the first case — averaging over each one time points — reduces to the metric ϵ_1. The results, as the number of times a generating model is selected out of all 400 data samples, are presented for the two suitability conditions in Figure 5.3.

In general, the fewer error points is used, the better the NME criterion identifies the generating class. Furthermore, using suitability condition 2 accentuates this tendency since it facilitates the distinction between the informed class and the others.

Summary

The preliminary results suggest that even in this relatively simple case the proposed selection criterion is sensitive to experimental manipulations, both to factors exogenous to the model (e.g., landscape suitability) and spatial measure used in error calculation.

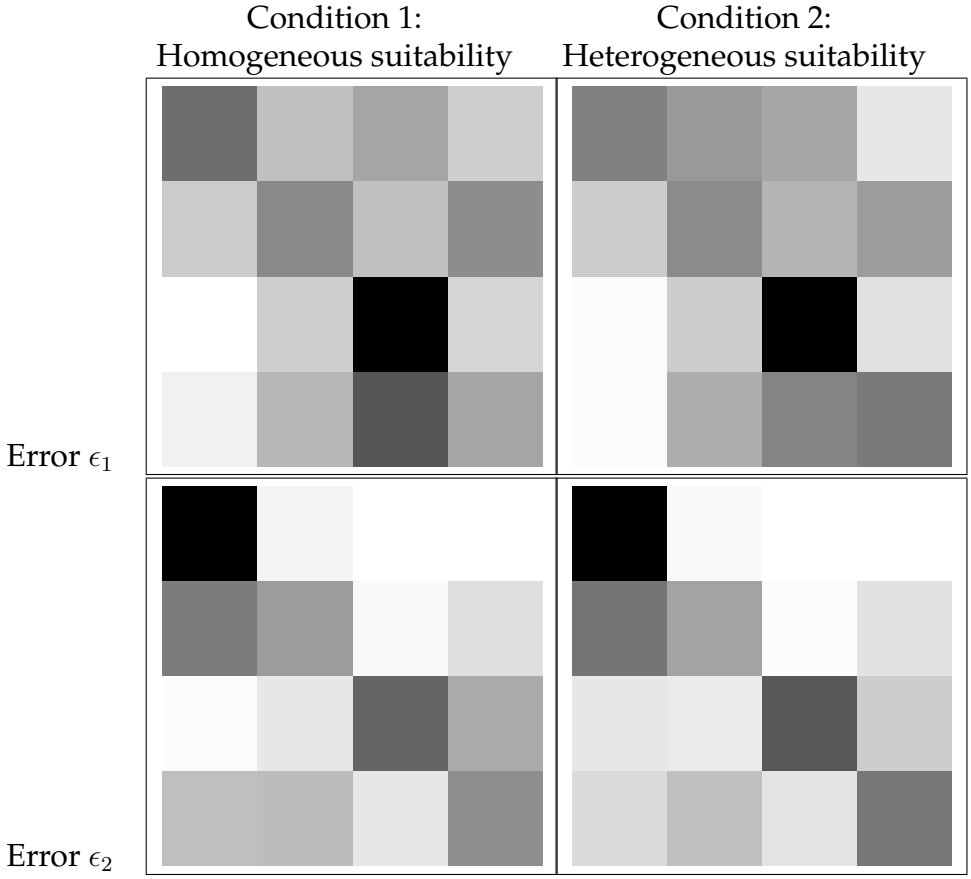


Figure 5.2: Confusion matrices, using the ERR criterion, for two error functions and two suitability conditions. Generating and candidate classes are in rows and columns, respectively, in the following order: random, ignorant, uninformed and informed.

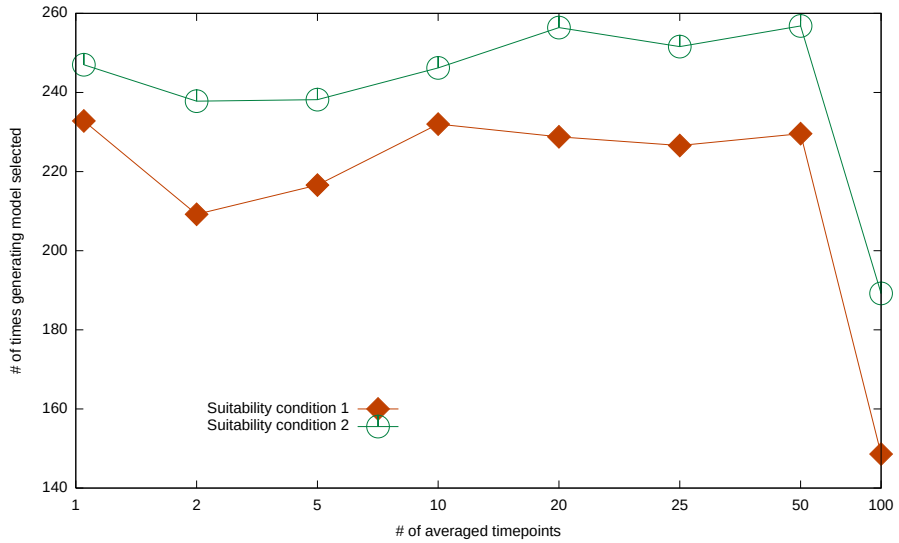


Figure 5.3: The number of times the generating model is selected as a function of the number of averaged time points.

This is by no means an unexpected result, and the effect is assumed to be more pronounced when the model class becomes more detailed and the error calculation scheme more sophisticated, i.e., instead of comparing two landscape covers directly, different spatial metrics are calculated from them and metrics are compared. Different spatial metrics (cf. section 3.6) are able to identify different, more or less coarse or subtle, patterns in the data, and consequently some of them may make few models look unreasonably bad, whereas others, e.g., easy to fit ones, may make all models look unjustifiably good. So, the question becomes, not only which model selection criterion to use, but also how to use it.

The next task is to analyze the influence of spatial metrics and other exogenous factors in the performance of the proposed selection criterion. In the following set of experiments I still use artificially generated data, but instead of aiming at selecting the true generating model, I am focused on the interrelation of landscape

and agent properties, given as exogenous to the model, and spatial metrics used to characterize these properties, and how the selection criterion's behavior is biased by them.

5.3 Experiment II

Acquisition of multiple samples of accurate land-cover data with a good resolution is difficult or at least time consuming. Therefore, in order to extensively test the model selection framework, I need to rely on data generated by an artificial system. In these experiments I use data generated by the same model classes that I use as candidate models. This is a common method in literature when comparing multiple model selection methods (Busemeyer & Wang, 2000; Myung & Pitt, 1997; Myung, 2000; Pitt et al., 2002). The experiments are run in several conditions by varying both the input to the model, i.e., the biophysical and agent characteristics, and the measure with which its performance is assessed, i.e., spatial metrics used in the landscape comparisons.

Data

This experiment employs the framework used by Evans *et al.* (in press) in laboratory experiments in which human participants make repeated spatial resource allocation decisions in an abstract environment, and receive numerical feedback of the success of their decisions. The environment is a 15×15 grid of cells, and it is occupied by nine participants each with a set of 5×5 cells. The participants have an unlimited supply of two resources they can allocate to their cells at each decision round. After the decision, they observe the payoff received from the allocation, the actual prices of resources that were used to calculate the payoff, and

the changes in the whole landscape; they can also see what decisions the other participants made on their cells.

Four types of data are adopted from this experimental framework:

Land-cover One of the starting landscapes, i.e., a distribution of resources (land-uses) on the landscape, used in the laboratory experiments is used as the initial land-cover map in the current experiment. It is always the same regardless of the experimental condition.

Biophysical variables consist of two suitability maps constructed for the resources (land-uses). One of the suitability maps is homogeneous — all the cells have the same suitability value — and the other one is heterogeneous. In heterogeneous map every parcel has exactly the same suitability configuration: they are either equal or mirror images of each other. Each parcel has two corners of high suitability and two corners of low suitability and monotonic slopes in between. Heterogeneous suitability map is shown in Figure 5.4. The mean suitability is the same in both maps.

Ownership The whole landscape is divided equally between nine landowners so that each has a square parcel of 5×5 cells located in a 3 by 3 grid. The parcel borders are displayed in Figure 5.4.

Economic data Two economic data series are generated, one for each land-use. The economic data gives the unit (per cell) price for the land-uses at each time point. The price trends are either monotonically increasing or monotonically decreasing.

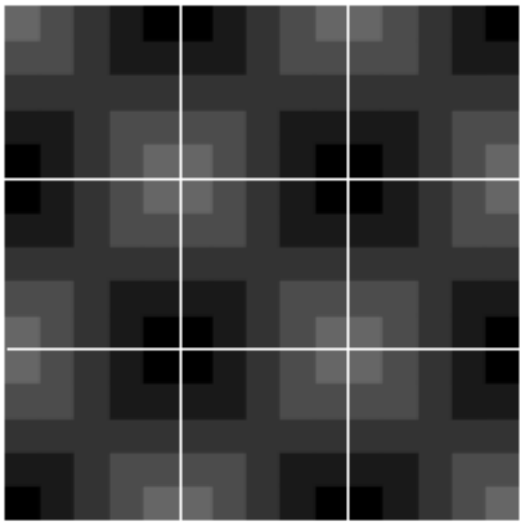


Figure 5.4: Heterogeneous suitability map. A lighter shade means higher suitability and darker shade lower suitability. White lines mark the parcel borders.

Method

The experimental methodology regarding the data generation and model fitting follows the designed sketched in section 5.1. The six decision strategies are, as explained in the Chapter 3: null model, random model, greedy model, Q-learner, individual experience-based attraction model iEWA, and social experience-based attraction model sEWA. This is the assumed order of flexibility of the strategies, determined by the intricacy of the decision process.

Two versions of these strategies are used; one in which all the agents have the same parameter values (in case of generating models) or are fitted a common set of parameter values (in case of candidate models), and one in which all the agents have their individual parameter values (generating models) or are fitted individual parameter values (candidate models). These are considered different model classes. Thus, in this experiment there are twelve model classes. Furthermore, it is

assumed that all collectively fitted classes are simpler than all individually fitted classes.

The $2 \times 4 \times 3$ design varies agent characteristics, spatial metrics and the landscape suitability as follows:

Agent characteristics are defined by the input parameters explained in Chapter 3, Section 3.4: household size, initial wealth and the size of social network for sEWA class. Agents are either homogeneous or heterogeneous. Heterogeneity is achieved by increasing the variation of the initial parameter values while keeping the mean constant. Furthermore, for homogeneous agents the spatial metrics are calculated at the landscape level, but for heterogeneous agents at the individual agent's parcel level and then aggregated over the agents.

Spatial metrics Four spatial metrics are used as described in Chapter 3, Section 3.6: 1. mean absolute difference 2. composition, 3. edge density, and 4. mean patch size.

Landscape suitability values for the two land-uses are varied in three conditions:

I Both land-use have homogeneous suitability, i.e., each cell on the landscape is equally good for both uses.

II Homogeneous map is used for one land-use, and heterogeneous map for another.

III Both land-uses have heterogeneous suitability, one has larger variation in suitability values than the other.

Finally, I want to highlight the distinction between model class, exogenous factors and error analysis methods, since these terms are used extensively in the rest

of this chapter.

Model classes between which the selection is done, consist of six learning strategies in two parameter fitting schemes: collective fitting and individual fitting.

Exogenous factors are input conditions that are common to all model classes in a single experiment, but are not considered to be a part of the class. These include landscape suitability and economic trends.

Error analysis methods, also common to all model classes in a single experiment, are the factors used to assess model performance. These consist of spatial metrics and error calculation schemes, i.e., whether the error is calculated at the landscape or at the individual parcel level. The latter factor goes hand in hand with the assumption of agent heterogeneity (i.e., distribution of agent characteristics), which is both the input to the model, but also a part of the scientific question attempted to answer with the model, namely does the assumption on different agent characteristics yield qualitatively different kinds of behaviors and consequently, different landscape outcomes, that the selection criterion can detect.

In the analysis of the results I am more interested in the performance of the selection criterion than actual land-use outcomes. Four types of analyses are carried out: first, to find out general tendencies in the criterion's behavior in selecting between model classes in different experimental conditions; secondly, to test the effect of exogenous factors and error analysis method, specifically the selection criterion's sensitivity to spatial metrics, landscape suitability and agent heterogeneity; and finally, to analyze the selection criterion's stability and consistency with varying sets of generating and candidate models.

Analysis of Confusion Matrices

Similarly to the Experiment I, the NME criterion's behavior is compared to the ERR criterion and also to leave-one-out cross validation[2], hereafter called the *CV criterion*. After fitting candidate models to data samples produced by generating models (sample size $N = 25$ for NME and ERR criteria, $N = 5$ for CV), the best model, according to the criterion, is selected for each data sample.

The information about the selected classes is collected in the set of confusion matrices. The $2 \times 4 \times 3$ (agent types $\times$ spatial metrics $\times$ suitability conditions) design generates 24 matrices. For the NME and ERR criteria the matrices tell how many times a specific candidate model class is selected when a certain generating model class is the source of the data, whereas for the CV criterion they show the selected model class for each generating model class. The selection is based on the minimum average error the candidate class makes on all data samples generated by each generating class.

The matrices are presented in Appendix A, section A.1 for all experimental conditions: Figures A.1 and A.2, for homogeneous and heterogeneous agents, respectively, with the NME criterion, Figures A.3 and A.4 with the ERR criterion, and Figures A.5 and A.6 with the CV criterion.

Description of Matrices

Table 5.2 shows a schematics view of a confusion matrix. The rows and columns in the matrix represents generating model classes and candidate model classes, respectively, in the following order: null model, random model, greedy model, Q-learner, individual EWA (iEWA) and social EWA (sEWA). The first six rows and

[2]In leave-one-out cross-validation one data sample is held out for validation, and the model parameters are calibrated to the remaining samples.

		Simpler (collective fit) null ⋯ sEWA	More flexible (indiv. fit) null ⋯ sEWA
Simpler (common param. values)	Null Random Greedy Q iEWA sEWA	n_{11}	n_{1n}
More flexible (indiv. param. values)	Null Random Greedy Q iEWA sEWA	n_{m1}	n_{mn}

Table 5.2: Interpretation guide for confusion matrices.

columns represent the models with a single set of parameter values common to all agents (cf. 'simpler' model classes in the discussion in this chapter), whereas the latter six rows and columns represent the models with individual parameter values (cf. 'more flexible' model classes in the discussion). Note, that in the subsequent experiments in which the generating and candidate model classes are varied, there may be fewer rows or columns, or both in the matrices.

The number n_{ij} in each cell indicates the number of times the criterion selects the candidate model in class j when the data is generated by the model in class i. This number is of all the data samples generated by the model in i. Instead of actual numbers, I use shades of gray to illustrate the matrices; the darker the color the larger the number.

Results and Discussion

The preliminary observations reveal the NME criterion's relatively strong tendency to select the generating model in most conditions for homogeneous agents except when mean patch size is used to calculate the errors (lowest row of matrices in Figure A.1. The overall preference for the generating model is weaker for heterogeneous agents; presumably because a number of individual behaviors are harder to recognize than aggregate behavior. The tendency to select the generating model is slightly less accentuated with the ERR criterion and non-existent with the CV criterion.

The confusion matrices with NME criterion also suggest a noticeable effect of spatial metrics in the selection preference for homogeneous agents but not for heterogeneous. Furthermore, there is some interaction between suitability conditions and spatial metrics in whether the selection criterion prefers a more flexible model class when a more flexible class generated the data. In all conditions the criterion tends to choose a simpler class when a simpler class generated the data. The influence of the spatial metrics and the suitability conditions is somewhat noticeable with the ERR criterion and distinct with the CV criterion.

Before going to more detailed analyses of the selection results, I discuss a few interesting issues:

- In the MDL sense, the null model seems the safest choice; it certainly is the simplest class and it fits perfectly few data samples, namely those generated by itself and those generated by classes that make very few changes to the landscape. However, the NME and ERR criteria seldom select the null model, unlike the CV criterion which in most conditions clearly prefers it.[3].

[3]The null model with individually fitted parameter values is never selected in any conditions.

- Both the NME and the ERR criterion's tendency to favor generating models, particularly the simplest and most flexible model classes, is also distinguishable with mean absolute difference, which is somewhat surprising. It is presumably the hardest metric to fit; literature indicates that not a single model has been reported that can predict the changes by exact location as accurately as the null model (Pontius et al., 2004).

- The NME and ERR criteria seldom select a more flexible class when a simpler class generated the data.

- Neither NME nor ERR criterion strongly prefers any single class, but whether they select a generating (or simpler or more flexible) class is contingent on which kind of regularities different spatial metrics detect in different suitability conditions.

- There are also differences in how easy it is for each spatial metrics to fit the data well. For instance, as opposed to the other metrics, an accurate composition measure can be obtained in multiple ways; several different land-cover configurations may have the same percentage of that land-cover, whereas there is only one way to make zero error when measured by mean absolute difference.

The interplay of these factors is subjected to statistical analyses discussed next.

This makes sense, since this is a model class from which the agents have been eliminated and consequently the number of parameters do not have any effect, i.e., both null models fit the data equally well. The fact that the collectively fitted null model class gets selected over the individually fitted class is just a byproduct of how the comparisons are made — both have the same NME score and the first is selected.

Analysis of Sensitivity

The main reason behind the analysis of the proposed criterion, as well as the alternative criteria it is compared to, is to find out if it can effectively guard us against too flexible models, i.e., models that easily overfit. Here I am interested in four separate issues regarding its behavior, and how the choice of the spatial metrics, the landscape suitability, agent heterogeneity and error measurement scheme affect them:

1. How many times is the generating model class selected?
2. How many times is a simpler model class selected overall?
3. How many times is a simpler model class selected when the generating class is simpler?
4. How many times is a more flexible model class selected when the generating class is more flexible?

Method

To answer these questions, four summary statistics are calculated from the 2 by 4 by 3 (agent characteristics $\times$ spatial metrics $\times$ suitability conditions) confusion matrices, and placed in the location corresponding to spatial metric and suitability condition in the respective summary matrix. The cases of homogeneous and heterogeneous agents are dealt with separately. So, two sets of 4 by 3 summary matrices are constructed.

Example 5.1. The summary statistic number 2 (see below) calculated from the confusion matrix resulting from the experiment that uses spatial metric number two (composition), suitability condition number one (homogeneous suitability) and homogeneous agents is placed in the location (2,1) in the summary matrix 2 for homogeneous agents (shown in boldface in Table 5.4), and so on.

The four summary statistics and the corresponding matrices, presented in Tables 5.3 - 5.6, are:

Statistic 1 The fraction of time the generating model class is selected. This number is found by summing up numbers on the diagonal of a confusion matrix.

Statistic 2 The fraction of time a simpler model class is selected overall whether the data is generated by a simpler or more flexible class. This number is found by summing up all numbers in the left side of the confusion matrix, i.e., in the columns labeled 'Simpler' in Table 5.2.

Statistic 3 The fraction of time a simpler model class is selected when a model in a simpler class generates the data. This number is found by summing up numbers in the upper left corner in the confusion matrix, i.e, only in the columns and rows labeled 'Simpler' in Table 5.2.

Statistic 4 The fraction of time a more flexible model class is selected when a model in a more flexible class generates the data. This number is found by summing up numbers in the lower right corner, i.e., only in the columns and rows labeled 'More flexible' in Table 5.2.

It seems that some spatial metrics and suitability conditions make the NME criterion favor simpler or more flexible model classes more than other conditions. For instance, with heterogeneous agents on the homogeneous landscape (suitability

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.34	.43	.44	.33	.32	.39
Composition	.33	.35	.38	.30	.32	.36
Edge density	.43	.43	.46	.30	.35	.38
Mean patch size	.37	.35	.29	.31	.33	.33

Table 5.3: Statistic 1: Fraction of time the generating model class is selected for each spatial metrics, suitability conditions and agent type.

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.77	.67	.69	.62	.69	.75
Composition	**.64**	.63	.61	.66	.65	.65
Edge density	.45	.53	.52	.60	.67	.75
Mean patch size	.58	.65	.73	.69	.67	.67

Table 5.4: Statistic 2: Fraction of time a simpler model class is selected for each spatial metric, suitability condition and agent type. The number in boldface corresponds to Example 5.1.

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.87	.90	.95	.80	.77	.90
Composition	.77	.73	.83	.89	.73	.83
Edge density	.54	.68	.74	.67	.70	.83
Mean patch size	.71	.74	.89	.90	.72	.85

Table 5.5: Statistic 3: Fraction of time a simpler model class is selected when a simpler class generates the data for each spatial metric, suitability condition and agent type.

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.33	.57	.56	.56	.39	.41
Composition	.50	.48	.61	.57	.44	.53
Edge density	.65	.61	.69	.47	.37	.33
Mean patch size	.55	.43	.43	.51	.38	.50

Table 5.6: Statistic 4: Fraction of time a more flexible model class is selected when a more flexible class generates the data for each spatial metric, suitability condition and agent type.

condition I) the criterion prefers a more flexible model class when a more flexible one generates the data (see Table 5.6; the percentages in the column corresponding to the suitability condition I are larger than in other conditions). On the other hand, edge density makes the criterion select more flexible model classes for homogeneous agents when a more flexible class generates the data (see Table 5.6; the percentages on the respective row and columns are consistently larger than for other metrics), whereas mean absolute difference makes it select simpler classes more often for most suitability conditions compared to other metrics (see Table 5.4).

These complex interactions are subjected to statistical analysis to test if the observed effects are significant. The reason I choose to use traditional statistical tools instead of model selection methods in testing the effects are twofold; first, I am not really doing model selection, but the questions I am interested in lend themselves naturally to be formulated as hypotheses; and secondly, the usage of standard and familiar tools do not unnecessarily complicate the analysis.

The specific questions I want to ask about the summary matrices are:

1. Are the differences between the number of times the generating (or simpler,

	Selected	Not selected	Row sum
Mean absolute difference	639	261	900
Composition	562	338	900
Edge density	451	449	900
Mean patch size	588	312	900
Column sum	2240	1360	3600

Table 5.7: Two-way contingency table for testing the statistical significance of the differences in the number of times a simpler class is selected for homogeneous agents.

more flexible etc.) model class is selected statistically significant if analyzed across the spatial metrics?

2. Are the differences between the number of times the generating (or simpler, more flexible etc.) model class is selected statistically significant if analyzed across the landscape suitability conditions?

I use χ^2 test to find out if these differences are statistically significant. Moreover, I am interested in if the error calculation scheme (landscape vs. parcel level) and agent heterogeneity make any difference in either of these cases.

Results and Discussion

For instance, in order to test the effect of spatial metrics in selecting a simpler model class for homogeneous agents, the two-way contingency table shown in Table 5.7 is constructed from Statistic 2 in Table 5.4.

The remaining fifteen two-way contingency tables are constructed in the same way. The results of the χ^2 tests are described below.

Statistic 1 (Generating class selected) The differences that are significant are spatial metrics for homogeneous agents ($\alpha = .005$), and the landscape suitabilities for heterogeneous agents ($\alpha = .025$).

Statistic 2 (Simpler class selected) Likewise, the differences that are significant are the spatial metrics for homogeneous agents ($\alpha = .005$) and the suitabilities for heterogeneous agents ($\alpha = .01$).

Statistic 3 (Simpler class selected when generating) For both agent types all differences are significant ($\alpha = .005$)

Statistic 4 (More flexible class selected when generating) For heterogeneous agents all differences are significant ($\alpha = .005$), for homogeneous agents only the spatial metrics are significant ($\alpha = .005$).

The results of these tests are summarized in Table 5.8 as the lowest significance level [4] α at which the test statistic χ^2 is higher than the critical value χ^2_α. In other words, it gives the upper limit to the probability that the differences this large or larger would be observed if the null hypothesis is true, the null hypothesis being that there is no difference in the number of times a generating (or simpler, more flexible, etc.) class will be selected given different spatial metrics and landscape suitabilities. I only list the significance levels less than or equal to 5%.

To summarize, the selection of a specific class, be it generating, simpler or more flexible, is more likely affected by the spatial metrics with homogeneous agents, and by the landscape suitability with heterogeneous agents. This is understandable for two reasons. First, for homogeneous agents the potential interrelation between landscape heterogeneity and agent heterogeneity[5] may be lost, since the

[4] This is the α level found in χ^2 tables, for instance in McClave (2003).

[5] Even the homogeneous agents vary somewhat but less than heterogeneous agents.

	Homogeneous agents		Heterogeneous agents	
	Spatial metrics	Suitability	Spatial metrics	Suitability
Statistic 1	$\alpha = .005$			$\alpha = .025$
Statistic 2	$\alpha = .005$			$\alpha = .01$
Statistic 3	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$
Statistic 4	$\alpha = .005$		$\alpha = .005$	$\alpha = .005$

Table 5.8: Summary table for χ^2 tests with the NME criterion. Empty entries indicate that the differences are not significant at any level.

errors are calculated on the landscape level. The aggregated spatial metrics may obliterate the potential individual land-use differences resulting from the heterogeneous suitability. On the other hand, the interaction between landscape heterogeneity and agent characteristics may allow more variability in decisions and consequent landscape outcomes. This spatial heterogeneity, characterized by various spatial metrics, may facilitate the selection criterion's task.

The results reported here are from the the experiments in which each candidate model generates 25 data samples, which makes a total of 300 data samples per analysis. Since this is not by any means a realistic number of samples that are usually available of real domains, the experiment has been repeated with smaller sample sets, e.g., with 5 and 10 samples generated by each model. As can be expected, some of the significant differences only appear with larger sample sizes.

Comparison to ERR Criterion

Before going to the next level of analysis, the same sensitivity tests as above are carried out with the ERR criterion. Instead of reporting the fraction of time a generating (or simpler or more flexible etc.) class is selected by this criterion, I list the differences to the fractions reported for the NME criterion. These are shown

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.03	.083	.1	.03	.02	.07
Composition	0	-.01	-.007	.03	-.01	-.007
Edge density	0	-.01	-.007	-.007	.03	-.01
Mean patch size	.03	-.07	-.003	-.007	.003	.03

Table 5.9: Statistic 1: Difference between the NME criterion and the ERR criterion in the fraction of time the generating model class is selected.

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.03	.04	.06	-.01	.04	.02
Composition	.04	.08	.03	-.003	-.01	.01
Edge density	-.03	.007	.04	.003	-.4	.003
Mean patch size	.06	.2	.2	.07	.01	.3

Table 5.10: Statistic 2: Difference between the NME criterion and the ERR criterion in the fraction of time a simpler model class is selected. The number in bold corresponds to Example 5.1.

in Tables 5.9 - 5.12 for each spatial metrics, suitability conditions and agent type. Positive differences indicate that the NME criterion ranks higher in the number of times the specific class (e.g., generating) is selected, i.e., it selects it more often.

Compared to the ERR criterion, the NME criterion selects a simpler class more often, especially if the generating class is simpler, whereas it seldom selects a more flexible class even if the generating class is more flexible. Both criteria select the generating class equally often.

The sensitivity analysis of the ERR criterion is summarized in Table 5.13. The greatest difference compared to the NME criterion is that spatial metrics seem to have more impact overall in which class is selected (generating, or simpler, etc.).

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	.07	.08	.08	.007	0	.07
Composition	.04	.07	.02	.007	-.05	-.007
Edge density	-.07	0	.01	-.007	-.04	-.007
Mean patch size	.07	.1	.2	.1	-.04	.2

Table 5.11: Statistic 3: Difference between the NME criterion and the ERR criterion in the fraction of time a simpler model class is selected when a simpler class generates the data.

Agent type:	Homogeneous			Heterogeneous		
Suitability condition:	I	II	III	I	II	III
Mean absolute difference	0	-.007	-.03	.03	-.08	.02
Composition	-.03	-.09	-.04	.02	-.03	-.03
Edge density	-.01	-.01	-.07	-.01	.03	-.01
Mean patch size	-.05	-.3	-.3	-.04	-.07	-.03

Table 5.12: Statistic 4: Difference between the NME criterion and the ERR criterion in the fraction of time a more flexible model class is selected when a more flexible class generates the data.

	Homogeneous agents		Heterogeneous agents	
	Spatial metrics	Suitability	Spatial metrics	Suitability
Statistic 1	$\alpha = .005$			$\alpha = .05$
Statistic 2	$\alpha = .005$		$\alpha = .005$	
Statistic 3	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$	
Statistic 4	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$

Table 5.13: Summary table of χ^2 tests with the ERR criterion, Empty entries indicate that the differences are not significant at any level.

Comparison to Cross-Validation

The cross-validation criterion is different from the other two criteria in two fundamental ways; first, instead of trying to find the best class for each data sample, it selects a model class that best explains all observed data, assuming they come from the same source, and secondly, the error a model makes on a data sample depends on its fit to other samples, whereas with the NME and ERR criteria all data samples are fitted independently. For these reasons there is little meaning in using the same analysis system with the CV criterion as was used with the other two criteria. On the other hand, the entries in the contingency tables would be too small for the χ^2 test to apply. However, Fisher's exact test (Fisher, 1922) could be used in this case with low expected values. Currently the comparison to the CV criterion is limited to an observational analysis of the confusion matrices, and any further statistical testing of the criterion is subject to future work.

Hold-out Analysis of MDL

In the first set of experiments, when exactly the same generating and candidate model classes are used, the proposed criterion identifies the generating class relatively reliably and favors simpler model classes in most of the experimental conditions. The natural question to ask is, how and if the results change if, for instance, the generating model is excluded from the candidate set. This corresponds to the situation with real data, when we do not have the privilege of knowing what the 'true' model is, and we cannot make any assumptions on its presence among the candidate classes.

Method

In order to test the selection criterion's sensitivity to the composition of the generating and the candidate sets the above analyses[6] are repeated by manipulating these sets in the following ways:

1. The set of candidate model classes is held as before. The generating model classes are grouped in sets of two: null and random, greedy and Q, and iEWA and sEWA, for both collectively and individually fitted classes with the exception of null and random classes, since in these two versions parameters do not make any difference. Each of these five pairs is used as a generating set in turn.

2. The same set of generating model class pairs is used so that in each experiment the respective generating pair is removed from the candidate set.

Statistical Analysis and Results

The confusion matrices resulting from the experiments manipulating the generating set are presented in Appendix A, section A.2. The statistical analysis is broken down by the generating model class pairs. Since, the relative simplicity or flexibility, quantified in the amount of computation, is about equal within each pair, I concentrate only to the first two summary statistics: the significance of differences in the number of times the generating model class is selected and in the number of times a simpler model class is selected regardless of the generating class. As above, I give the lowest significant level (α value) less than or equal to 5% at which the differences are significant.

[6]The NME scores are recalculated using the manipulated sets and the χ^2 test are rerun.

Null and random The confusion matrices for homogeneous and heterogeneous agents are presented in Figures A.7 and A.8, respectively. χ^2 tests indicate that for both agent types the differences in the number of times both the generating class (both at $\alpha = .005$) and a simpler class is selected are significant across spatial metrics ($\alpha = .005$ and $\alpha = .01$ for homogeneous and heterogeneous agents, respectively). The criterion selects the generating class 100% of time when the data is generated by the null model, but only 33% of time when it is generated by the random model.

Greedy and Q (collective fit) The confusion matrices are presented for homogeneous and heterogeneous agents in Figures A.9 and A.10, respectively. The differences in the number of times the generating model is selected are significant for both agents across the suitability conditions ($\alpha = .005$). The differences in the number of times a simpler class is selected are significant across the spatial metrics ($\alpha = .005$) also for both agent types. In addition, these differences are significant across the suitability conditions for heterogeneous agents ($\alpha = .005$). The greedy class is selected 69% of the time, when it generates the data, and the Q learner only 23% of the time.

The χ^2 test gives somewhat different results for individually fitted classes. For homogeneous agents the differences in the number of times the generating class is selected become significant for spatial metrics ($\alpha = .005$), and the number of times a simpler class is selected become significant for suitability ($\alpha = .05$). For heterogeneous agents the only significant difference remaining is in the number of times a generating class is selected ($\alpha = .01$). The confusion matrices are presented in Figures A.13 and A.14. Individual fitting also drops the number of times the criterion selects the generating model; the greedy class is selected only 33% of the time and the Q learner only 21% of

the time.

iEWA and sEWA (collective fit) The confusion matrices are presented for homogeneous and heterogeneous agents in Figures A.11 and A.12, respectively. The differences in the number of times a generating model is selected is significant for heterogeneous agents only across suitability conditions ($\alpha = .005$) and for homogeneous agents across spatial metrics ($\alpha = .05$). For the homogeneous agents the differences in the number of times a simpler class is selected are significant along both dimensions ($\alpha = .005$), but for heterogenous agents the only across the spatial metrics ($\alpha = .01$). The percent of time the generating model is selected is 25% for the iEWA and only 12.5% for sEWA.

Again individual fitting changes the significances drastically. The changes are that for homogeneous agents the only differences that are significant are across spatial metrics in the number of times a simpler class is selected ($\alpha = .025$), and for the heterogeneous the differences in the number of times a simpler class is selected are significant across suitability conditions ($\alpha = .01$). The confusion matrices are presented in Figures A.15 and A.16. Individual fitting reduces the number of time the generating model is selected to 17.8% for iEWA, but increases it to 17% for sEWA.

These results are summarized in Table 5.14. Two observations can be made. First, even if the selection criterion shifts its preference for more flexible classes, when the more flexible classes are the generating classes (evident in the confusion matrices by observation), it is increasingly harder for the criterion to select the generating class. Secondly, the same trend can be noted as before; the criterion is more sensitive to spatial metrics when agents are homogeneous and to suitability conditions when agents are more heterogeneous, especially when more flexible

Gen. class	Statistic	Homogeneous agents		Heterogeneous agents	
		Spatial metrics	Suitability	Spatial metrics	Suitability
Null & random	1	$\alpha = .005$		$\alpha = .005$	
	2	$\alpha = .005$		$\alpha = .01$	
Greedy & Q	1		$\alpha = .005$		$\alpha = .005$
(c)	2	$\alpha = .005$		$\alpha = .005$	$\alpha = .005$
Greedy & Q	1	$\alpha = .005$	$\alpha = .005$		$\alpha = .01$
(i)	2	$\alpha = .005$	$\alpha = .05$		
iEWA & sEWA	1	$\alpha = .05$			$\alpha = .005$
(c)	2	$\alpha = .005$	$\alpha = .005$	$\alpha = .01$	
iEWA & sEWA	1				$\alpha = .005$
(i)	2	$\alpha = .025$			$\alpha = .01$

Table 5.14: Summary table for χ^2 tests for data using full candidate model classes, and reduced set of generating classes. Empty entries indicate that the differences are not significant at any level. (c=collective parameter values, i=individual parameter values)

classes generated the data. For the simplest classes spatial metrics are more critical for both agent types, whereas for the medium flexibility classes there is practically no difference.

This sort of manipulative analysis, of course, is impossible with real data since the generating classes are not available. Therefore, it is extremely important to use realistic but artificial data sets to thoroughly test the potential hazards in the selection process that one may not be aware of.

The results of the experiments in which the candidate set was also manipulated are presented next. The confusion matrices of these experiments are presented in Appendix A, section A.3. Again, the statistical analysis is broken down by the generating model class pairs. Since the generating models are excluded from the candidate sets, this set of analyses only tests the significance in the number of times

a simpler class is selected.

Null and random The confusion matrices for homogeneous and heterogeneous agents are presented in Figures A.17 and A.18, respectively. χ^2 tests indicate that for both agent types the differences in the number of times a simpler class is selected are significant across spatial metrics ($\alpha = .005$).

Greedy and Q (collective fit) The confusion matrices are presented for homogeneous and heterogeneous agents in Figures A.19 and A.20, respectively. For both agent types the differences in the number of times a simpler class is selected are significant across both dimensions ($\alpha = .005$).

Individual fitting changes the results somewhat. The differences across the spatial metrics remain significant for homogeneous agents ($\alpha = .005$) and across suitability conditions for heterogeneous agents ($\alpha = .005$).

iEWA and sEWA (collective fit) The confusion matrices are presented for homogeneous and heterogeneous agents in Figures A.21 and A.22, respectively. For the homogeneous agents the differences in the number of times a simpler class is selected are significant across both dimensions ($\alpha = .005$). No other differences are significant.

Using the individually fitted classes abolish all other significances but across spatial metrics for homogeneous agents ($\alpha = .005$).

Summary of these results is presented in Table 5.15.

Discussion

Except for a couple of cases, if generating models are included in the candidate set, the overall preference is for simpler models regardless of the spatial metrics or

Gen. class	Statistic	Homogeneous agents		Heterogeneous agents	
		Spatial metrics	Suitability	Spatial metrics	Suitability
Null & random	2	$\alpha = .005$		$\alpha = .005$	
Greedy & Q (c)	2	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$	$\alpha = .005$
Greedy & Q (i)	2	$\alpha = .005$			$\alpha = .005$
iEWA & sEWA (c)	2	$\alpha = .005$	$\alpha = .005$		
iEWA & sEWA (i)	2	$\alpha = .005$			

Table 5.15: Summary table for χ^2 tests for data using sets of candidate model classes from which the generating classes are removed. Empty entries indicate that the differences are not significant at any level. (c=collective parameter values, i=individual parameter values)

landscape suitability. In general, if the generating model is omitted, there is more variation in the classes the criterion selects, and a visible shift towards selecting more flexible classes, especially with the classes with collective parameter values. In the individual parameter case there is not much difference if the generating class is excluded or not; the overall pattern in the confusion matrices is the same.

If simpler classes are omitted, such as null and random or greedy and Q, focus changes to more flexible classes depending on the suitability condition. On the other hand, if the more flexible classes are excluded the simpler models are selected, or the selection preference is more evenly spread over all the candidate classes.

In general, the selection patterns observed in the confusion matrices are more variable if the generating classes are excluded from the candidates. Also spatial metrics and suitability conditions seem to have more influence to the criterion's behavior if the generating class is excluded. For instance, if comparing the leftmost columns in Figures A.11 and A.21 (homogeneous suitability condition I), omitting the generating class shifts the selection criterion's focus clearly towards the more

flexible end of the scale. Furthermore, the exclusion of the generating class increases the number of significant differences for simpler classes but decreases it for more flexible classes.

NME criterion and Model Classes

In the above experiments the NME scores were used to select between classes. It is not very clear from all experimental conditions how the magnitudes of the numerators and the denominators relate to each other. In order to study this, the values of numerators and denominators of the NME scores given to the model classes are plotted against each other over all the suitability conditions: (c) marks collectively fitted classes, and (i) individually fitted. The scatter plots are shown in Figures 5.5 - 5.8.

Again, there are observable differences in the spatial metrics how the components of the NME score relate to each other:

Homogeneous agents For edge density and mean patch size larger denominators mean larger variation in numerator values, whereas for mean absolute difference the trend is quite the opposite; a larger denominator indicates exclusively larger numerators — there is more variation in numerator for smaller denominator values. For composition the relation is less obvious.

Heterogeneous agents For mean absolute difference the pattern is similar to the previous case with homogeneous agents, but composition and mean patch size show a slight tendency of larger numerators to associate with larger denominators. For edge density, the numerators concentrate in the smaller end of the scale regardless of the magnitude of the denominator.

Summary

Only a subset of tests that can be run on the model selection criterion have been conducted. Other considerations — i.e., things that can be varied — include for instance, how the errors are aggregated over time and space. Of course, one can also use a more extensive set of spatial metrics and decision algorithms. The metrics and algorithms chosen for the current experiments represent a relatively wide spectrum of sophistication and complexity.

The results strongly support the hypotheses that the initial conditions, assumptions of agent heterogeneity, and the error analysis methods together with the spatial metrics make a significant difference in the selection criterion's behavior. Therefore I suggest that whenever a scientist wants to compare several competing complex systems using real data and eventually select one for her purposes, she should:

1. Compare models along multiple dimensions and use several different measures for error.
2. Use a set of sufficiently different initial conditions or input values within the scope of the modeled domain and the scientific question she is interested in.
3. Vary the candidate model set substantially in order to see if the selection criterion is stable and consistent.

These measures help to ensure that both the selection process and the selected model are guarded against (often) *ad hoc* or arbitrary choices made when constructing the models or designing the experiments, and thus not influenced by factors not relevant regarding the goal of modeling process.

5.4 Experiment III

The final and third evaluation of the model selection framework is conducted with real data.

Background

The forest cover changes in two townships, Indian Creek and Van Buren, in rural South-central Indiana between the years 1940 and 1998 motivates this modeling study. The available data indicates that the forest cover has undergone a significant increase within the first 15 years of the study period and after that a modest, but gradual increase. The overall increase of forest cover is around 20% in both townships.

Two spatial metrics have been used to characterize the forest composition and pattern in these two townships: percentage of forest of the total landscape area and the length of the forest edge, respectively (Evans & Kelley, 2004; Laine & Busemeyer, 2004b). Figure 5.9 shows the monotonic increase of the percentage of forest and non-monotonic increase in forest edge in Indian Creek. Figure 5.10 shows the changes in the same metrics in Van Buren.

The change has not been unidirectional nor uniform; for instance, both deforestation and afforestation can be seen in the both townships, as can be seen in Figures 5.11 and 5.13. The parcel borders and the steepness of slopes are displayed in Figures 5.12 and 5.14 for Indian Creek and Van Buren, respectively. The relationship between the ownership and the direction of forest cover change is not evident, whereas the steepest slopes have experienced the highest rate of afforestation.

Theories in land and agricultural economics assume that land-use decision

preferences are primarily formed by comparing expected financial benefits from different activities to the potential monetary costs of carrying out these activities. Koontz (2001) conducted an interview study among South-central Indiana land-owners in which he tried to explicate their motives driving land-use decision making. The survey results suggest that non-monetary benefits also play a significant role, especially if the land is not the land-owner's primary source of income.

The goal of the current modeling study is to explain spatial patterns in South-central Indiana by postulating a small set of individual characteristics and learning mechanisms to decision makers and the payoff scheme that combines both monetary and non-monetary benefits. Simulating land-owners' yearly land-use decisions model makes predictions about the set of alternative land-uses, from which the changes in forest cover are predicted. The alternative land-uses are farmland and fallow, i.e., the land left unused as a result of off-farm employment.

Data

Three types of data from these two Indiana townships are used in this experiment: forest-cover data, slope and soil data and land ownership data. In addition to these data sets, farm product and timber prices, wages and forest growth data are imported as exogenous forces.

Landscape Data

The time series of land-cover data were acquired by remote sensing — historical areal photographs or satellite imaging — of the years 1939, 1958, 1967, 1975, 1980, 1987, 1993 and 1998. The slope data was extracted from the topographic maps and the ownership information from the historic parcel maps from the modeled period

(Laine & Busemeyer, 2004b). However, since the current model does not support parcellization, only the parcel boundaries of the year 1957 are used.

These data were encoded into layered raster GIS (Geographic Information System) representation with $50m \times 50m$ resolution. The landscape is divided into a grid of cells of equal size and each layer records one type of information for each cell. The Indian Creek landscape consists of 195 $\times$ 189 cells, and Van Buren has 192$\times$190 cells. A group of cells belonging to an individual landowner is called a *parcel*. There are 190 and 615 landowners in Indian Creek and Van Buren, respectively. Parcel sizes in Indian Creek vary between 12 and 1936 cells, the average size being 185 cells, whereas Van Buren parcel sizes are between one and 362 cells, the average size being 59 cells.[7]

Economic Data

Relatively few data sets of economic variables are available for the modeled period. Crop prices for farming revenues are aggregated from corn and soybean prices per produced unit. This price was derived from the US Census database (Evans & Kelley, 2004). The timber pricing is discussed below, but it is also aggregated from the common hardwood species in the area to form a single unit price for thousand board feet. The off-farm labor wages for the years 1980-1998 are derived from the federal minimum hourly wage and the average yearly wage per job in Monroe County, Indiana. The wages from 1940 to 1980 are extrapolated using the observed correlation between consumer price index and the a middle wage class of Indiana University for 1980-2001, together with the above mentioned two

[7]Since some of the landscapes were reconstructed from areal photographs, all the land-uses could not be accurately and reliably identified. Therefore both landscapes contain 'no data' cells whose land-cover is unknown, or alternatively identifiable but not relevant for the current study. One such area is the Bloomington airport in Van Buren. Because of these 'no data' cells the mean parcel size does not equal to the landscape size divided by the number of land-owners.

wages. Aggregate economic variables are used, since the landscape data does not allow the discrimination of tree species or agricultural crop types, and accurate occupational data of the landowners in the modeled time period is not available.

Forest Data

Since the focus is in the change patterns of the forested land, either deforestation or afforestation, and only forest-cover data is available, the model requires a plausible forest growth model. Moreover, the agent decision making is contingent on forest growth, and vice versa. There are three types of decisions the agents can make about their forested land: 1) let it grow, 2) harvest trees, or 3) cease the current land-use activity and leave the land unused, after which trees start growing back. The harvesting decision depends on either expected or past revenues from harvesting, which in turn depends on the commercial timber value in an agent's parcel. In order to incorporate these decisions in the agent's portfolio, forest growth is modeled as an exogenous process with some simplified assumptions.

Forest Initialization and Growth The starting forest (in 1940) is all 40-year old red oak. The trees of that age are about 11 inches in diameter, which gives 0.66 square feet of basal area per tree. At the 100% stocking rate, the tree density based on the timber management objective, there are 180 trees of this size per acre (Gingrich, 1967; Miles & G. J. Brand, 2001)[8], i.e., 111 trees per $50m \times 50m$ cell (also called a stand).

Two variables are tracked of the forest: basal area per tree and the number of trees per stand. All trees have homogeneous growth of the basal area, which is

[8] All the forest data are expressed in terms of inches, feet and acres, and translated into SI-units for the model.

given in square feet per year by the following equation:[9]

$$g(ft^2/yr) = ((\beta_1 B)^{\beta_2} - \beta_3 B) \times (\beta_4 + \beta_5 SI + \beta_6 CR),$$

where β's are growth coefficients, B is the basal are in square inches, SI is site index, which gives the total height to which dominant trees will grow on the specific site at some predetermined age (determined by the soil type and quality on stand). CR is the crown ratio, i.e., percent of a tree's total height occupied by live crown. After the growth function is applied to stand, the mortality is implemented implicitly by "killing" a number of trees in order to keep the stand below the crowding threshold.

Calculating the Revenues The only tree harvesting decision available to agents is to clear-cut the whole cell. The harvesting occurs only if the trees are more than 11 inches in diameter; below that they do not have any commercial value.

The revenues from a harvested cell is calculated by first translating the trees' basal area into board feet (Avery & Burkhart, 2002). The price for 1000 board feet in 1982 dollars is determined by $p = 2.05t + 132.15$, where t is the year. This function is based on the price \$171.1 of 1000 board ft. in 1957, modified by the 1.2% annual increase in the real price (Hoover & Preston, 2004).

Method

Since most of the data sets, such as the suitabilities and economic trends, are given, unlike in the Experiment II, extensive experimental manipulation is not possible. The only things that are varied in this experiment are the agent heterogeneity and the spatial metrics. The same spatial metrics are used as in Experiment II,

[9] Vicky Meretsky, personal communication

Landscape:	Indian Creek		Van Buren	
	Selected	NME (μ)	Selected	NME (μ)
Mean abs. difference	sEWA (c)	.25 (.4125)	random	.499 (.5875)
Composition	iEWA (c)	.05 (.4152)	Q (c)	.12 (.5848)
Edge density	sEWA (i)	.35 (.4627)	sEWA (c)	.05 (.5373)
Mean patch size	sEWA (c)	.103 (.4061)	iEWA (c)	.49 (.5939)

Table 5.16: Selected model classes and their NME scores for homogeneous agents with landscape level fit (mean scores in parenthesis, c=collectively fitted, i=individually fitted).

namely mean absolute difference, composition, edge density and mean patch size.

Unlike in previous experiments, the actual values of the selection criteria are also analyzed.

Hypotheses

Since the real data has many more agents and more heterogenous parcels than the artificial domain used in the Experiment II, the agent heterogeneity is assumed to make a larger difference in landscape outcomes. Likewise, the more complex suitability map together with the forest growth dynamics and the economic considerations of the timber value, may result in more varying land-use outcomes.

Results

The selected models together with the respective NME scores and their means are presented in Tables 5.16 and 5.17 for homogeneous and heterogeneous agents, respectively. The number of decimal points is determined by how many decimals are needed to distinguish between the NME scores.

Landscape:	Indian Creek		Van Buren	
	Selected	NME (μ)	Selected	NME (μ)
Mean abs. difference	sEWA (i)	.193 (.2485)	iEWA (c)	.59 (.7515)
Composition	greedy (c)	.04 (.1799)	Q (c)	.39 (.8201)
Edge density	Q (i)	.154 (.2050)	sEWA (c)	.67 (.7950)
Mean patch size	greedy (i)	.674 (.7778)	Q (c)	.03 (.2222)

Table 5.17: Selected model classes and their NME scores for heterogenous agents with parcel level fit (mean scores in parenthesis, c=collectively fitted, i=individually fitted)).

For homogeneous agents only one time out of eight is the individually fitted model class selected, whereas for heterogeneous agents three times out of eight. This is roughly what can be expected; when there is more variation in the agent population, there is potentially something to be gained by fitting the agents individually. In other words, the benefit attained in better fit outweighs the cost in extra flexibility.

The fact that the model's NME scores as well as their means are more equal for the two landscapes for homogeneous agents than they are for heterogeneous agents supports the same interpretation. Particularly, the agent heterogeneity is required in order to explain the varying land-cover patterns. However, there are many more agents in Van Buren, which are potentially harder to fit.

The scatter plots in Figures 5.15 - 5.18 show the numerators of the NME scores plotted against denominators for all spatial metrics and model classes, where (c) marks collectively fitted classes, and (i) individually fitted. These confirm the observation that in general the errors are larger with Van Buren than with Indian Creek data; with heterogeneous agents the difference is more considerable than with homogeneous agents.

Two trends are observable; the errors are either clustered around smaller and

large values, so that larger numerators associate to larger denominators, or there is almost a linear and continuous relationship between numerators and denominators. Homogeneous agents show slightly more clustering than heterogeneous agents. Furthermore, heterogeneous agents in all model classes generate equal errors for Indian Creek with the exception of mean patch size which shows quite an opposite trend. These results imply that the agent heterogeneity may explain more variation in the land-cover changes in Indian Creek than the actual decision algorithms.

Summary

In general, the selection criterion selects simpler models, i.e., collectively fitted classes, for homogeneous agents with both data sets. However, with heterogeneous agents it predominantly selects individually fitted classes for Indian Creek, but collectively fitted for Van Buren. There are two possible explanations to this: either the agents heterogeneity plays a bigger role in Indian Creek and some of the models classes are able to capture it, or the number of agents in Van Buren are hard to fit, so the selection criterion resorts to making a safe decision and selects simpler model classes.

Finally, null and random classes are seldom selected, unlike with artificial data, when they were preferred a substantial amount of time. This implies that the real landscapes are not stationary, but they undergo very characteristic changes which cannot be captured by a random process.

To summarize, with these real data sets the proposed model selection criterion exhibits relatively stable and consistent behavior, similar to what was observed with artificial data.

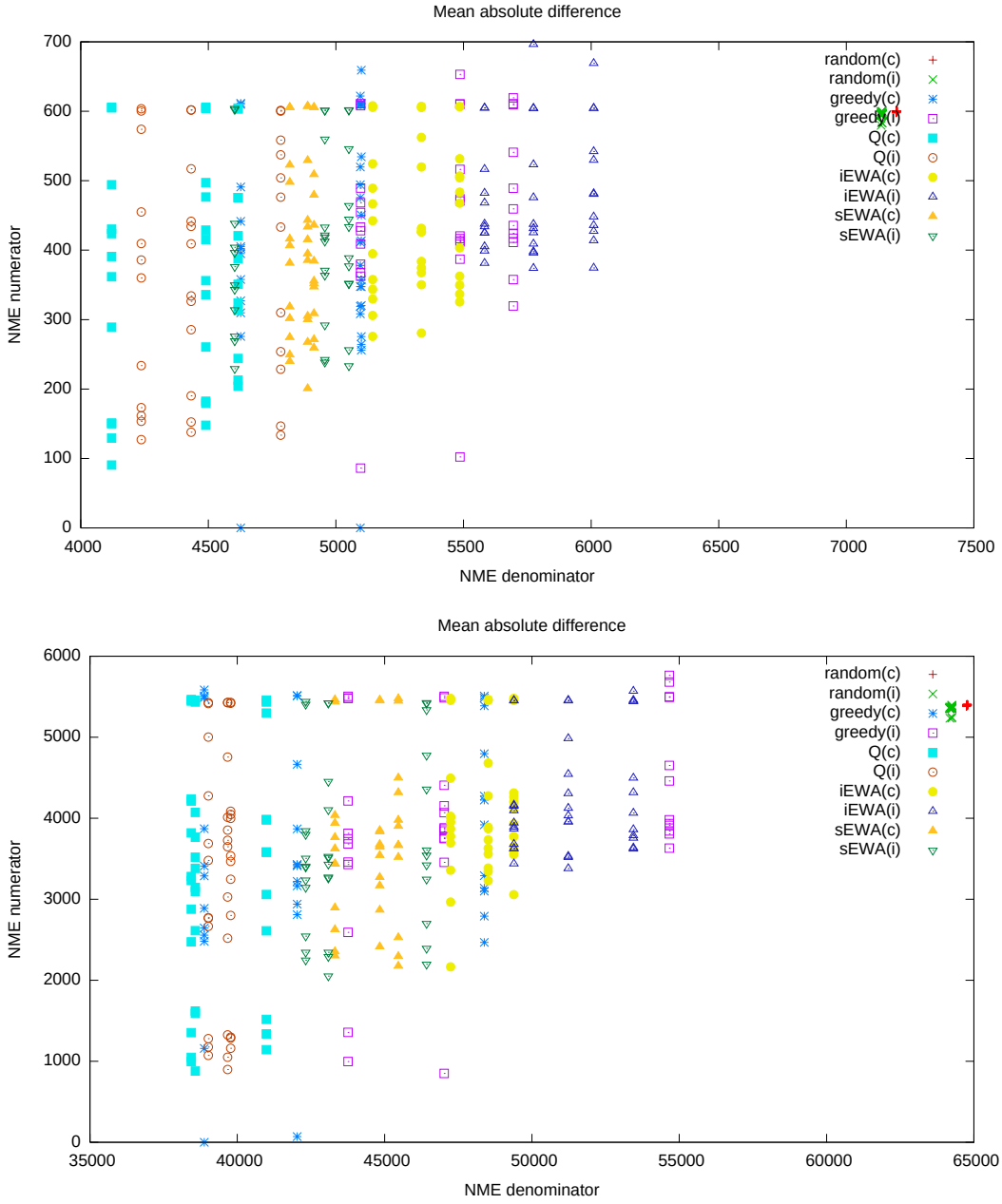


Figure 5.5: The numerator of the NME score plotted against the denominator for each model class for homogeneous agents (top) and heterogeneous agents (bottom).

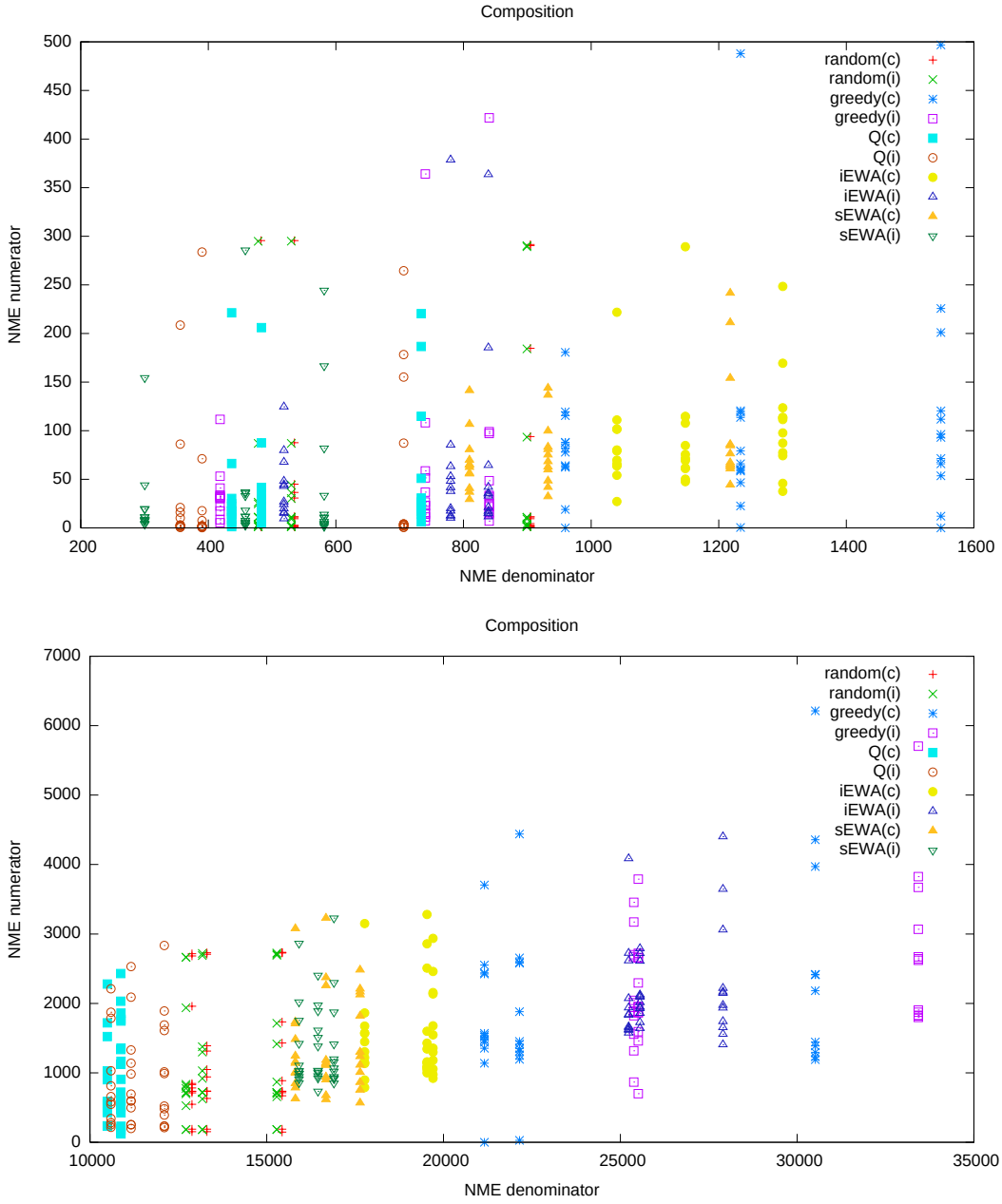


Figure 5.6: The numerator of the NME score plotted against the denominator for each model class for homogeneous agents (top) and heterogeneous agents (bottom).

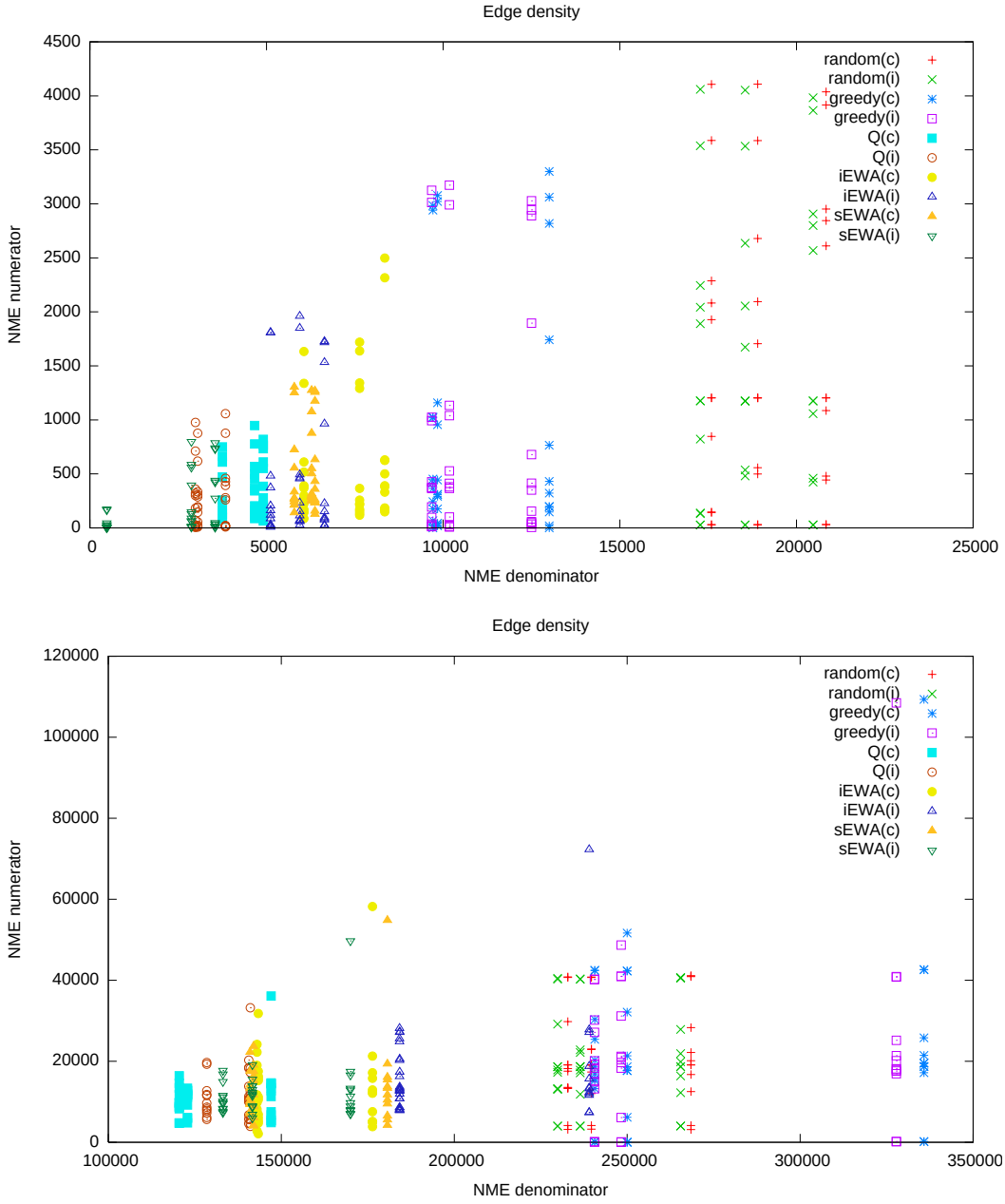


Figure 5.7: The numerator of the NME score plotted against the denominator for each model class for homogeneous agents (top) and heterogeneous agents (bottom).

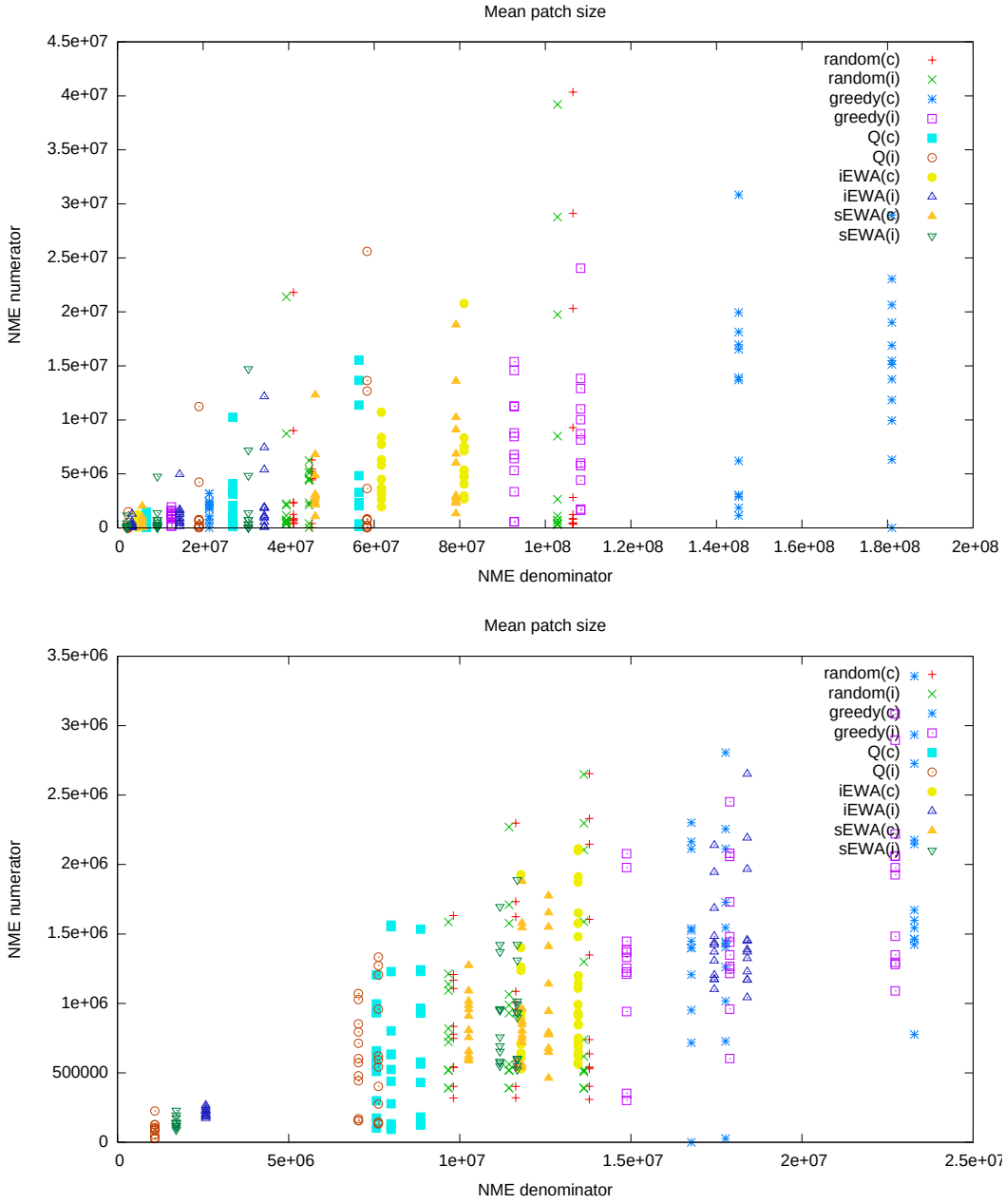


Figure 5.8: The numerator of the NME score plotted against the denominator for each model class for heterogenous agents (top) and heterogeneous agents (bottom).

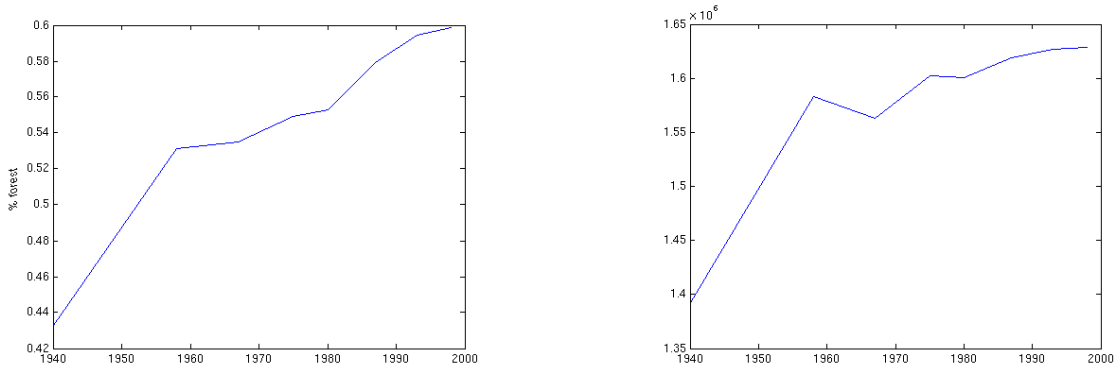


Figure 5.9: Quantitative changes in composition (left) and forest edge length (right) in Indian Creek township from 1940 to 1993.

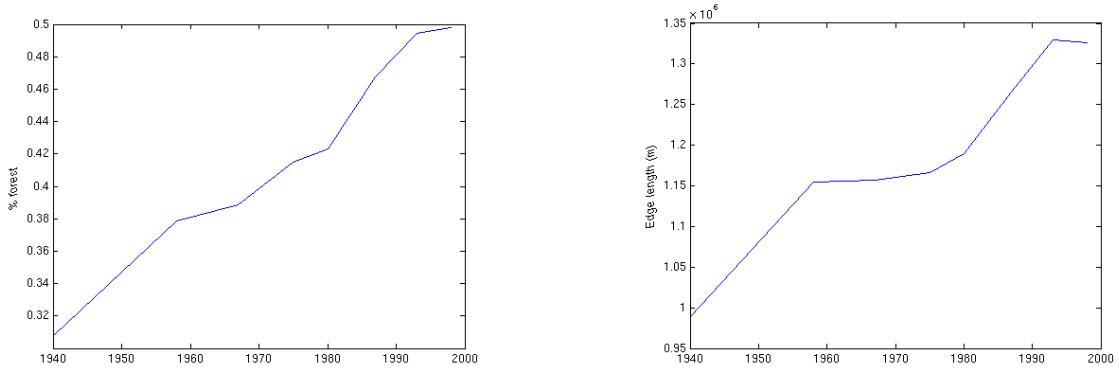


Figure 5.10: Quantitative changes in composition (left) and forest edge length (right) in Van Buren township from 1940 to 1993.

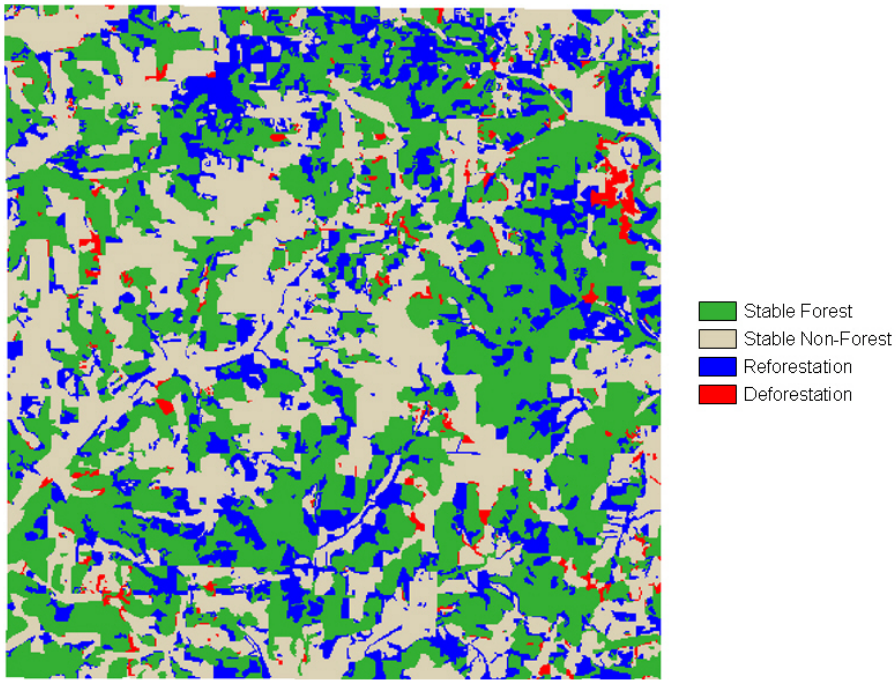


Figure 5.11: Deforestation, afforestation and stable forest cover in Indian Creek from 1940 to 1993.

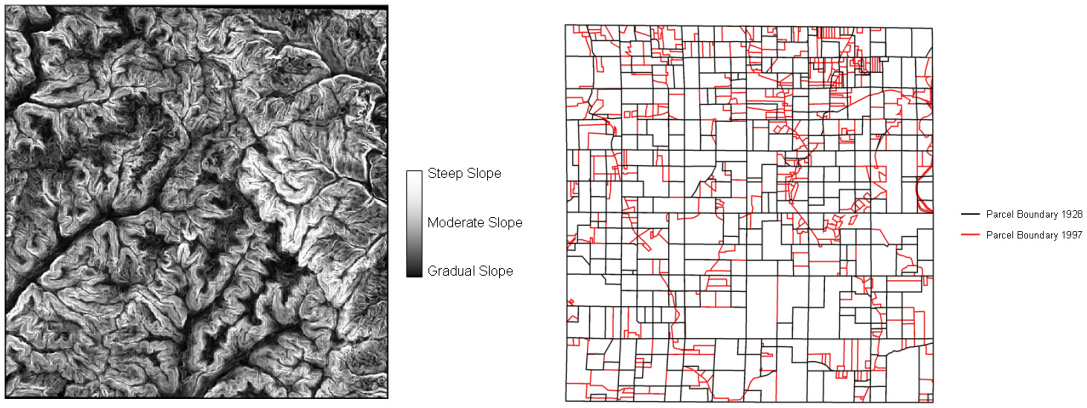


Figure 5.12: Indian Creek slope steepness (left) and parcel borders (right) in 1928 (red line) and 1997 (black line).

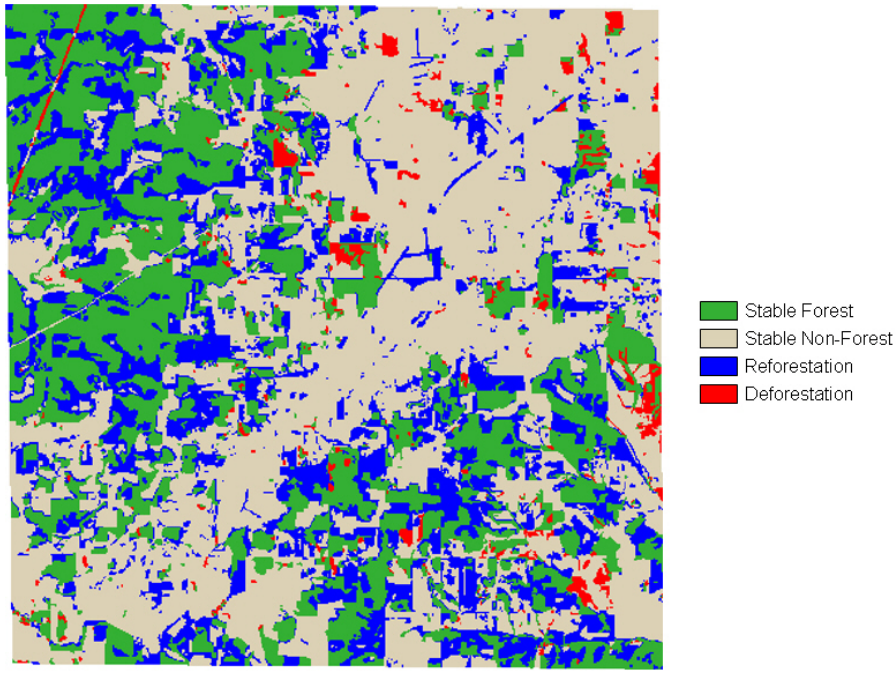


Figure 5.13: Deforestation, afforestation and stable forest cover in Van Buren from 1940 to 1993.

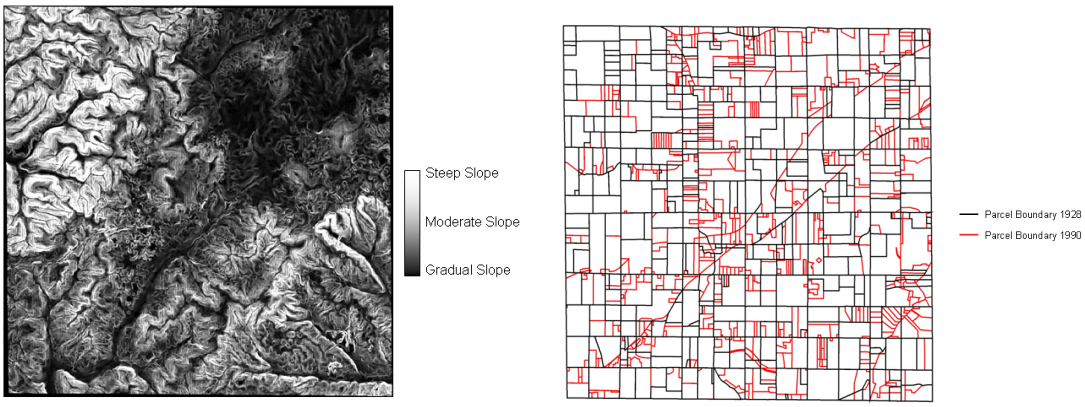


Figure 5.14: Van Buren slope steepness (left) and parcel borders (right) in 1928 (red line) and 1997 (black line).

Figure 5.15: NME numerator vs. denominator with mean absolute difference for heterogeneous agents.

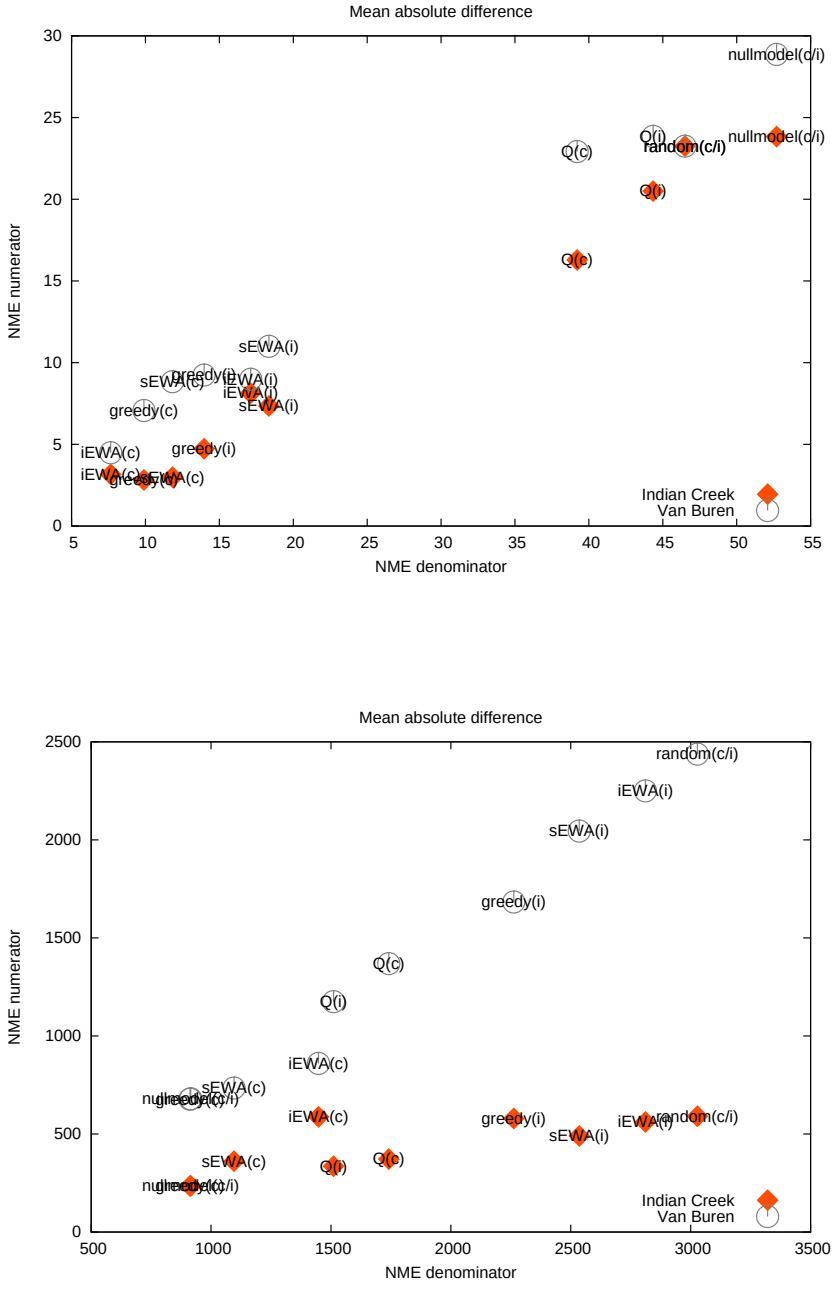
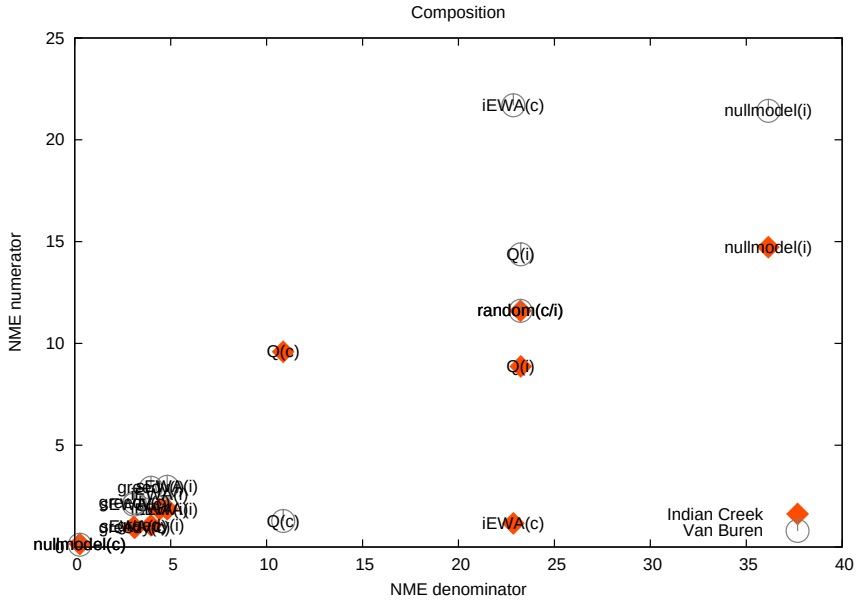


Figure 5.16: NME numerator vs. denominator with composition for homogeneous agents (top) and heterogeneous agents (bottom).



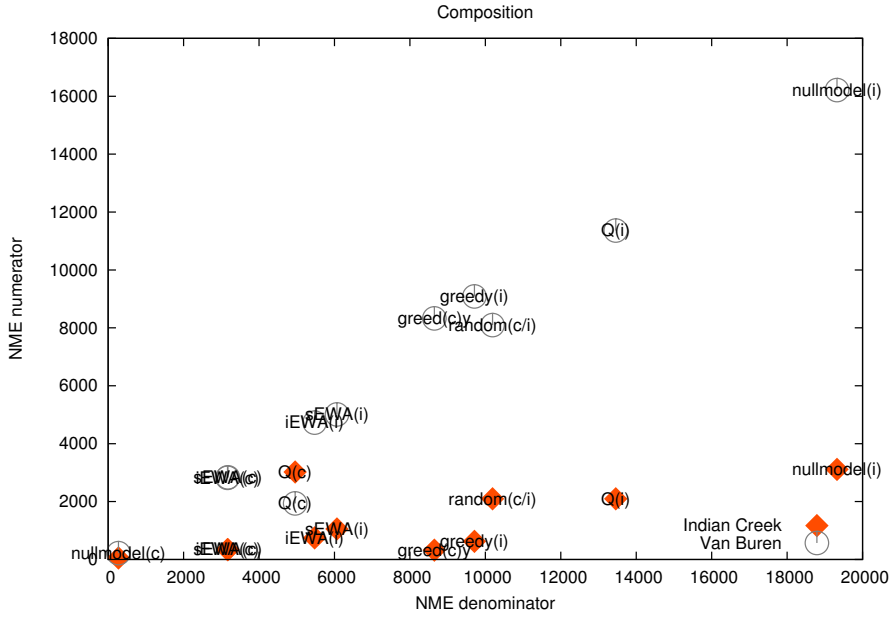


Figure 5.17: NME numerator vs. denominator with edge density for homogeneous agents (top) and heterogeneous agents (bottom).

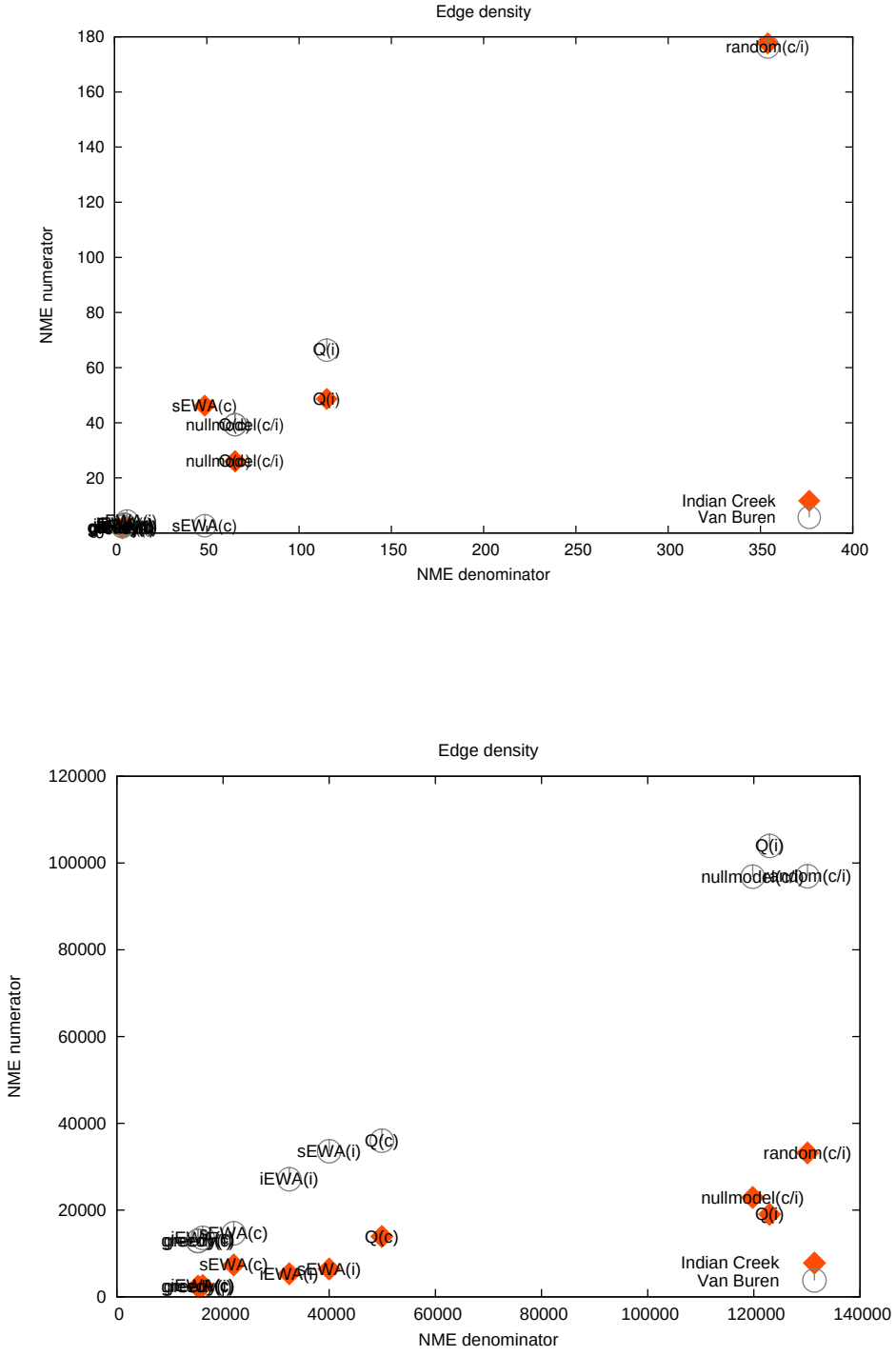
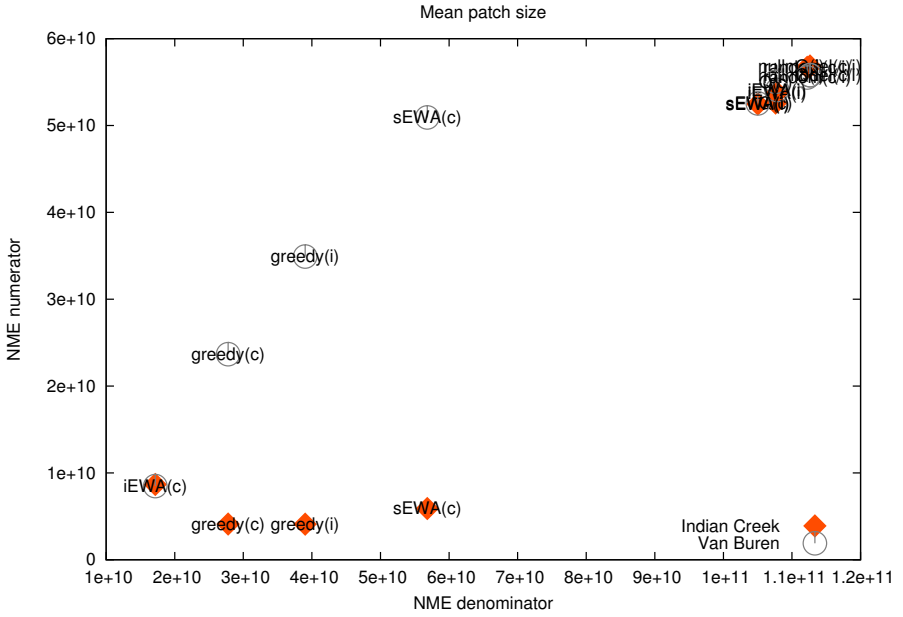
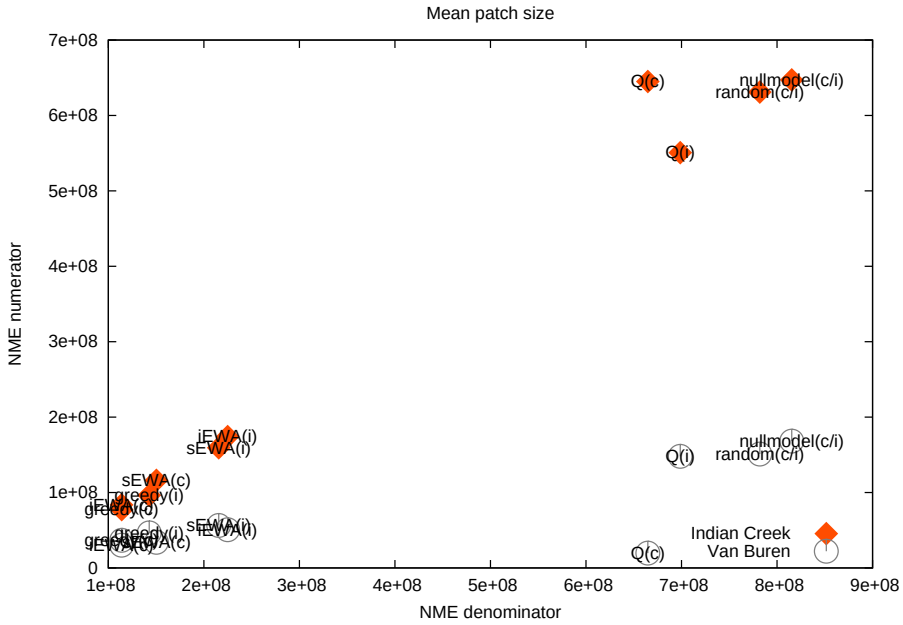


Figure 5.18: NME numerator vs. denominator with mean patch size for homogeneous agents (top) and heterogeneous agents (bottom).





6

Discussion and Future Work

Land-cover is not only a considerable factor in determining global climate, it is also important for the well being of terrestrial and aquatic species. These days Earth's land-cover is going through changes at a faster pace than ever, and most of these changes are human initiated; there is practically no area on Earth that has not been or is not being directly or indirectly altered by human land use. Besides the Ice Age and some other hypothesized cataclysms (e.g., one that supposedly led to the extinction of dinosaurs) it is hard to envision changes with an impact of equal extent, nor conceive types of contemporary changes that would be favorable for biodiversity in the long run.

Pervasive land-use and consequent land-cover changes have had and continually have adverse impact on local, regional and global level by destroying natural ecosystems and causing irreversible changes in global climate. In order to slow down the effect of destructive practices, factors driving land-use decisions need to be examined and understood. Human population growth directly burdens the natural environment. However, in order to understand the full extent of the impact land-use change has on ecological systems, and also to educate land-owners

about sustainable land-management practices, the socio-economical, political, psychological, and demographic forces driving the change need to be explained.

Computer modeling is gaining popularity among scientists working in fields in which manipulative experimentation is not an option, such as studying land-use and land-cover change. Combined with other methods, for instance household surveys and analysis of census data, computer models offer a relatively effortless method for testing implications of policies, predicting land-cover changes and exploring interactions between, for instance, socio-economic and bio-ecological factors in land-use change.

Since computer models are often used to inform real-world decision making, it is important that these models are based on sound principles, and their plausibility and adequacy to the task is rigorously assessed; i.e., it is pivotal to have a right model to the task. Consequently, the evaluation, validation and selection methods are as crucial as the models themselves.

This dissertation proposes a criterion for selecting between agent-based models of land-use and land-cover change, which is based on algorithmic coding theory. Artificial and real-world data is used to demonstrate that the criterion behaves sensibly under multiple experimental manipulations of the error function, exogenous forces and the set of candidate model classes.

Models of land-use and land-cover change are often validated against real landscape data as opposed to data gathered about human decisions. Different spatial metrics — measures to characterize physical landscapes — are used in the validation; some of them quantify composition whereas some of them concentrate on land-cover patterns. Different metrics supposedly depict different kinds of regularities in the landscape surface. Consequently, a model selection criterion, presumed to identify models' disposition to detect regularities in the data, is assumed

to be sensitive to the metrics. This is the case also with the proposed criterion and the model classes used in this study. The criterion is sensitive to variations of spatial metrics as well as to exogenous factors, such as landscape suitability and agent heterogeneity, but does not break down; it consistently prefers simpler model classes over more flexible ones, and even with the manipulated candidate model classes, it makes consistent selections.

6.1 Contributions

The specific contributions of this research are discussed one by one.

Question 1. *What are good measures to be used to distinguish the performance of different adaptive spatially explicit agent-based models?*

There is no straightforward answer to this question. Given the relatively complex class of land-use and land-cover change models, and potentially a large set of interacting forces driving the change, a safe solution is to use a battery of spatial metrics that characterize different aspects of landscape composition and configuration.

Question 2. *What is an appropriate selection criterion to choose a model that best explains the available data?*

This study offers a practical criterion, called *Normalized Minimum Error (NME) principle* for selecting among complex adaptive systems that do not lend themselves to the usage of traditional methods. The only prerequisites to the usage of the proposed method are an error function and a minimization algorithm.

The proposed selection criterion is based on *Minimum Description Length (MDL)* principle, which is a general method of doing inductive inference. The method

does not assume that a 'true model' exists, but selects a model class with which the most regularities can be extracted from data. The proposed criterion is based on a simple idea that non-probabilistic model classes can be interpreted as probabilistic ones by replacing the probabilities as measures of fit with errors, and associating the errors with code lengths.

The criterion's relation to another formulation of MDL principle is demonstrated. This principle is called *Normalized Maximum Likelihood (NML)*, and it is based on using probabilities as measures of fit. The NML criterion has been proved optimal in the sense that it defines a unique model that minimizes the discrepancy, measured in the description length, between the selected and the best-fitting model. However, as opposed to the NME criterion, the NML criterion cannot be applied to many practically interesting model classes. In general, the results of this study supply evidence for the MDL principle's, and particularly the NME criterion's applicability in relatively complex real-world domains, such as the model class of agent-based land-use and land-cover models.

The proposed criterion is compared to two other methods: first to a modified NME, called ERR criterion, that only uses the numerator, i.e., the error values, as selection criterion, and secondly to leave-one-out cross validation. While the NME criterion clearly outperforms the ERR criterion by consistently selecting simpler model classes in most experimental conditions, the CV technique proves impossible to use in practice; there are seldom enough data samples available to carry out the cross-validation process, and even with a small number of samples (five was used in the current study) the calibration takes a huge amount of time to run.

The framework of TRAPP^2 class, in which the selection criterion is embedded, also provides an extensible platform for studying the behavior of models belonging to this class of complex adaptive systems.

Question 3. *How does the choice of the performance measure influence the behavior of the model selection criterion?*

The main contribution of this research is the thorough analysis of the selection criterion's sensitivity, or on the other hand, its propensity to brake down or to be inconsistent with a manipulation of variables external to the criterion and the candidate model classes.

The criterion is tested in multiple experimental settings by varying (i) the spatial metrics used in landscape comparison, (ii) initial conditions that determine agent and landscape heterogeneity, and (iii) generating and candidate model sets. The results of these analyses confirm the stability of the criterion, and increase confidence on its capability of behaving sensibly with real-world data.

6.2 Caveats

The proposed model selection criterion cannot be analyzed in isolation without regarding the error function it uses. The current study uses two error functions and four spatial metrics. Three of these metrics — composition, edge density and mean patch size — are so called summary statistics; they characterize a single aspect of the land-cover, whereas the fourth one, mean absolute difference, calculates a location by location difference between land-covers of two landscapes. This metric uses more information of the landscapes than the other three that do not consider

location.

Summary statistics are supposedly easier to fit, since there are several possible ways to get them right, e.g., several different land-cover configurations may have the same composition. Consequently, there are fewer ways of getting them wrong, too. However, there are very few, actually only one way of getting the location-by-location comparison correct, and a considerable number of ways of getting it wrong.

The literature provides us with evidence that, somewhat counterintuitively, location-by-location comparison is not that difficult after all. Pontius *et al.* (2004) argue that not a single model has been reported that is able to predict the location of land-cover changes better than a null model, a model that predicts no change. Another surprising issue of this observation is that the most trivial model class of all fares so well. Although Pontius *et al.* do not argue that the null model necessarily is the best model to explain land-cover changes in general, in certain respect it is; it is computationally simple, and it does not have a single free parameter.

This is exactly what the proposed selection criterion is looking for; a model class that is simple and contains a model that fits the data well. Since the changes over time in the real landscapes are usually small, a model that predicts few changes should perform well. Why does not the NME criterion then select the null model more often?

In Experiments II and III 'all possible data' was replaced by 'all available data' for practical reasons. While the null model fitted well the two Indiana data sets and to a considerable number of artificial data sets, it is improbable that it would fit well 'all possible data sets', especially those that undergo significant number of changes over time. With 'all possible data' its flexibility score would end up lower (i.e., the denominator of the NME criterion higher because of larger errors)

allowing it to be selected more often.

Thus, the criterion's tendency not to select the null model class is a by-product of how the experiments are carried out, not the criterion itself. For Experiment II, with a reasonable number of generated data samples, this deviation from the theoretical framework is likely not detrimental, but certainly it distorts the results of Experiment III.

Of course one can use 'all possible data' only in some restricted cases when data size and the number of permutations are small. In other cases one needs to come up with a different approach; either (i) to use 'all available data' and bare the consequences, or (ii) to generate 'all possible data' or at least a representative subset of them.

In the former case 'all possible data' translates into 'all plausible data', but there is an unwanted consequence, which can be demonstrated with Experiment III. If only the available data is used, the NME criterion never selects a model class that fits equally well multiple data samples even if their underlying process can be assumed to be the same. For instance, even if both Indiana landscapes exhibit some idiosyncrasies, they can be assumed to be generated by 'the same process'; they are physically linked, subject to the same weather conditions and under the same county rules, just to name few common factors. However, the NME criterion penalizes a model class that fits well both of these data samples, as if it fitted all data well.

However, a scientist working on a problem sometimes has more than one data set available and wants to find a single mechanism that explains them all. Not choosing the same model class for these data samples implies that the NME criterion penalizes for generalizability rather than flexibility. This certainly is not what we want. However, we should keep in mind that the criterion only penalizes for

excess flexibility when the denominator enumerates 'all possible data'. A model class that fits well all of them, definitely is overly flexible. On the other hand, a model class that fits well a small number of samples and produces large errors with the remaining ones potentially generalizes well. The proposed criterion fares well in both of these cases when used properly.

6.3 Directions for Future Work

This research represents preliminary stages in studying of model selection methods for complex adaptive systems of which the class of agent-based land-use and land-cover change is an example. Several future directions can be envisioned regarding both the modeling framework and the model selection criterion.

First, the current framework was deliberately kept simple and abstract to highlight the properties of the selection criterion. The most obvious extensions to the framework are, first, to allow the emergence of a larger set of land-uses, and secondly, to make the actual physical land-cover distinct from the agent land-use. Regarding the domain of land-cover change the assumption of two alternative land-uses which coincide with the resulting cover is somewhat an oversimplification.

Secondly, relatively straightforward although commonly used spatial metrics were chosen for the current study, partly because of almost exclusive usage of artificial data. Real landscapes are assumed to exhibit regularities that may not be captured by the current metrics. Consequently, more intricate spatial metrics may be appropriate. Moreover, the error function may need to be adjusted so that it better applies to cases with real-world data that is not always available on a fine temporal scale, for instance yearly. The effect of calculating the errors every other,

third, or tenth round, a feature that is already implemented in the framework, was not studied in this dissertation.

The learning strategies chosen for this study represent quite general reinforcement learning based strategies familiar from economics and psychology. The context and the choice options involved in land-use decision-making differ significantly from the situation a decision-maker in a psychological laboratory experiment encounters. Therefore, more specialized decision strategies may be called for in future studies of model selection among agent-based models of LUCC.

In the current study the scarcity of landscape change data was overcome by utilizing artificially generated data. Another direction to explore and test the selection criterion's performance is to use human decision data collected in laboratory experiments. Evans *et al.* (in press) designed and implemented a computerized platform for studying human land-use decision making in an abstract spatial context. The pioneering experiments that have been carried out provide real decision data that can be used to compare different decision strategies and to select between them. This is a natural extension to explore the proposed criterion's behavior.

In order to fully understand the relation between theoretical underpinnings of the proposed criterion and the underlying practical issues discussed above, the actual meaning of 'all possible data' needs to be explicated. In reality 'all possible data' seldom means 'all possible permutations' of the data, but the actual outcomes are restricted in myriad ways. For instance, it is hard to conceive that a real landscape would undergo sporadic changes between any two time points. On the other hand, the generation of all possible data sets becomes practically infeasible even for relatively small landscapes if multiple land-uses and land-covers are allowed. The study of how to constrain possible and plausible data sets offers another potential for future research.

Moreover, in order to comprehend the full potential of the criterion and its possible shortcomings, it needs to be compared to other applicable model selection methods. The work in this direction is under way, and preliminary results are reported in this dissertation.

The proposed model selection criterion's feasibility was demonstrated by showing its close resemblance to the normalized maximum likelihood (NML) principle. Particularly, the likeness between the 'average' version of the NME criterion and the 'average' version of the NML criterion was addressed. A subject for future studies is to extend the NME criterion to have this 'average' interpretation, and test experimentally its adequacy and the deviation from the original formulation.

Other methods of using errors in model selection have been introduced. The *Prequential* method, proposed by Dawid (1984), uses series of errors made so far to predict the occurrence of next instance in a sequence. Grünwald (1998, 1999) has presented a method called *entropification* that can be used to associate non-probabilistic model classes with probabilistic ones. Lee (2006) applies this method for psychological models with 0/1 loss function. Using the same observation as is used in the current study, that the loss function can only take a finite number of values, he derives a practically computable formulation for the probability distribution given by entropification. The very same insight could be easily used to apply the entropification to the class of agent-based models of LUCC. An empirical comparison of the entropification method and the proposed NME criterion offers another interesting direction for future research.

Bibliography

Agarwal, C., Green, G. M., Grove, J. M., Evans, T. P., & Schweik, C. M. (2002). *A review and assessment of land-use change models: Dynamics of space, time and human choice* (Tech. Rep. No. NE-297). Department of Agriculture, Forest Service, Northeastern Research Station.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrox & F. Caski (Eds.), *Second International Symposium on Information Theory* (p. 267-281). Akademiai Kiado, Budapest, Hungary.

Andras, P., Roberts, G., & Lazarus, J. (2003). Environmental risk, cooperation, and communication complexity. In E. Alonso, D. Kudenko, & D. Kazakov (Eds.), *Adaptive agents and multi-agent systems* (Vol. 2636, p. 49-65). Springer-Verlag.

Avery, T. E., & Burkhart, H. E. (2002). *Forest measurements.* Boston, MA: McGraw-Hill.

Axelrod, R. (1984). *Evolution of cooperation.* New York: Basic Books.

Baker, W. L. (1989). A review of models of landscape change. *Landscape Ecology*, *2*(2), 111-133.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle and coding and modeling. *IEEE Transaction on Information Theory*, *44*(6), 2743-2760.

Bartlett, M., & Kazakov, D. (2004). The role of environmental structure in multi-agent simulations of language evolution. In *Proceedings of the Fourth Symposium on Adaptive Agents and Multi-Agent Systems (AAMAS-4), AISB convention.* Leeds, UK.

Bayarri, M. J., Berger, J. O., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., & Tu, J. (2002). A framework for validation of computer models. In *Proceedings of the Workshop of Foundations for V&V in the 21th Century.* Society for Modeling and Simulation International.

Berger, T. (2001). Agent-based spatial models applied to agriculture: A simulation tool for technology diffusion, resource use changes and policy analysis. *Agricultural Economics*, *25*, 245-260.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. Launer & G. Wilkinson (Eds.), *Robustness in statistics.* New York: Academic Press.

Boyce, M. S. (2002). Statistics as viewed by biologists. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(3), 306-312.

Bradbury, R. (2002). Futures, predictions and other foolishness. In M. A. Janssen (Ed.), *Complexity and ecosystem management: The theory and practice of multi-agent systems.* Edward Elgar, Cheltenham, U.K.

Brown, D. G., Page, S., Riolo, R., Zellner, M., & Rand, W. (2005). Path dependence and the validation of agent-based spatial models of land use. *International Journal of Geographical Information Science*, *19*(2), 153-174.

Brown, D. G., Pijanowski, B. C., & Duh, J.-D. (2000). Modeling the relationship between land use and land cover on private lands in the Upper Midwest, USA. *Journal of Environmental Management*, *59*, 247-263.

Brown, D. G., Riolo, R., Robinson, D. T., North, M., & Rand, W. (2005). Spatial process and data models: Toward integration of agent-based models and GIS. *Journal of Geographical Systems*, *7*, 25-47.

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.

Busemeyer, J. R., & Myung, I. J. (1987). Resource allocation decision making in an uncertain environment. *Acta Psychologica*(66), 1-19.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171-189.

Camerer, C., & Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827-874.

Carpenter, J. P., Harrison, G. W., & List, J. A. (Eds.). (2005). *Field experiments in economics* (Vol. 10). Elsevier.

Casti, J. L. (1997). Can you trust it? On the reliability of computer simulation and the validity of models. *Complexity*, *2*(5), 8-11.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science*, *148*, 754-759.

Chater, N. (2005). A minimum description length principle for perception. In P. D. Grnwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications.* Bradfor Books.

Chater, N., & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19-22.

Cioffi-Revilla, C., & Gotts, N. M. (2003). Comparative analysis of agent-based social simulations: Geosim and FEARLUS models. *Journal of Artificial Societies and Social Simulation*, *6*(4).

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: Wiley.

Dawid, P. (1984). Statistical theory: The prequential approach. *Journal of Royal Statistical Society A*, *147*, 278-292.

de Rooij, S., & Grünwald, P. (2006). An empirical study of minimum description length model selection with infinite parametric complexity. *Journal of Mathematical Psychology*, *50*, 180-192.

Deadman, P., & Gimblett, R. H. (1994). The role of goal-oriented autonomous agents in modeling people-environment interactions in forest recreation. *Mathematical and Computer Modeling*, *20*(8).

Deadman, P., Robinson, D., Moran, E., & Brondizio, E. (2004). Colonist household decisionmaking and land-use change in the Amazon rainforest: An agent-based simulation. *Environment and Planning: Planning and Design*, *31*, 693-709.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, *7*, 509-520.

Evans, T., Sun, W., & Kelley, H. (in press). Spatially explicit experiments for the exploration of land-use decision-making dynamics. *International Journal of Geographical Information Science*.

Evans, T. P., & Kelley, H. (2004). Multi-scale analysis of a household level agent-based model of landcover change. *Journal of Environmental Management*, *72*, 57-72.

Fisher, R. A. (1922). On the interpretation of x^4 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, *85*(1), 87-94.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*, 205-231.

Gingrich, S. F. (1967). Measuring and evaluating stocking and stand density in upland hardwood forests in the central states. *Forest Science*, *13*, 38-53.

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, *44*, 153-170.

Grimm, V. (1999). Ten years of individual-based modelling in ecology: What have we learned and what could we learn in the future? *Ecological Modelling*, *115*, 129-148.

Grünwald, P. (1998). *The minimum description length principle and reasoning under uncertainty.* Unpublished doctoral dissertation, University of Amsterdam.

Grünwald, P. (1999). Viewing all models as 'probabilistic'. In *Twelfth Annual Conference on Computational Learning Theory (COLT '99)* (p. 171-182). New York, NY, USA: ACM Press.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133-152.

Grünwald, P. (2005). Minimum description length principle tutorial. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications.* Bradfor Books.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of American Statistical Association*, *96*(454), 746-774.

Hoffman, M., Kelley, H., & Evans, T. (2002). Simulating land cover change in South-central Indiana. In M. E. Janssen (Ed.), *Complexity and Ecosystem Management: The theory and practice of multi-agent systems.* Edward Elgar, Cheltenham, U.K.

Hoover, W. L., & Preston, G. (2004). *2004 Indiana forest products price report and trend analysis* (Tech. Rep. No. FNR-177W). Lafayette, IN: Purdue University Department of Forestery and Natural Resources.

Huigen, M. G. A. (2004). First principles of the MameLuke multi-actor modelling framework for land-use change, illustrated with a Philippine case study. *Journal of Environmental Management*, *72*, 5-12.

Irwin, E. G., & Bockstael, N. E. (2002). Interacting agents, spatial externalities and the evolution of residential land use patterns. *Journal of Economic Geography*, *2*, 31-54.

Itami, R., & Gimblett, H. (2001). Intelligent recreation agents in a virtual GIS world. *Complexity International Journal*, *08*.

Jager, W., Janssen, M., Vries, H. D., Greef, J. D., & Vlek, C. (2000). Behaviour in commons dilemmas: *Homo economicus* and *homo psychologicus* in an ecological-economic model. *Ecological Economics*, *35*, 357-379.

Janssen, M. A. (2004). Agent-based modelling. In J. Proops & P. Safonov (Eds.), *Modelling in ecological economics.* Edward Elgar.

Janssen, M. E. (Ed.). (2002). *Complexity and ecosystem management: The theory and practice of multi-agent systems.* Edward Elgar, Cheltenham, U.K.

Jenerette, G. D., & Wu, J. (2001). Analysis and simulation of land-use change in the central Arizona - Phoenix region, USA. *Landscape Ecology*, *16*, 611-626.

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution*, *19*(2), 101-108.

Kaelbling, L. P., & Littman, M. L. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237-285.

Karasmaa, N. (2003). The spatial transferability of the Helsinki metropolitan area mode choice models. In *Selected Proceedings of the 9th World Conference on Transport Research (WCTR9).* Elsevier Science Ltd.

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.

Kelley, H., & Evans, T. (under review). The relative influences of land-owner and landscape heterogeneity in an agent-based model of land-use. *Journal of Economic Dynamics and Control*.

Koontz, T. M. (2001). Money talks — But to whom? Financial versus nonmonetary motivations in land use decisions. *Society and Natural Resources*, *14*, 51-65.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79-86.

Laine, T., & Busemeyer, J. (2004a). *Agent-based model of land-use decision making* [Proceedings of the NAACSOS (North American Association for Computational Social and Organizational Science) Conference]. `http://www.casos.cs.cmu.edu/events/conferences/2004/2004_proceedings/Laine_Tei.pdf`.

Laine, T., & Busemeyer, J. (2004b). 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). Lawrence Erlbaum Associates.

Laine, T., & Gasser, M. (2003). *How communicative pressure affects phonology: Modelling sound change in a population of communicators.* Presentation in the 8th International Cognitive Linguistics Conference. Logroño, Spain.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*(3), 555-580.

Lendasse, A., Simon, G., Wertz, V., & Verleysen, M. (2005). Fast bootstrap methodology for model selection. *Neurocomputing*, *64*, 161-181.

Lendasse, A., Wertz, V., & Verleysen, M. (2003). Model selection with cross-validation and bootstraps — application to time series prediction with RBFN models. In (p. 573-580). Berlin, Germany: Springer-Verlag.

Ligtenberg, A., Bregt, A. K., & van Lammeren, R. (2001). Multi-actor-based land use modelling: spatial planning using agents. *Landscape and Urban Planning*, *56*, 21-33.

Manson, S. M. (2000). Agent-based dynamic spatial simulation of land-use/cover change in the Yucatan peninsula, Mexico. In *4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs.* Banff, Alberta, Canada.

Manson, S. M. (2002). Validation and verification of multi-agent models. In M. A. Janssen (Ed.), *Complexity and ecosystem management: The theory and practice of multi-agent systems.* Edward Elgar, Cheltenham, U.K.

McClave, J. T., & Sincich, T. (2003). (Ninth Edition ed.). New Jersey: Prentice Hall.

Menczer, F., & Belew, R. K. (1996). From complex environments to complex behaviors. *Adaptive Behavior*, *4*(3-4).

Miles, P. D., & G. J. Brand, e. a. . (2001). *The forest inventory and analysis database description and users manual version 1.0* (Tech. Rep. No. GTR NC-218). St. Paul, MN: U.S. Dept. of Agriculture, Forest Service, North Central Research Station.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190-204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79-95.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167-179.

Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308-313.

Parker, D., Manson, S., Janssen, M., Hoffman, M., & Deadman, P. (2003). Multi-agent system models for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers*, *93*(2), 316-340.

Parker, D. C., & Meretsky, V. (2004). Measuring pattern outcomes in an agent-based model of edge-effect externalities using spatial metrics. *Agriculture Ecosystems and Environment*, *101*, 233-250.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421-425.

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.

Pontius, R. G., Huffaker, D., & Denman, K. (2004). Useful techniques of validation for spatially explicit land-change models. *Ecological Modelling*, *79*, 445-461.

Quadrat-Ullah, H. (2005). Structural validation of system dynamics and agent-based simulation models. In *Proceedings of the 19th European Conference on Modelling and Simulation (ECMS05).* Riga, Latvia.

Rand, W., Brown, D. G., Page, S. E., Riolo, R., Fernandez, L. E., & Zellner, M. (2003). Statistical validation of spatial patterns in agent-based models. In *Proceedings of Agent-based Simulation.* Montpellier, France.

Rieskamp, J., Busemeyer, J., & Laine, T. (2003). How do people learn to allocate resources? Comparing two learning models. *Experimental Psychology: Learning, Memory & Cognition*, *29*(6).

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465-471.

Rissanen, J. (1989). *Stochastic complexity in scientific inquiry.* World Scientific Publishing Co. Pte. Ltd.

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.

Salmon, T. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, *69*(6), 1597-1628.

Schelling, T. (1978). *Micromotives and macrobehavior.* New York: W.W. Norton & Co.

Schneider, L. C., & Pontius, R. G. (2001). Modeling land-use change in the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment*, *85*, 85-94.

Schwarz, G. (1978). Estimating the dimension of the model. *The Annals of Statistics*, *6*, 461-464.

Sillanpää, M. J., & Corander, J. (2002). Model choice in gene mapping: what and why. *Trends in Genetics*, *18*(6), 301-307.

Stephens, P. A., Buskirk, S. W., Hayward, G. D., & Rio, C. M. D. (2005). Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, *42*, 4-12.

Strong, D. R., Whipple, A. V., Child, A. L., & Dennis, N. (1999). Model selection for a subterranean trophic cascade root-feeding catepillars and entomopathogenic nematodes. *Ecology*, *80*(8), 2750-2761.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction (adaptive computation and machine learning).* MIT Press.

Tesfatsion, L. (2002). Agent-based computationl economics: Growing economies from the bottom up. *Artificial Life*, *8*(1), 55-81.

van Daalen, C. E., Dresen, L., & Janssen, M. A. (2002). The roles of computer models in the environmental policy life cycle. *Environment Science & Policy*(238), 1-11.

Vicsek, T. (2002). Complexity: The bigger picture. *Nature*, *418*(131).

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3/4), 279-292.

Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods and Research*, *33*(2), 167-187.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*, 41-61.

Curriculum Vitae

July 14, 1966	Born, Helsinki, Finland
1996	Master of Science, University of Helsinki
1995-1999	Instructor, Computer Science and Cognitive Science, University of Helsinki
1997-1998	Young Researcher's Fellowship, University of Helsinki Research Foundation
1999-2000	ASLA-Fulbright Study Grant
1999-2002	Associate Instructor, Computer Science, Indiana University
2002-2006	Research Assistant, CIPEC, Indiana, University

Publications

- Laine, T. (2005), A Selection Criterion for a Class of Agent-Based Spatial Decision Models. Presentation in the NAACSOS (North American Association for Computational Social and Organizational Science) Conference. Notre Dame, South Bend, IN, June 26-28.

- Laine, T. & Busemeyer, J. (2004), Comparing Agent-Based Learning Models of Land-use Decision Making. In *Proceedings of the Sixth International Conference on Cognitive Modelling, ICCM 2004* (Marsha Lovett, Christian Schunn, Christian Lebiere, Paul Munro, eds.), pp. 142-147.

- Laine, T. & Busemeyer, J. (2004), Agent-Based Model of Land-use Decision Making. In *the NAACSOS (North American Association for Computational Social*

and Organizational Science) Conference 2004 Proceedings, URL: `http://www.casos.cs.cmu.edu/events/conferences/2004/2004_proceedings/Laine_Tei.pdf`.

- Laine, T., Gasser, M. (2003), How Communicative Pressure Affects Phonology: Modelling Sound Change in a Population of Communicators. Talk at the 8th International Cognitive Linguistics Conference, Logrono, Spain, July 2003.

- Rieskamp, J., Busemeyer, J. & Laine, T. (2003), How do people learn to allocate resources? Comparing two learning models. *Journal of Experimental Psychology: Learning, Memory & Cognition*. Vol. 29, No.6.

- Laine, T., Kalakoski, V. (2001), Modelling Taxi Drivers' Learning and Exceptional Memory of Street Names. *Proceedings of the Fourth International Conference on Cognitive Modeling, ICCM-2001* (Erik M Altman, Axel Cleeremans, Christian D Schunn, Wayne D Gray, eds.), pp. 133-138.

- Saariluoma, P., Laine, T. (2001), Novice construction of chess memory. *Scandinavian Journal of Psychology*, Vol. 42, No. 2, pp. 137-146.

- Kurhila, J., Laine, T. (2000), Individualized special education with cognitive skill assessment. *British Journal of Educational Technology*, Vol. 11, No. 2, pp. 163-170.

- J.Kurhila, T.Laine (1999), Individualized special education with cognitive skill assessment. In *Proceedings of the 9th International PEG Conference, Intelligent Computer and Communication Technology: Teaching & Learning for the 21th Century* (Linda Baggot and Jon Nichol, eds.). University of Exeter, School of Education, England.

- Laine, T., Hyötyniemi, H., Saariluoma, P. (1998), The Foundations of Simulative Theorizing. *Proceedings of the 13th Biennial European Conference on Artificial Intelligence, ECAI'98* (Henri Prade, ed.), John Wiley & Sons, New York, pp. 109-113.

- Laine, T. (1998), Logic, Ontologies and Mental States - Report on the ECAI-98 Conference. *AI Communications 11(3-4)*, pp. 229-232.

- Saariluoma, P. Laine, T. (1998), Chess players' early recall of chess positions: An empirical and simulative investigation. *Proceedings of the Second European Conference on Cognitive Modelling, ECCM '98* (Frank E Ritter, Richard M Young, eds.), Nottingham University Press.

Outline of Studies

Major field: Computer Science Studies in Programming Languages, Intelligent Systems, Natural Language Processing

Major field: Cognitive Science Studies in Decision Making and Reasoning, Evolution and Learning, Philosophy of Computation

A

Results of Experiment II

A.1 Confusion matrices 1

The confusion matrices produced in the first phase of the experiments, using the NME criterion, the ERR criterion and the CV criterion, with all model classes are presented in this section for each $2 \times 4 \times 3$ experimental conditions: two agent types – homogeneous and heterogeneous, four spatial metrics and three landscape suitability conditions.

The confusion matrices for homogeneous and heterogeneous agents are shown in Figures A.1 and A.2, respectively, using the NME criterion, in Figures A.3 and A.4 using the ERR criterion, and in Figures A.5 and A.6 using the CV criterion. The rows of matrices in all figures represent the four spatial metrics and the columns represent the three suitability conditions. For instance, the first matrix on the second row is from the experiment with suitability condition I (homogeneous landscape) and spatial metric 2 (composition).

In each matrix the rows and columns, respectively, represent the generating and candidate classes in the following order: null, random, greedy, Q, iEWA, and sEWA — first the classes with collective parameter values, followed by the classes with individual values.

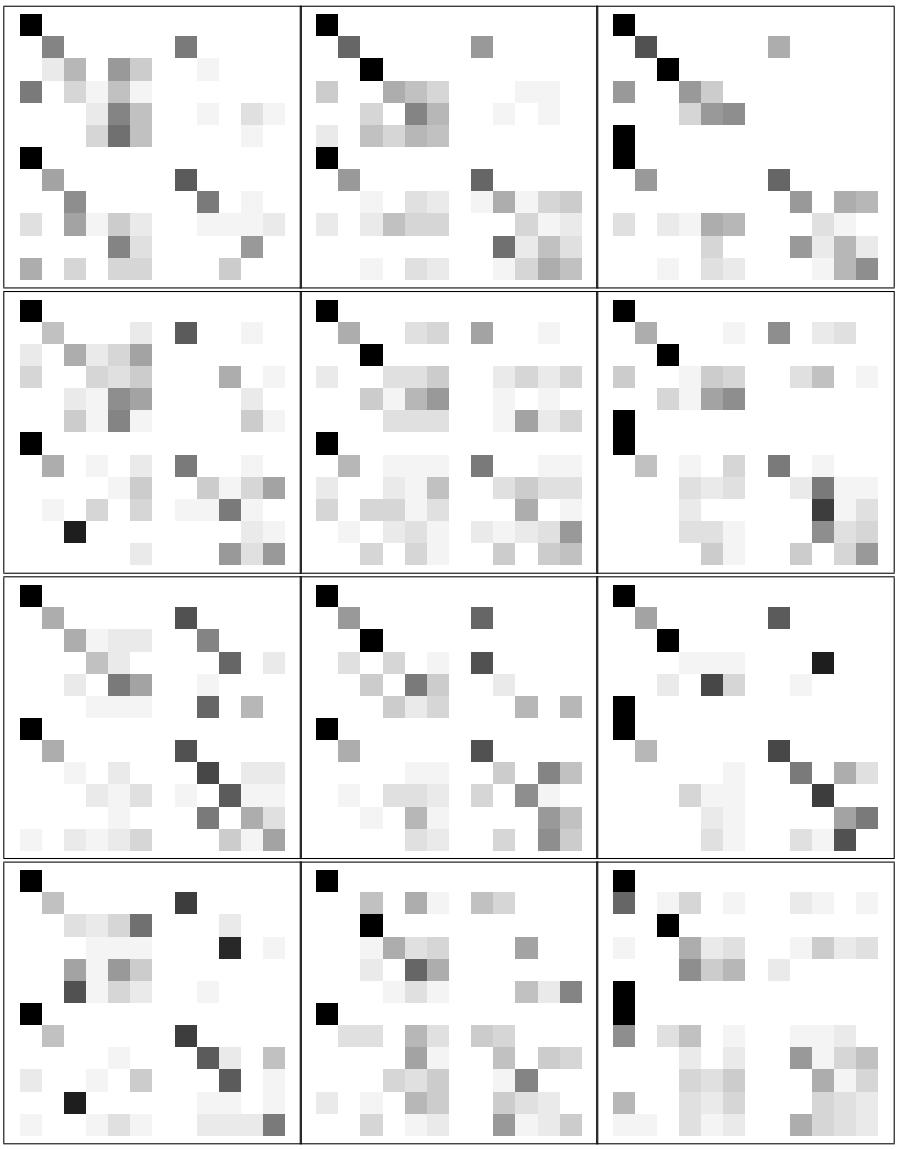


Figure A.1: Selection results for homogeneous agents using the NME criterion.

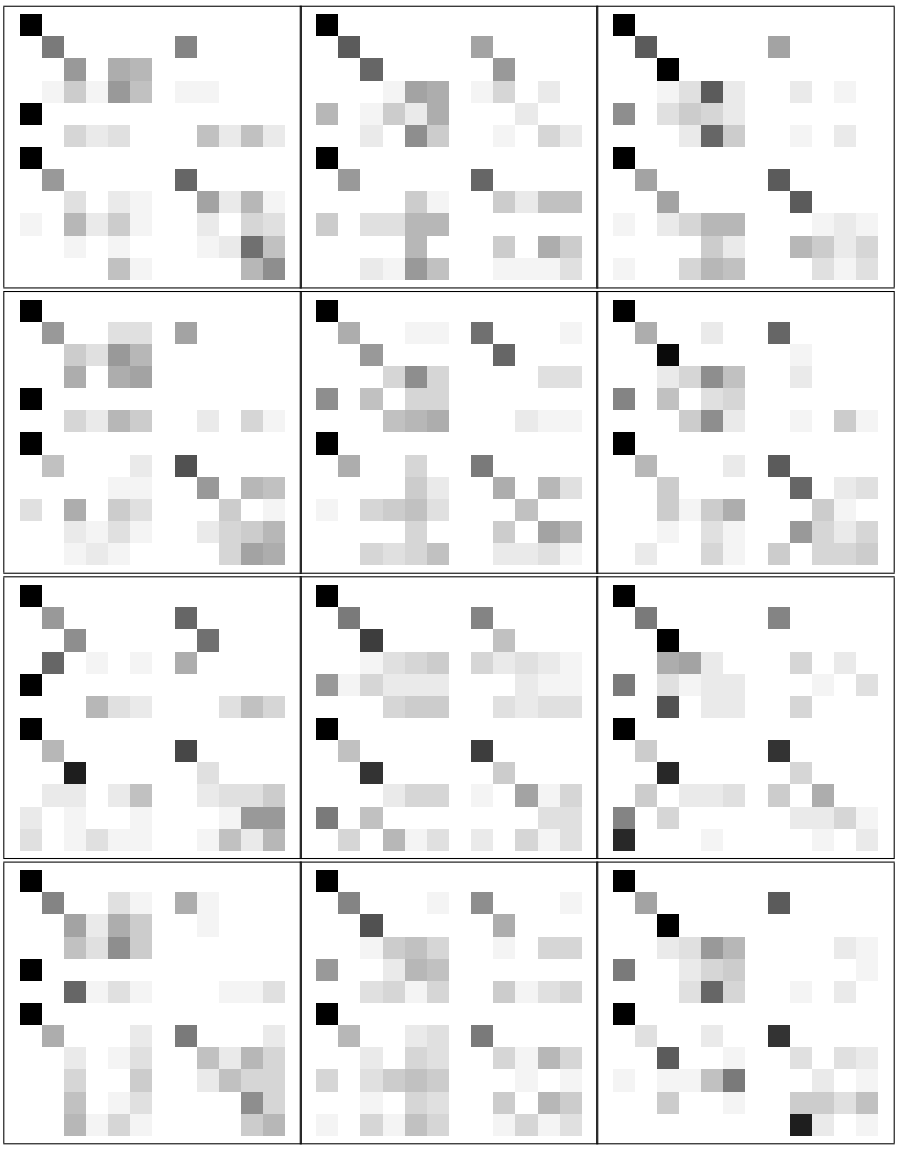


Figure A.2: Selection results for heterogeneous agents using the NME criterion.

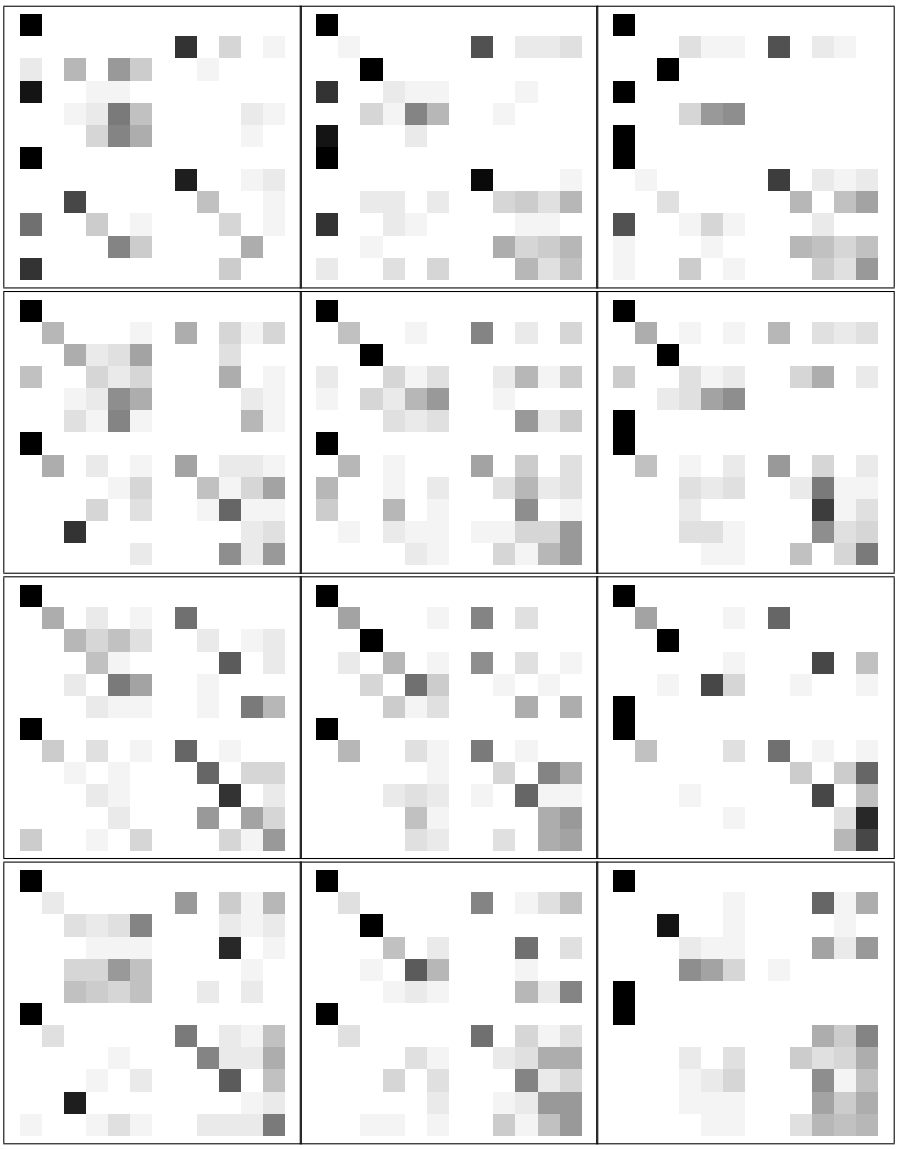


Figure A.3: Selection results for homogeneous agents using the ERR criterion.

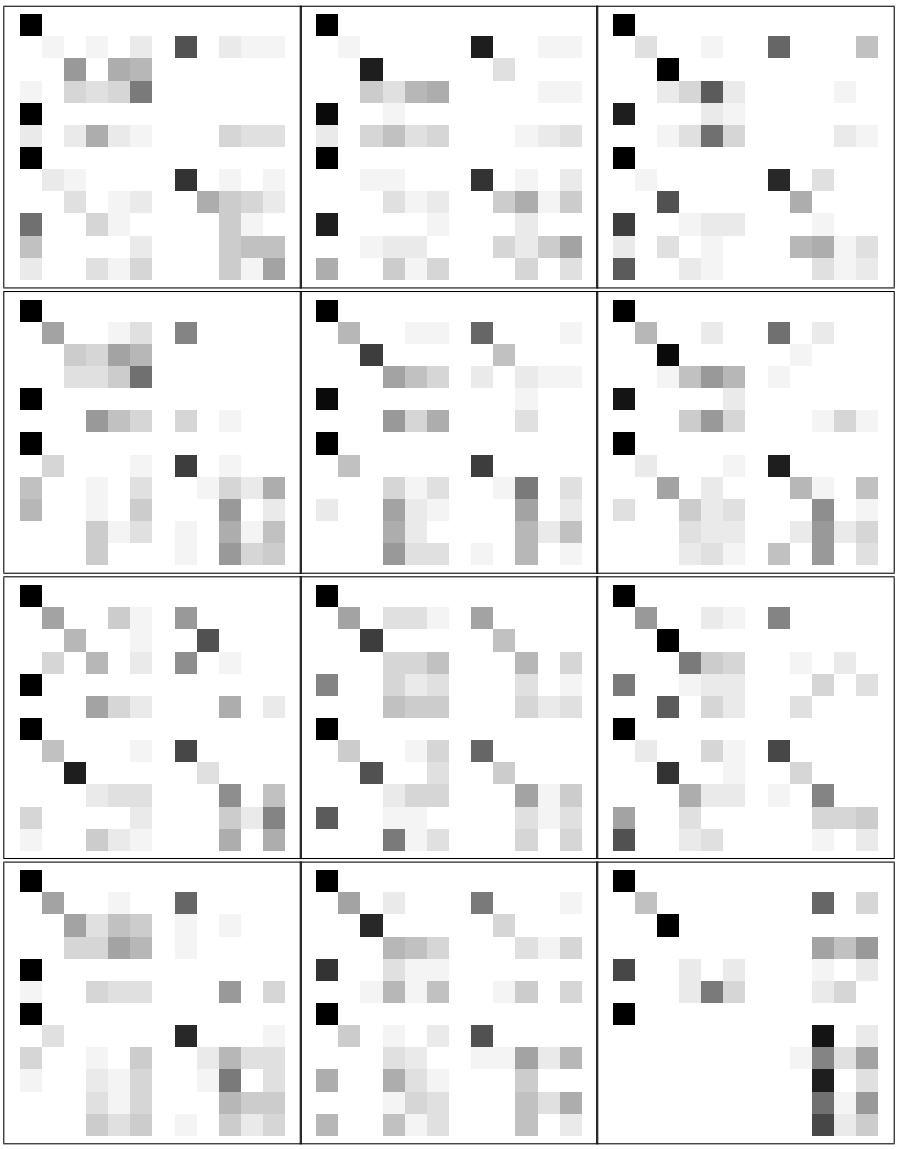


Figure A.4: Selection results for heterogeneous agents using the ERR criterion.

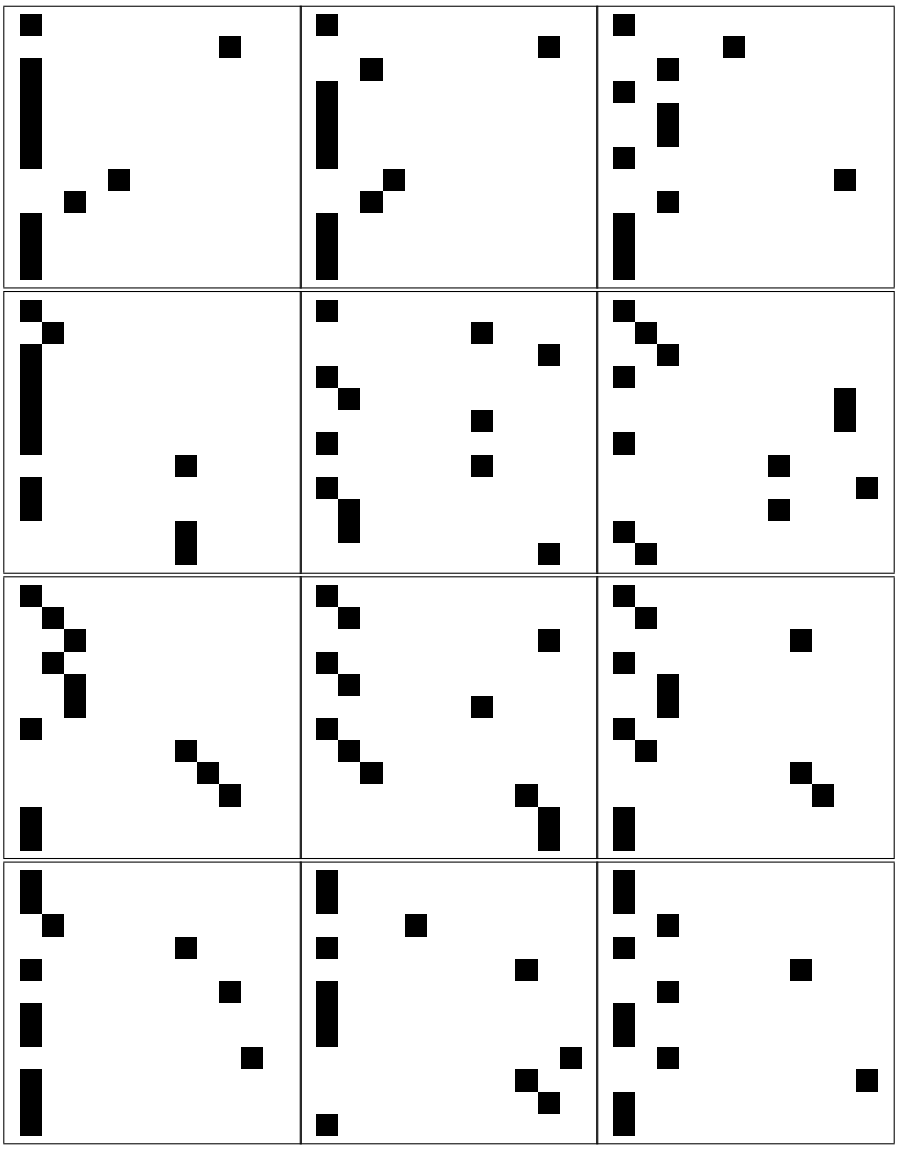


Figure A.5: Selection results for homogeneous agents using the CV criterion.

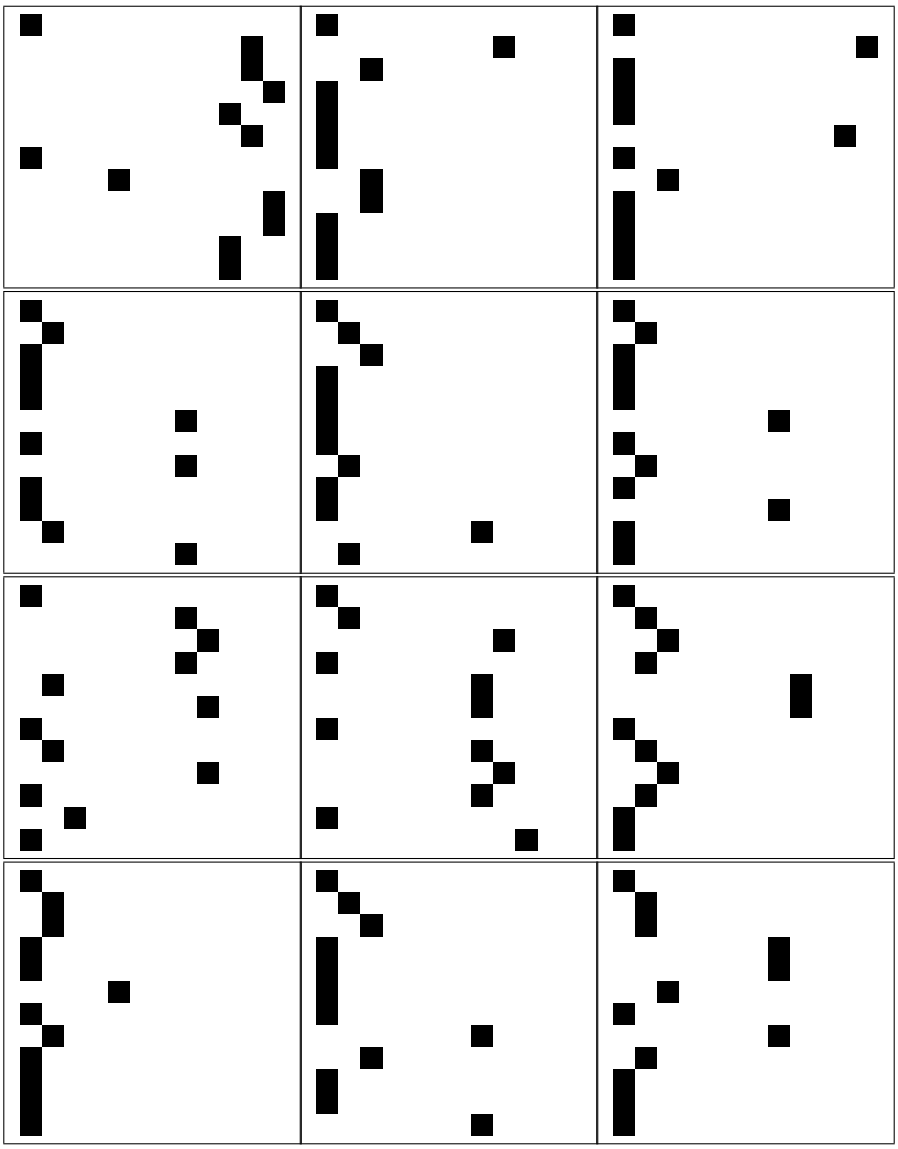


Figure A.6: Selection results for heterogeneous agents using the CV criterion.

Figure A.7: Homogeneous agents: generating classes are null and random.

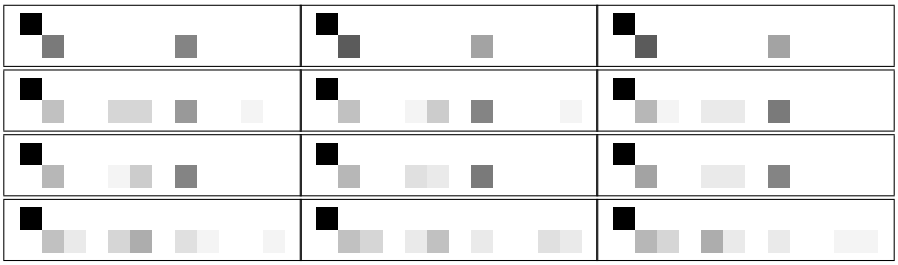


Figure A.8: Heterogeneous agents: generating classes are null and random.

A.2 Confusion matrices 2

The confusion matrices in Figures A.7 - A.16 are generated by the experiments with a complete set of the candidate model classes. The generating model class sets are varied so that in each experiment there only two generating models: null and random, greedy and Q, or iEWA and SEWA with common parameter values for each agent (Figures A.7 and A.12), followed by greedy and Q, and iEWA and sEWA with individual parameter values (Figures A.13 and A.16). Since the parameter fitting scheme does not make any difference for the null and random classes, these versions of the matrices are omitted.

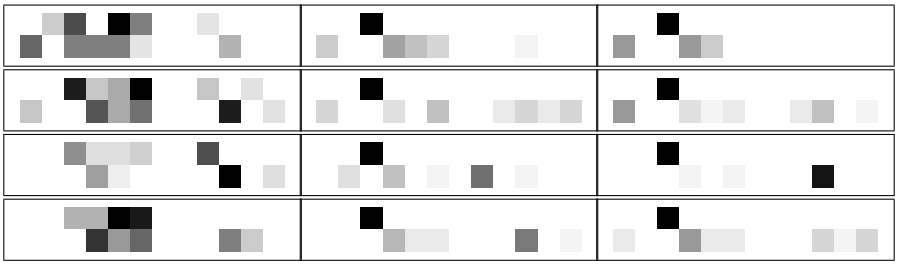


Figure A.9: Homogeneous agents: generating classes are greedy and Q (collective parameter values).

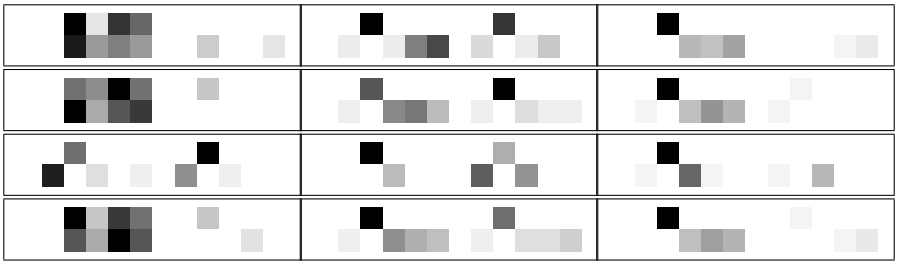


Figure A.10: Heterogeneous agents: generating classes are greedy and Q (collective parameter values).

A.3 Confusion matrices 3

The confusion matrices in Figures A.17 - A.26 are generated by the experiments in which the generating model classes are excluded from the candidate class set. Therefore, there are only ten columns in these matrices as opposed to twelve. The generating model classes are varied as above: null and random, greedy and Q, or iEWA and SEWA with common parameter values in Figures A.17 - A.22, followed by greedy and Q, and iEWA and sEWA with individual parameter values in Figures A.23 - A.26. The individually fitted null and random class are omitted again.

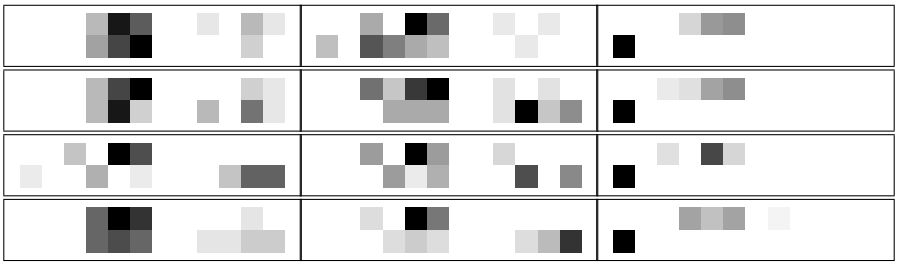


Figure A.11: Homogeneous agents: generating classes are iEWA and sEWA (collective parameter values).

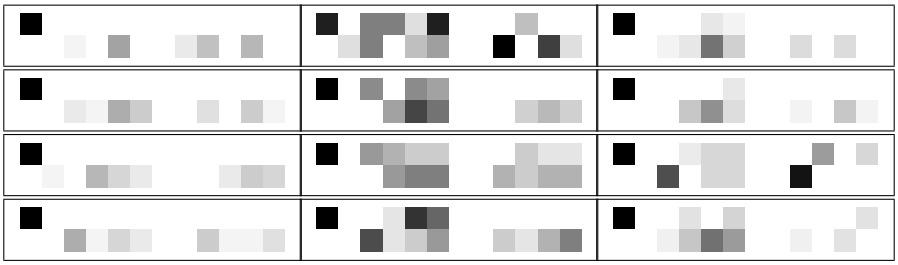


Figure A.12: Heterogeneous agents: generating classes are iEWA and sEWA (collective parameter values).

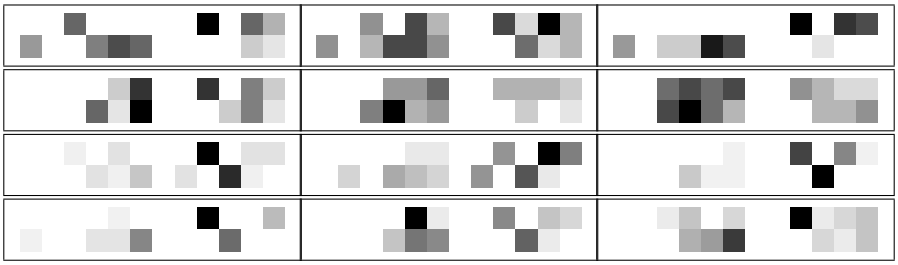


Figure A.13: Homogeneous agents: generating models greedy and Q (individual parameter values).

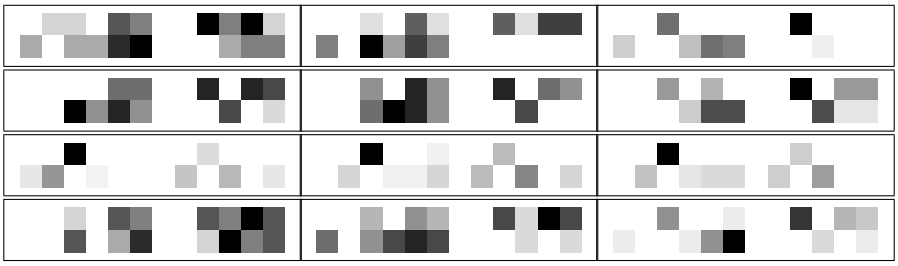


Figure A.14: Heterogeneous agents: generating models greedy and Q (individual parameter values).

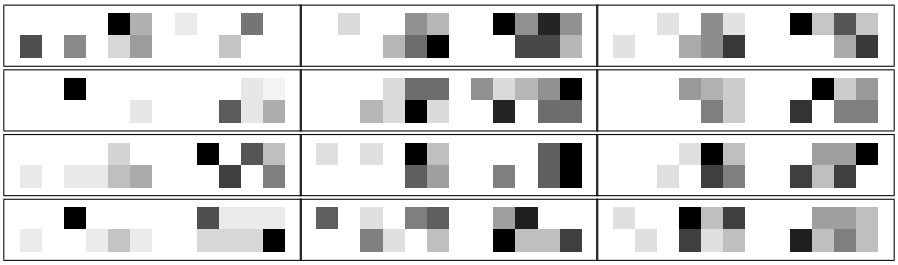


Figure A.15: Homogeneous agents: generating models iEWA and sEWA (individual parameter values).

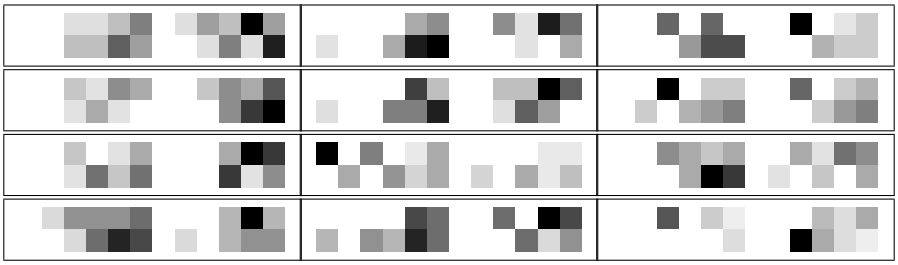


Figure A.16: Heterogeneous agents: generating models iEWA and sEWA (individual parameter values).

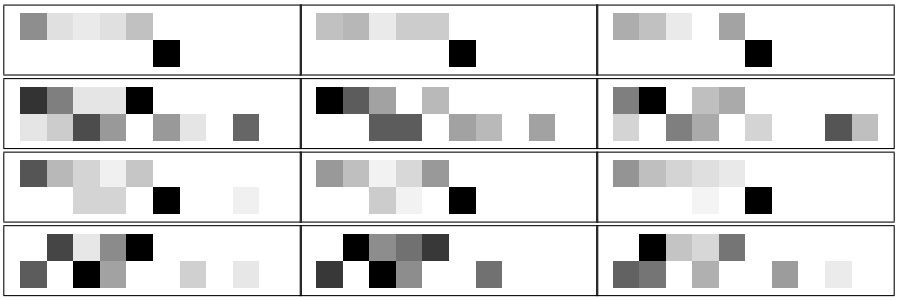


Figure A.17: Homogeneous agents: generating classes, excluded from candidates, are null and random.

Figure A.18: Heterogeneous agents: generating classes, excluded from candidates, are null and random.

A.4 Error histograms

The histograms in Figures A.27 and A.28 show, for homogeneous and heterogeneous agents respectively, the distributions of squared errors with artificial data from Experiment II. The distributions are aggregated over the following candidate model classes: random, greedy, Q, iEWA and sEWA, both collectively and individually fitted. Null model is omitted since it is not of real importance or interest.

Summary statistics of the error values are presented in Tables A.1 and A.2. The rightmost column gives the number of unique error values of all 9000 values. The

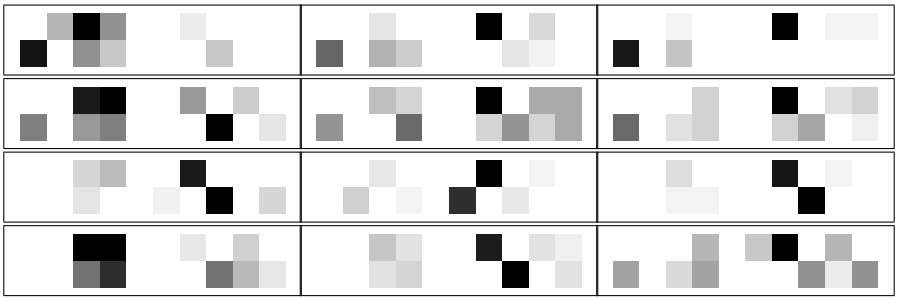


Figure A.19: Homogeneous agents: generating classes, excluded from candidates, are greedy and Q (collective parameter values).

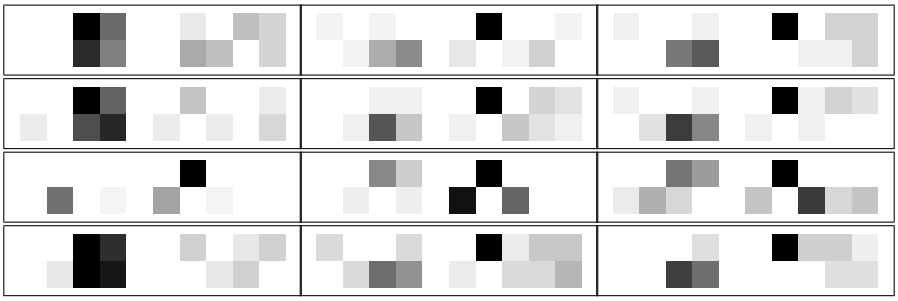


Figure A.20: Heterogeneous agents: generating classes, excluded from candidates, are greedy and Q (collective parameter values).

minimum error is not listed since it is zero for all spatial metrics.

Spatial metric	μ	σ	Median	Max	Unique values
Mean abs. difference	18.15	7.47	19.73	48.99	3399
Composition	2.51	4.66	.4970	48.67	3239
Edge density	29.64	42.49	9.45	204.76	5974
Mean patch size	1.65×10^5	3.36×10^5	2.48×10^4	2,474,944	7094

Table A.1: Summary statistics of the squared error values for spatial metrics, aggregated over all model classes.

Spatial metric	μ	σ	Median	Max	Unique values
Mean abs. difference	162.54	64.74	170.32	438.28	3270
Composition	60.24	48.90	51.07	437.36	7028
Edge density	654.97	581.49	560.46	5.26×10^3	7469
Mean patch size	3.99×10^4	3.25×10^4	3.34×10^5	2.74×10^5	7455

Table A.2: Summary statistics of the squared error values for spatial metrics, aggregated over all model classes.

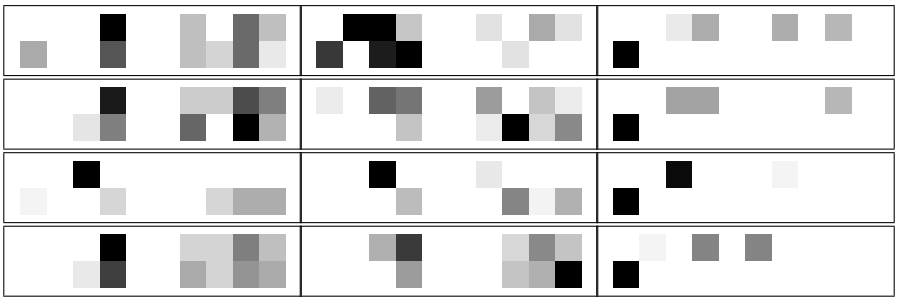


Figure A.21: Homogeneous agents: generating classes, excluded from candidates, are iEWA and sEWA (collective parameter values).

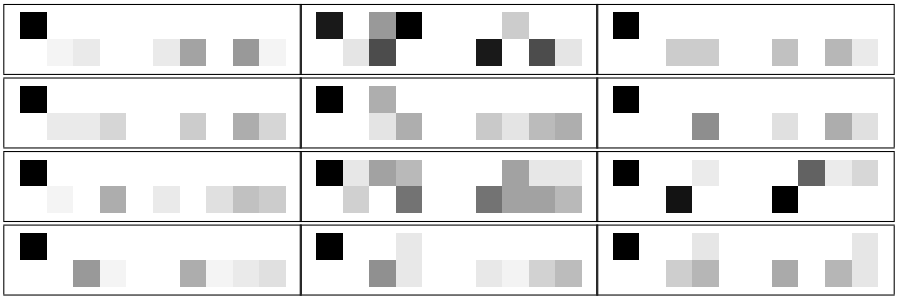


Figure A.22: Heterogeneous agents: generating classes, excluded from candidates, are iEWA and sEWA (collective parameter values).

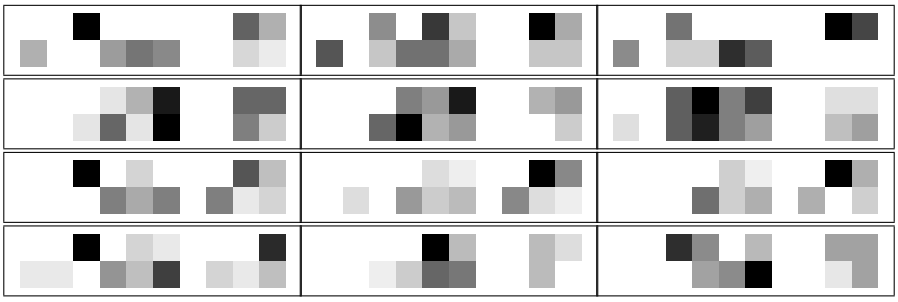


Figure A.23: Homogeneous agents: generating classes, excluded from candidates, are greedy and Q (individual parameter values).

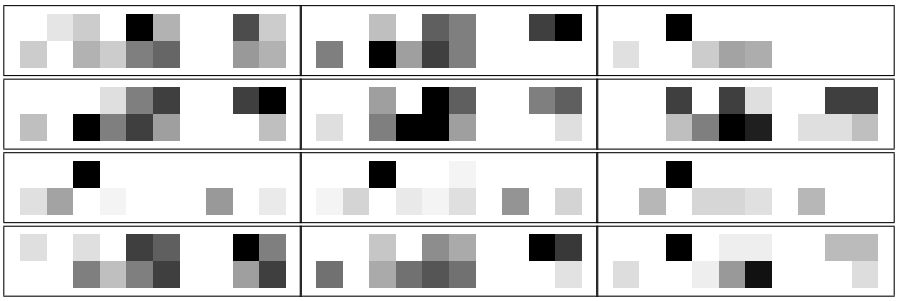


Figure A.24: Heterogeneous agents: generating classes, excluded from candidates, are greedy and Q (individual parameter values).

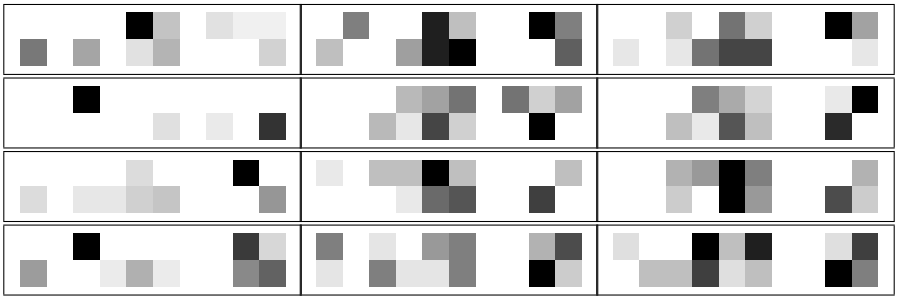


Figure A.25: Homogeneous agents: generating classes, excluded from candidates, are iEWA and sEWA (individual parameter values).

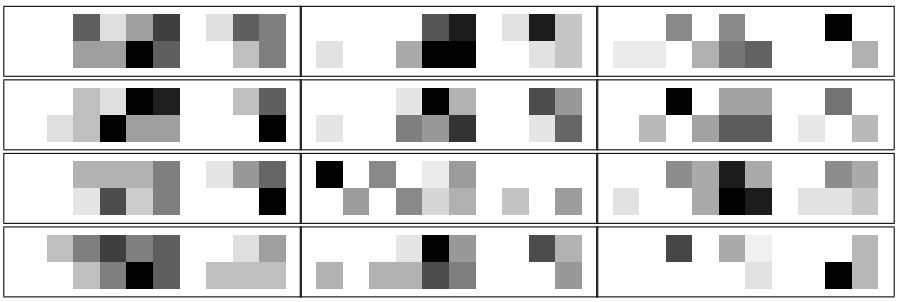


Figure A.26: Heterogeneous agents: generating classes, excluded from candidates, are iEWA and sEWA (individual parameter values).

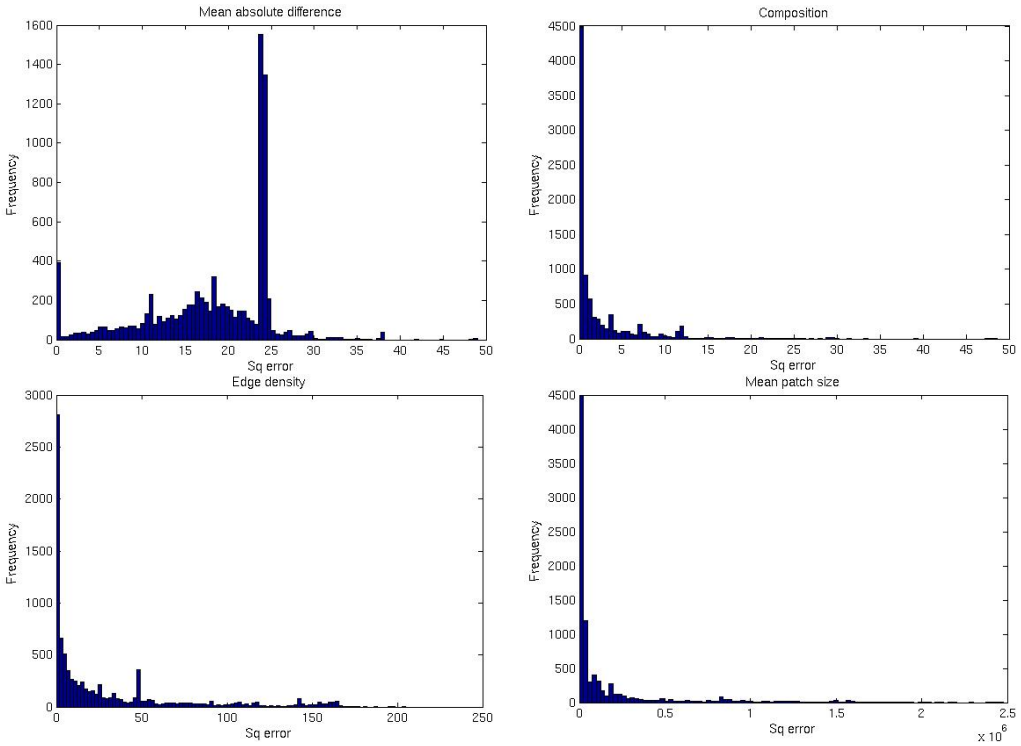


Figure A.27: The error distributions with homogeneous agents in artificial data.

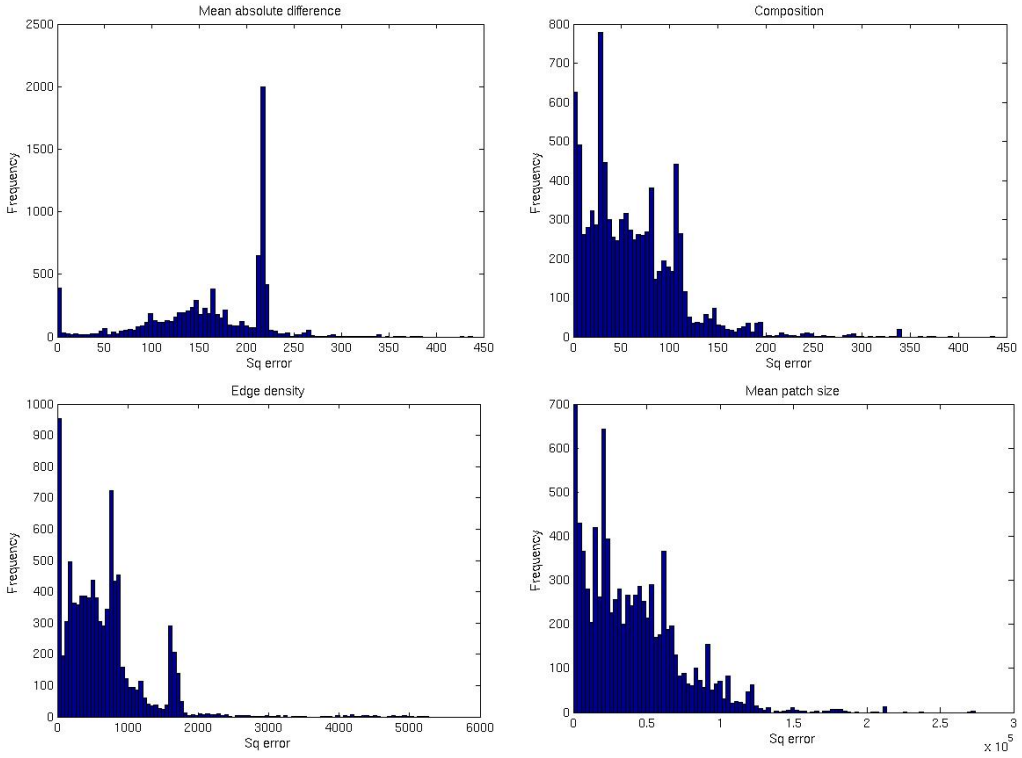


Figure A.28: The error distributions with heterogeneous agents in artificial data.

B

Results of Experiment III

B.1 Error Histograms

The squared error distributions for different spatial metrics with Indiana data are presented in Figures B.1 and B.2 for homogeneous and heterogeneous agents, respectively.

As noted in Chapter 5.4 the model classes make much more error with Van Buren data than with Indian Creek. Since these distributions are constructed over both data sets, the error distributions are either bi-polar or resemble a uniform distribution, whereas the distributions with artificial data peak at either small or median values.

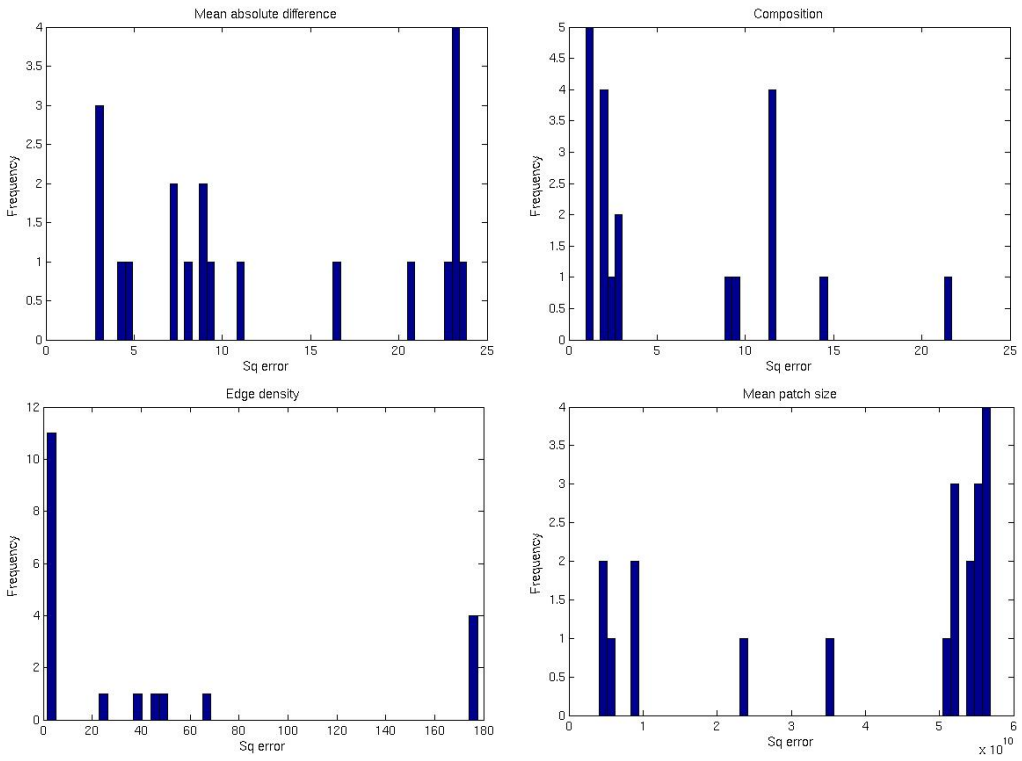


Figure B.1: The error distributions with homogeneous agents in Indiana data.

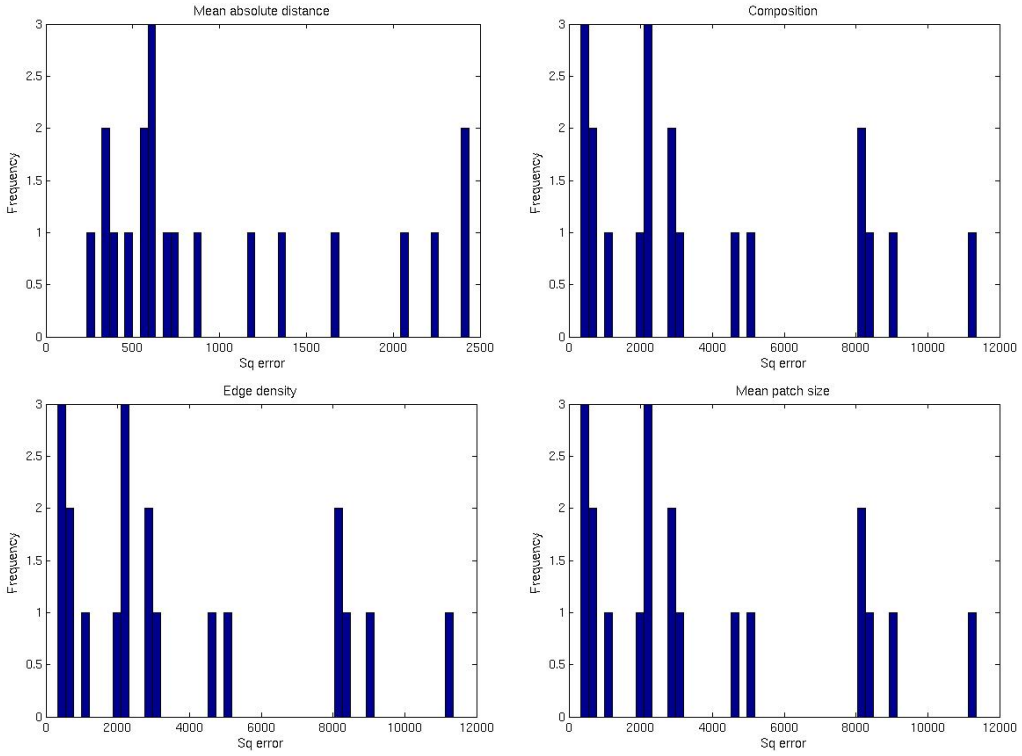


Figure B.2: The error distributions with heterogeneous agents in Indiana data.