Use of Agent-based Simulation Data in Assessing the Inferential Power of Statistical Methods

Ninghua(Nathan) Wang[1]， Li An[2]

[1]Joint Ph.D. Program in Geography
San Diego State University and UC Santa Barbara
Email: wangn@rohan.sdsu.edu

[2]Department of Geography
San Diego State University
Email: lan@mail.sdsu.edu

Inferential statistical methods play an important role in uncovering underlying mechanisms behind land-use changes. A variety of methods have been proposed for this task, e.g., ordinary least square linear regression (Allen and Barnes 1985), logistic regression (Wu and Yeh 1997), and survival analysis (An and Brown 2008). While valuable insights those methods can bring, their usefulness under different conditions has rarely been evaluated. Most land-cover change processes involve complex features, such as feedbacks, chaos, and emergence (Lambin and Geist 2006), and it remains unclear how statistical models function in presence of those effects. Second, most statistical models are borrowed from other fields so their assumptions may not be fulfilled in geographical applications. Furthermore, data availability in terms of spatial and temporal resolutions may severely constrain the performance of certain models. Given these limitations, an assessment of the inferential power of statistical methods seems necessary. However, traditional assessment based on empirical case studies is inadequate because of unknown underlying mechanism, limited conditions, and data incompleteness. A novel approach is thus called for.

We demonstrate an innovative use of data generated by a land-use change agent-based simulation model coupled with a Monte Carlo sampling of its parameter space. Agent-based simulation data, as opposed to empirical data, can be used to better assess statistical method in at least three ways:

First, a simulated dataset minimizes uncertainties in method assessment. Uncertainties arise when assessing statistical methods applied to real data, because it is only possible to hypothesize what the underlying mechanisms are, by using various theories, experts' opinions, and statistical evaluation methods. It is very difficult to know for certain, or to manipulate with experimental control, the underlying mechanisms generating a real-world data set. However, by using simulation data the underlying mechanisms, drivers, and parameters are completely known and the ability of statistical methods to recover them can be accurately assessed (Hirzel, Helfer and Metral 2001).

Second, comparing statistical methods on specific case studies, as has been done in previous work (Malanson 2005, Pontius et al. 2008), limits the application situations under which the methods can be explored. Different land-change processes involve different driver sets (Rindfuss et al. 2008, An et al. 2011), operating at various speeds (Lambin and Geist 2006), and having different levels of information access by land managers and users (Manson 2006). These factors lead to diverse land-change pathways. Acquiring an empirical dataset that spans a diverse set of pathways is very difficult, if possible at all, but relatively easy for a simulation dataset, where alteration of the simulation program's initial parameters produces a range of pathways (Rindfuss et al. 2007, Clarke et al. 2007). As such, we can map the relative performance of each statistical technique across application situations.

Third, the simulation approach enables full control data production and ensures data completeness (Neel, McGarigal and Cushman 2004, Epperson et al. 2010). Though data completeness is a key consideration in selecting a method (Harrell 2001), it is often limited spatially and temporally in real datasets, and in terms of resolution and extent of data coverage. To study the sensitivity of methods to these kinds of limits, we test methods on datasets sampled at different frequencies.

Though computer simulation data has been widely used in statistics literature to test alternative methods, it is a relatively new practice in geography. In statistics, simulation data are mostly generated from predetermined distributions of parameters (e.g. uniform, normal, Weibull) but such distributions have limited application in geography. Instead, individual behaviors and interactions are key factors of many geographical processes, where agent-based simulation could generate data better reflecting geographical reality (Turner, Stephens and Anderson 1982, Grimm and Railsback 2005). In recent years, such data have been applied to assess metrics for animal movement analysis (Miller 2012) and models for landscape genetics (Epperson et al. 2010). If an inferential method could reveal the roles of spatial variables that are used in the agent-based simulation, it will increase our confidence that this inferential method could reveal the roles of real spatial variables.

There is a risk that virtual land-change processes generated by one simulation program may lack generality, limiting the scope of conditions under which conclusions about the efficacy of statistical methods will apply. In this study, we replicated all our experiments using a completely different simulation program (see below).

The procedure of assessing inferential methods involved 4 steps (Figure 1). With a specific set of predictor variables and parameters, we (1) simulated a virtual land-change process, (2) sampled the data that resulted from that process spatially and temporally following specific rules, (3) exposed the sampled data to statistical methods of interest, and (4) evaluated those methods by comparing their coefficients to the actual parameters we programmed in the ABM. This evaluation procedure was repeated 1,000 times over different model variables and parameter values, both selected by the Monte Carlo technique.

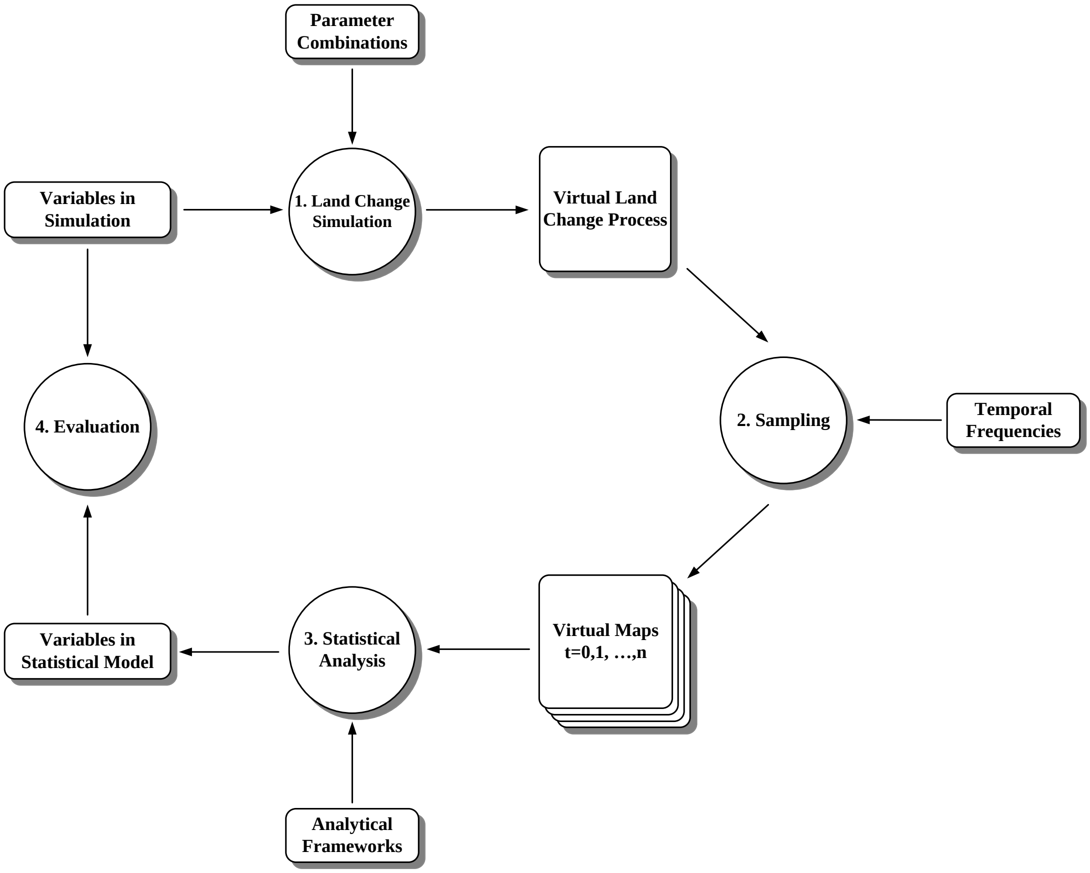

Figure 1

Case Study

In this case study we compared two statistical methods: logistic regression and survival analysis. Logistic regression is a very popular statistical method for inferring predictor variables in land-change processes (Coomes, Grimard and Burt 2000, Cheng and Masser 2003, Huang, Zhang and Wu 2009, López and Sierra 2010, Wyman and Stein 2010, Huang, Cai and Peng 2007) because it can associate binary land status, e.g. changed or not, to a suite of biophysical and/or socioeconomic variables. Despite many advantages, logistic regression suffers difficulties in modeling processes with multiple time measures. If both land types and independent variables change values over time, logistic regression has to use either the value at one time or the average over time, which results in loss of information and reduction in degrees of freedom (An and Brown 2008). In contrast, survival analysis, as a method explicitly dealing with the occurrence and timing of events, shows promise in analyzing multi-temporal data. Theoretically, survival analysis should outperform logistic regression in analyzing multi-temporal data. However, survival analysis is relatively more complex than logistic regression, and has requirements on higher time resolution of data. As such, the relative performance of each statistical framework needs to be compared under different circumstances, as a way to improve their use in analyses of land-change processes.

To compare their relative performances, we employed the SOME (Brown et al. 2005) land-use change simulation (also replicated findings with IDEAL (Ligmann-Zielinska and Sun 2010)) and tested these two methods for a wide range of conditions which were constructed as following:

> Step 1: SOME can include up to 4 predictors variables in simulation. We varied the number of variables included, in particular, time-dependent variables (i.e. variables change value over time) and/or time-independent variables, and tested if those variables can be effectively detected.
>
> Step 2: Two parameters out of a dozen from SOME, *numresidents* and *numtests*, are swept over their ranges. *numresidents* controls the amount of land change while *numtests* controls the level of determinism of land development. We acknowledge that in real world there may be more land-change possibilities that cannot be controlled by these two parameters. However we would like to choose them because 1) they determine the quantity and spatial distribution of land change, and 2) they affect the validity of some statistical assumptions when their values go to extremes.
>
> Step 3: We used the Monte Carlo technique to sample 1,000 specific land-change conditions from the full set of possible parameter combinations for simulations on our workstation (Dell Precision T5400 Intel Xeon 3.16GHz Qual-core 16G memory).

Of the 1,000 Monte Carlo experiments, survival analysis achieved a success rate 60% higher than logistic regression (Figure 2, success rate is defined as the number of experiments over 1,000 that a statistical method correctly detects the variables based on statistical significance). This result confirms the theoretical speculation that survival analysis generally performs better than logistic regression. Moreover, while survival analysis and logistic regression performed equally well in detecting time-independent variables, survival analysis outperformed logistic regression in detecting time-dependent variables (Figure 3). This may arise from survival analysis' more effective use of temporal information. However, survival analysis was more likely to detect the corresponding land-change variables at higher frequencies (Figure 4), whereas logistic regression was recommended when samples in time are sparse.

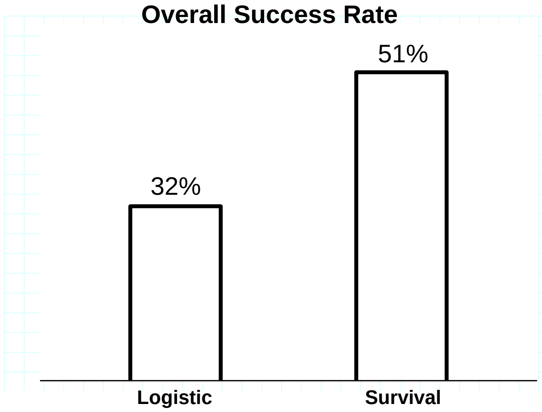

Figure 2

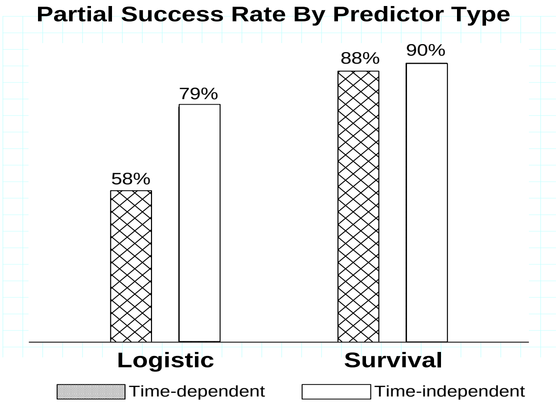

Figure 3

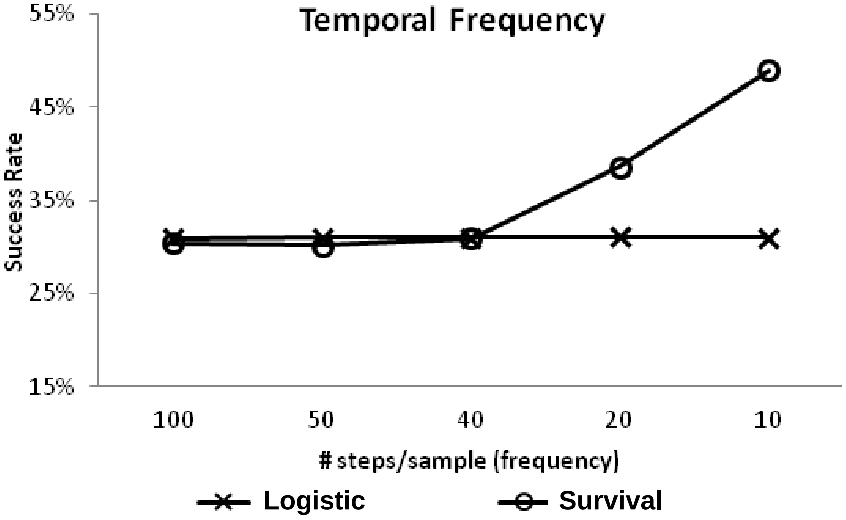

Figure 4

Conclusion

The simulation approach is useful when assessing the inferential power of statistical models in land-change studies. One advantage that cannot be otherwise obtained is its endowment of a land-change researcher's better control of experiment processes (thus less "noise" or unknown processes), allowing for potentially more accurate assessment of different statistical methods when applied in different land-change situations. Second, a rich and diverse set of land-change processes can be generated, as opposed to one or a few that might be contained in an empirical setting. In response to the risk that virtual processes do not reflect the wide range of possible underlying processes very well, we validated our findings using another simulation program, which was independently developed based on different assumptions and conceptualizations, and the match of results corroborates our findings.

Reference:

Allen, J. & D. Barnes (1985) The causes of deforestation in developing countries. *Annals of the Association of American Geographers,* 75, 163-184.

An, L. & D. Brown (2008) Survival Analysis in Land Change Science: Integrating with GIScience to Address Temporal Complexities. *Annals of the Association of American Geographers,* 98, 323-344.

An, L., D. G. Brown, J. I. Nassauer & B. Low (2011) Variations in development of exurban residential landscapes: timing, location, and driving forces. *Journal of Land Use Science,* 6, 13-32.

Cheng, J. & I. Masser (2003) Urban growth pattern modeling: a case study of Wuhan city, PR China. *Landscape and Urban Planning,* 62, 199-217.

Clarke, K. C., N. Gazulis, C. Dietzel & N. C. Goldstein (2007) A decade of SLEUTHing: lessons learned from applications of a cellular automaton land use change model. *Classics in IJGIS: Twenty years of the International Journal of Geographical Information Science and Systems*, 413-427.

Coomes, O., F. Grimard & G. Burt (2000) Tropical forests and shifting cultivation: Secondary forest fallow dynamics among traditional farmers of the Peruvian Amazon. *Ecological Economics,* 32, 109-124.

Epperson, B., B. McRae, K. Scribner, S. Cushman, M. Rosenberg, M. Fortin, P. James, M. Murphy, S. Manel & P. Legendre (2010) Utility of computer simulations in landscape genetics. *Molecular Ecology*.

Grimm, V. & S. F. Railsback. 2005. *Individual-based modeling and ecology*. Princeton university press.

Harrell, F. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Verlag.

Hirzel, A., V. Helfer & F. Metral (2001) Assessing habitat-suitability models with a virtual species. *Ecological modelling,* 145, 111-121.

Huang, B., L. Zhang & B. Wu (2009) Spatiotemporal analysis of rural–urban land conversion. *International Journal of Geographical Information Science,* 23, 379-398.

Huang, Q., Y. Cai & J. Peng (2007) Modeling the spatial pattern of farmland using GIS and multiple logistic regression: a case study of Maotiao River Basin, Guizhou Province, China. *Environmental Modeling and Assessment,* 12, 55-61.

López, S. & R. Sierra (2010) Agricultural change in the Pastaza River Basin: A spatially explicit model of native Amazonian cultivation. *Applied Geography,* 30, 355-369.

Lambin, E. F. & H. Geist. 2006. *Land-use and land-cover change: local processes and global impacts*. Springer Verlag.

Ligmann-Zielinska, A. & L. Sun (2010) Applying time-dependent variance-based global sensitivity analysis to represent the dynamics of an agent-based model of land use change. *International Journal of Geographical Information Science,* 24, 1829-1850.

Malanson, J. (2005) Comparison of the structure and accuracy of two land change models. *International Journal of Geographical Information Science,* 19, 745-748.

Manson, S. M. (2006) Bounded rationality in agent-based models: experiments with evolutionary programs. *International Journal of Geographical Information Science,* 20, 991-1012.

Miller, J. A. (2012) Using Spatially Explicit Simulated Data to Analyze Animal Interactions: A Case Study with Brown Hyenas in Northern Botswana. *Transactions in GIS,* 16, 271-291.

Neel, M. C., K. McGarigal & S. A. Cushman (2004) Behavior of class-level landscape metrics across gradients of class aggregation and area. *Landscape Ecology,* 19, 435-455.

Pontius, R., W. Boersma, J. Castella, K. Clarke, T. de Nijs, C. Dietzel, Z. Duan, E. Fotsing, N. Goldstein & K. Kok (2008) Comparing the input, output, and validation maps for several models of land change. *The Annals of Regional Science,* 42**,** 11-37.

Rindfuss, R. R., B. Entwisle, S. J. Walsh, L. An, N. Badenoch, D. G. Brown, P. Deadman, T. P. Evans, J. Fox & J. Geoghegan (2008) Land use change: complexity and comparisons. *Journal of Land Use Science,* 3**,** 1-10.

Rindfuss, R. R., B. Entwisle, S. J. Walsh, C. F. Mena, C. M. Erlien & C. L. Gray (2007) Frontier land use change: synthesis, challenges, and next steps. *Annals of the Association of American Geographers,* 97**,** 739-754.

Turner, M. E., J. C. Stephens & W. W. Anderson (1982) Homozygosity and patch structure in plant populations as a result of nearest-neighbor pollination. *Proceedings of the National Academy of Sciences,* 79**,** 203-207.

Wu, F. & A. G. O. Yeh (1997) Changing spatial distribution and determinants of land development in Chinese cities in the transition from a centrally planned economy to a socialist market economy: a case study of Guangzhou. *Urban Studies,* 34**,** 1851.

Wyman, M. & T. Stein (2010) Modeling social and land-use/land-cover change data to assess drivers of smallholder deforestation in Belize. *Applied Geography,* 30**,** 329-342.