Personal perspectives on ABM:

Development of the Pseudo-history Approach

Li An

Questions: Sampling is a fundamental way to collect data for scientific investigations in various disciplines. For instance, scientists have been employing sampling data to test hypotheses, investigate new phenomena, and establish new theories. However, people have seldom, if ever, asked questions regarding the geographic size of the sampling frame, the time span of data collection (for longitudinal studies), and the frequency of data collection. Decisions related to these questions are usually made based on the researcher's experiences, data availability, or "common sense" without exploring whether the relationships under investigation can be captured by data collected using the chosen sampling strategy. An example in Geography is that researchers choose spatial data sets (e.g., satellite images) of their study site over a certain time span and at a certain time interval, which are largely decided by factors such as costs and availability of the data. This lack of examining the match between the extent, scale, and rate of the real processes (often not directly observable) under investigation and the corresponding sample strategy may substantially undermine the validity of the subsequent analysis and findings. Many reasons may account for this problem, such as uncertainties in the processes and the related system structures of interests, difficulties in observing and quantifying the spatial and temporal scales of some phenomena, and costs of data collection in some large areas, over long time, and at short time interval.

Justification: The advent of fast computers in the last decades and advancement in software have brought forward new opportunities to address the above sampling strategy questions using agent-based spatial models. Rather than solely depending on real data that are sometimes subject to the above constraints, researchers can use computers to generate an artificial digital "world", and let whatever processes of scientific interests proceed on this 'world". Theories, experiences, and hypotheses could be used to guide or affect the directions, strengths, and interactions of such processes. The emergent patterns on the artificial world, sometimes unexpected and surprising, may arise from the behaviors of many agents (autonomous entities or objects that have goals, some degree of knowledge, and actions), the interactions among themselves, and their relationships with the environment(s). This type of "realistic" mapping of real-world entities and processes onto a computer model facilitate a totally new scientific investigation in comparison with traditional analyses based on empirical data: computational simulation. A digital artificial laboratory can be constructed to take major relevant entities and processes under certain research objectives and eliminate unimportant details and noises, thus reducing uncertainties, difficulties, costs, which would otherwise prohibit many data collection, analysis, and scientific investigation activities. Thus Computational simulations undoubtedly will bring a new era in addressing the above questions relating to the match between the extent, scale, and rate of real processes (but often not directly observable) and our sampling strategy.

Because the researcher controls the processes that underlie the observable patterns through setting the parameters or choosing the algorithms, we can freely choose varying

sampling strategies by setting the simulation time frame, automating sampling at the designated intervals, and choosing a certain geographic area within our artificial world. Analyzing data obtained through such strategies with the knowledge of the processes ongoing should be able to reveal the effectiveness of different sampling strategies. However, such a crucially important area has seldom, if ever, been touched. For instance, a web of science search under this combination of key words "(sample or sampling) and (artificial world or computational simulation)" resulted in two irrelevant records (this search is not exclusive, but may be still suggestive).

The pseudo-history ABM approach: A new approach developed by An and Brown (in preparation); see CV, the pseudo-history analysis, can be used as a socio-environmental laboratory that accommodates all the sampling strategy related needs, conduct computer-based experiments, and advance scientific research related to the effectiveness of sampling strategy. The current pseudo-history model encapsulates a hypothetical "landscape" that is available for human residence and service center development at the beginning. The "landscape" can vary by size, and several variables (parameters) control its environmental and socioeconomic features such as soil quality and distance to service center. As time goes on, i.e., as computer internal clock moves, "homebuyers" (computer objects) with various characteristics and preferences (parameters) enter the "landscape", evaluate different locations (controlled by algorithms and equations) as candidate locations, and choose locations that meet their objectives of maximizing residence utility.

Given such a hypothetical "landscape" and the residence-choice related processes on it, the following tasks can be implemented in order to achieve the above objectives: (1) further verify the current model, including debugging, (2) let some residence decision-related parameters correlat with changes in the environmental or socioeconomic parameters, (3) add new features to the model, such as letting the residence-choice and land-use processes have higher correlation with time and space, automating the model in taking samples at varying sizes, spans, and time intervals, and writing the sample results in appropriate formats, and finally (4) examine what sampling strategies in relation to geographic sampling frame, sampling span, and sampling interval would best disclose model inputs that have generated the sample data. Furthermore, because the researcher knows the true mechanisms ongoing and their temporal/spatial scales (established as parameters and rules in the ABM), and thus many uncertainties are removed, the pseudo-history approach can be used to test what statistical methods can best detect the underlying rules accommodated in the ABM.

Significance: This approach is significant for several reasons. First, it explores the effectiveness of varying sampling strategies, which would benefit scientists who use sampled data regardless of what fields they come from. Second, it will advance methods and theories of the Geocomplexity, an increasingly recognized field that connects Geography, computer science and engineering, and complexity science, witnessing tremendous applications in many theoretical and empirical studies. Last, this approach will add crucial components to a spatial modeling and simulation curriculum, and give graduate students invaluable experiences in computational modeling and simulation.