Tsukuba Geoenvironmental Sciences, Vol. 8, pp. 3-7, Dec. 28, 2012

Introducing new measures of accuracy for land-use/cover change modeling

Ronald C. ESTOQUE[*, **] and Yuji MURAYAMA[*]

Abstract

Regardless of the method used to model land-use/cover (LUC) change, there is a need to assess the accuracy of the modeling. However, the LUC change modeling profession is still confronted with some important issues, including the problems on the focus of accuracy assessment, parameters to indicate overall accuracy, parameters for comparing different modeling results and the minimum accuracy standard. The purpose of this paper is to introduce new measures of accuracy, namely the HOC (hits to observed change), MOC (misses to observed change) and FOC (false alarms to observed change) ratio indices, and discuss their potential relevance to the broad field of LUC change modeling. Firstly, the paper reviews existing validation parameters and presents a conceptual analysis of these parameters, including the three proposed new indices. And secondly, the paper discusses the potentials of these indices relative to the current important issues of LUC change modeling accuracy assessment.

Key words: accuracy, modeling, ratio indices, validation

1. Introduction

Accuracy assessment, also known as validation, is a vital component of any LUC modeling exercise. Regardless of the method used to model LUC change, there is a need to assess the accuracy of the results. Validation is a process by which the quality of model parameters is evaluated by comparing the model's simulation results to a valid reference map or to an independent data set for the end time of the simulation interval (Pontius and Malanson, 2005; Vliet *et al*., 2011).

The concepts of error due to quantity and error due to location (now called error due to allocation) were first proposed by Pontius (2000), and have been among the most important topics in the LUC change modeling profession in recent years. To account for both types error in model validation and solve some conceptual problems observed on the standard Kappa index, different Kappa variants have been formulated (Pontius, 2000), but none have proved highly satisfactory (Pontius and Millones, 2011).

* Faculty of Life and Environmental Sciences, University of Tsukuba, Japan

** Don Mariano Marcos Memorial State University, the Philippines

Klug *et al*. (1992) and Perica and Foufoula-Georgiou (1996) proposed figure of merit (FoM) to assess the agreement of LUC changes rather than just the LUC categories by taking the ratio of the intersection of the observed change and simulated change to the union of the observed change and simulated change. FoM has been applied to compare different modeling results (Pontius *et al*., 2008).

Chen and Pontius (2010) introduced four components of correctness and error, *viz*. null successes (correct due to observed persistence simulated as persistence), hits (correct due to observed change simulated as change), misses (error due to observed change simulated as persistence) and false alarms (error due to observed persistence simulated as change). The concepts of these components have been adopted and implemented in recent studies (*e.g*. Ahmed and Ahmed, 2012; Sloan and Pelletier, 2012; Thapa and Murayama, 2011).

Indeed, literature shows that the concepts of errors due to quantity and allocation, null successes, hits, misses and false alarms have been applied in previous studies. However, these are usually expressed relative to the whole landscape under investigation. The ensuing problem is that it is often difficult to make comparisons of accuracy assessment results among different LUC change modeling studies. It is because landscape size varies across study sites or areas. In this regard, there is a need to focus only on the observed and simulated changes when assessing accuracy. Needless to say, the central core of LUC change modeling is on the simulation or projection of LUC change (observed change) and not on the lands that did not change. Accuracy assessment based on the observed and simulated changes can potentially provide the basis for effective comparison of different LUC change modeling results. However, validation parameters tailored for this type of accuracy assessment and those that can directly provide answers to the following basic but very important questions are still lacking. "How much of the observed change was simulated correctly and missed?" How much false alarms were simulated relative to the observed change?" These questions, although basic, are valid because they are applicable to all LUC change modeling studies that produce hard predictions or simulations, regardless of the model used. In this type of modeling, each pixel in the simulation and reference maps belongs to exactly one LUC category.

This paper introduces three new measures of accuracy

that take into account the observed and simulated changes only, namely the HOC (hits to observed change), MOC (misses to observed change) and FOC (false alarms to observed change) ratio indices. Their differences from the other established validation parameters, like FoM and the components of correctness and error, are discussed. Their potentials as comparison parameters and in setting a minimum accuracy standard are explored. Similar to the concepts of hits, misses and false alarms, these indices focus on LUC change modeling that produces hard predictions or simulations.

2. Conceptual analysis

Consider Figures 1a and 1b as the LUC maps at Time 1 (t_1) and Time 2 (t_2), respectively, both containing two categories, green and yellow. The hypothetical goal was to simulate the LUC change from green to yellow from t_1 to t_2 (Fig. 1c) using Model 1 and Model 2. Their results (Figs. 1d, 1e) were compared using the following validation methods and parameters: (1) simple comparison of the actual and simulated quantities of LUC categories; (2) components of correctness and error: null successes, hits, misses and false alarms; (3) FoM; and (4) ratio indices: HOC, MOC and FOC (Eqs. 1–3).

$$HOC = \frac{H}{H + M} \tag{1}$$

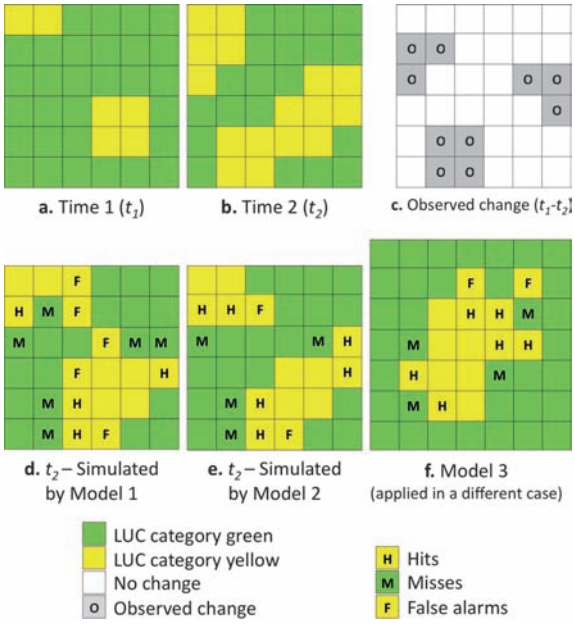

Fig. 1 LUC change modeling examples used to examine the validation methods and parameters presented in this paper. The hypothetical goal was to simulate the LUC change from green to yellow from t_1 to t_2 using Model 1 and Model 2.

$$MOC = \frac{M}{H + M} \tag{2}$$

$$FOC = \frac{F}{H + M} \tag{3}$$

where HOC, MOC and FOC are, respectively, the individual ratio of hits (H), misses (M) and false alarms (F) against the observed change (summation of H and M).

By comparing the quantity of the yellow category on the reference map at t_2 (Fig. 1b) with the quantities in the simulated maps (Figs. 1d, 1e), it can be seen that Model 1 had a quantity that was much closer to the reference value than Model 2 (Table 1a). However, this validation method could give a completely misleading assessment because in the example, Model 2 had two more pixels simulated correctly than Model 1 (Figs. 1d, 1e). Thus, it is very important to look at the four components of correctness and error, which show a completely different assessment results. Relative to the whole landscape, Model 2 had higher null successes and hits, and lower misses and false alarms than Model 1 (Table 1b). The comparison between Model 1 and Model 2 in this instance, using these four components as comparison parameters, is clearly possible because both models used the same information, such as landscape size. The issue on the applicability of the four components of correctness and error as comparison parameters arises when Models 1 and 2 are compared with Model 3, which was applied in a different case study area with a different landscape size (Fig. 1f). Figures 1e (Model 2) and 1f (Model 3) had

Table 1. Validation results of the modeling examples.

a. Comparison of the quantities of category yellow (% landscape)

	t_2 (Model 1)	t_2 (Model 2)	t_2 (Reference)
Category yellow	41.67	38.89	44.44

b. Components of correctness and error (% landscape)

	t_2 (Model 1)	t_2 (Model 2)	Model 3
Null success	58.33	66.67	75.51
Hits	11.11	16.67	12.24
Misses	16.67	11.11	8.16
False alarms	13.89	5.56	4.08

c. Figure of merit (FoM)[a]

	t_2 (Model 1)	t_2 (Model 2)	Model 3
FoM	26.67	50.00	50.00

d. Ratio indices (relative to the observed change)

	t_2 (Model 1)	t_2 (Model 2)	Model 3
HOC	0.40	0.60	0.60
MOC	0.60	0.40	0.40
FOC	0.50	0.20	0.20

[a] FoM = [H/(H+M+F)] × 100

the same quantities (no. of pixels) of observed change, hits, misses and false alarms. However, Table 1b shows that Models 2 and 3 had different accuracy assessment results. Hence, Figure 1e cannot be compared with Figure 1f using the four components as comparison parameters because the values are based on two different landscape sizes.

Nonetheless, the FoM, which was calculated by taking the ratio percentage of hits from the summation of the hits, misses and false alarms, offers the solution. The hypothetical example shows that, regardless of the size of the two landscapes, Figures 1e and 1f had the same FoM (Table 1c) because they had the same LUC change characteristics (*i.e*. quantities of observed change, hits, misses and false alarms). This shows that FoM can be used to effectively compare simulation results. FoM only takes into account the observed and simulated changes.

The three ratio indices, *viz*. HOC, MOC and FOC also produced the same result as FoM, that is, Figures 1e and 1f had the same HOC, MOC and FOC ratio indices. These indices (Table 1d) are also not influenced by the size of the landscape. However, if our goal was to answer the very important basic questions posed in the introductory part, FoM, as a summary statistic, cannot directly provide the answers. Only the three ratio indices, HOC, MOC and FOC, can provide direct answers to these questions.

Although false alarms are not part of the observed change, they contribute to the error due to allocation. The comparison of HOC against FOC can help in the assessment of a modeling result, for example, by comparing the rate of hits vs. rate of false alarms per unit area of observed change. Furthermore, although MOC can be derived from the value of HOC, and vice versa (*i.e*. MOC=1–HOC; HOC=1–MOC), all these three indices are explicitly used for the purposes of clarity. The potentials of the three ratio indices and their distinctions from FoM and the components of correctness and error are further discussed in detail in the next section.

3. Discussion

3.1. Focus of accuracy assessment

The comparison of the actual and simulated quantities of LUC categories (Huang and Cai, 2007; Kamusoko *et al*., 2011) or in combination with landscape indicators (Guan *et al*., 2011) can provide useful information to the specified end-users. However, a relatively high agreement in quantity could be due to the null successes and/or due to the false alarms filling in for the misses, as illustrated in Figure 1 and Table 1a. Consequently, it is difficult to determine the real quantity of the observed change that was simulated correctly. Thus, the errors due to quantity and allocation are both important for accuracy assessment.

We believe that the validation of LUC change modeling should not deviate from the central core of the subject matter: LUC change. After all, after the model has been calibrated and validated, we want to simulate or project the changes that would potentially occur in the future, given the past and present conditions. The accurate projection of these changes would enable development planners, environmental managers and policy-makers to examine whether such changes might potentially cause socio-economic and environmental chaos, and if so, determine the appropriate measures to be undertaken.

3.2. Parameters to indicate overall accuracy

Some of the recent attempts to apply the concepts of errors due to quantity and allocation include those by Khoi and Murayama (2010), that compared the agreement and disagreement due to quantity and allocation of the simulated model expressed as a percentage of the landscape with the null model. However, the overall accuracy of the modeling was based on the overall agreement, regardless of the ratios of observed change and the lands that did not change relative to the whole landscape. Ahmed and Ahmed (2012) used the components of quantity and allocation disagreements to compare the results of three LUC change models. As mentioned earlier, these components can be used as comparison parameters as long as the different models are applied in only one or in a common case study area. However, the overall accuracy was also measured based on the overall agreements and disagreements, which included the null successes of the lands that did not change. Sloan and Pelletier (2012) presented the accuracy of their forest-cover change modeling based on the agreement between the reference and the simulated maps, which, in effect, also included the null successes of the lands that did not change.

These attempts contribute to the growing application of these validation methods. However, it is still extremely difficult to capture and comprehend "how much of the observed change in each of these studies was simulated correctly and missed", and "how much false alarms were simulated relative to the observed change", whether per transition (Ahmed and Ahmed, 2012; Khoi and Murayama, 2010) or per stratum or as a whole (Sloan and Pelletier, 2012).

In the hypothetical example presented (Fig. 1e and Table 1b), Model 2 had a hits rate of 16.67% and a total error of 16.67%. However, since these values were relative to the whole landscape, we argue that it is not appropriate to report that the overall accuracy of the simulation of Model 2 was 16.67% (hits) because the denominator from which this value was computed included the lands that did not change. As illustrated in the example, at the same success

rate, the larger the area that did not change is, the lower this value would become (Fig. 1f, Table 1b). Moreover, it is also misleading to report that the overall accuracy of the simulation of Model 2 was 83.33% (*i.e.* 100 minus the total error or null successes plus hits, also referred to as the overall agreement in the studies mentioned) because it also included the null successes. A model for a landscape that has a percentage change of 4% between two time periods can have null successes and overall agreement close to 96%, even if the model will not be able to simulate correctly a part of the observed change. It is in this context that the combination of null successes and hits is not an appropriate measure of overall accuracy for any given LUC change modeling output.

The FoM of Model 2 was 50.00%, relatively higher when compared with that of Model 1 (Table 1c). Since FoM takes into account only the observed and simulated changes, we believe it provides the measure of overall accuracy. However, as mentioned earlier, if our goal was to answer the very important basic questions posed in the introductory part, FoM, as a summary statistic, cannot directly provide the answers. Only the three ratio indices: HOC, MOC and FOC can provide direct answers to these questions. They can also provide additional information not directly available from the other four components. Thus, these indices are also vital in the overall accuracy assessment report of LUC change modeling studies. Furthermore, if these indices were to be applied to evaluate LUC change modeling that dealt with multiple transitions, it would be possible to determine which transition was better modeled and which transition contributed the largest error. Thus, these indices can potentially help in the refinement of the model calibration process.

3.3. Parameters for comparing different modeling results

Pontius *et al*. (2008) applied FoM to compare different modeling outputs, while Thapa and Murayama (2011) and Sloan and Pelletier (2012) used it to measure the accuracy of their modeling outputs. However, Vliet *et al*. (2011) argued that because FoM does not include a reference level, it is not possible to interpret the absolute FoM value, and results of different models cannot be compared. However, we contend that FoM can be used as a comparison parameter when comparing outputs of different models because it considers only the observed and simulated changes, which are the more ideal reference level as opposed to the whole landscape. As illustrated in Figure 1 and Table 1c, the three components used to calculate FoM can be used as direct input in the calculation without first having them expressed relative to the whole landscape. This confirms the independency of FoM from the size of landscape. However, a positive correlation of FoM with the percentage net change (Pontius *et al*., 2008) poses a limitation. It means that modeling results for a landscape with extremely low percentage net change cannot be effectively and accurately compared with modeling results for a landscape with an extremely high percentage net change.

The three ratio indices, namely HOC, MOC and FOC, have the potential to be used as comparison parameters because they, too, are not influenced by the size of the landscape; only the observed and simulated changes of the phenomenon being modeled are considered. In fact, the very important basic questions which only these indices can directly answer can be used as common basic criteria for the comparison of LUC change modeling outputs. However, it must be pointed out that the components of correctness and error (hits, misses and false alarms), from which the three indices are derived, can only be determined through a proper validation technique, that is, a 3-map comparison technique. It is done by cross-tabulating the t_2 simulated LUC map with the t_1 and t_2 reference LUC maps. It is also worth mentioning that a validation process does not evaluate the model itself, rather the quality of the calibration undertaken on the model. Thus, the intention of the three ratio indices is not to literally evaluate models, but to examine and screen modeling outputs based on the quality of the calibration undertaken.

3.4. Minimum accuracy standard

Another issue that the LUC change modeling profession faces is the generalization of the modeling accuracy standard. It has been argued that a universal definition of good or acceptable can be misleading (Pontius *et al*., 2007). Hence, the accepted level of accuracy for any LUC change modeling output may vary according to the underlying purpose (Vliet *et al*., 2011). However, this should not be a hindrance to researchers and scientists to explore and challenge the issue. While we also agree that a criterion (accuracy level) for validating any LUC change modeling output should be defined in relation to the purpose of the research, this argument also causes tremendous problems, if not worse than those in having a minimum accuracy standard. It is because the acceptability of the accuracy of any modeling output will always be subjective. The problem is that there is no common appreciation among us as to the level of accuracy that a particular purpose should have, or the type of purpose that would require a certain level of accuracy. If one individual wants to model LUC change, then we believe this person is interested on the "change", which should be the case. In this context, we do not see any other "major purpose" other than to model the "change".

Since the three indices are applicable to any LUC change modeling that produces hard predictions or simulations,

we argue that they have the potential to be used in setting a minimum accuracy standard for this type of modeling. For example, the HOC ratio can be compared with the MOC and FOC ratios through the concept of "preponderance of evidence". For a particular LUC change modeling result to meet the minimum accuracy standard, its HOC ratio should always be higher than the individual ratio of the MOC and FOC. A practical reason for this is that it is unlikely that planners or policy-makers will be convinced to use a model that has a calibration level not capable of producing a HOC that is higher than its MOC and FOC. If this had been the case, only Models 2 and 3 of the three hypothetical models used to illustrate the validation concepts applied and examined in this paper (Fig. 1, Table 1) would have met the minimum accuracy standard. This implies that, hypothetically, Model 1 needs to be re-calibrated.

4. Concluding remarks

There is an urgent need for LUC change modelers and scientists around the world to address the issues presented, *viz*. focus of accuracy assessment, parameters to indicate overall accuracy, parameters for comparing different modeling results, and the minimum accuracy standard. In this paper, we introduced new measures of accuracy, being the HOC, MOC and FOC ratio indices, and discussed their potentials in contributing to the resolutions of the issues mentioned. The immediate future plan for this study is to further examine the potentials and possible limitations of the proposed new measures of accuracy.

Acknowledgement

We thank the reviewers for their valuable comments and suggestions on an eanlier version of the paper.

References

Ahmed, B., and Ahmed, R. 2012. Modelling urban land-cover growth dynamics using multi-temporal satellite images: A case study of Dhaka, Bangladesh. *ISPRS International Journal of Geo-Information*, **1**(1): 3–31.

Chen, H., and Pontius Jr., R. G. 2010. Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. *Landscape Ecology*, **25**(9): 1319–1331.

Guan, D. J., Li, H. F., Inohae, T., Su, W., Nagaie, T., and Hokao, K. 2011. Modelling urban land use change by the integration of cellular automaton and Markov model. *Ecological Modelling*, **222**(20–22): 3761–3772.

Huang, Q., and Cai, Y. 2007. Simulation of land use change using GIS-based stochastic model: The case study of Shiquian County, Southwestern China. *Stochastic Environmental Research and Risk Assessment*, **21**(4): 419–426.

Kamusoko, C., Oono, K., Nakazawa, A., Wada, Y., Nakada, R., Hosokawa, T., Tomimura, S. *et al*. 2011. Spatial simulation modelling of future forest cover change scenarios in Luangprabang Province, Lao PDR. *Forests*, **2**(3): 707–729.

Khoi, D. D., and Murayama, Y. 2010. Forecasting areas vulnerable to forest conversion in the Tam Dao National Park Region, Vietnam. *Remote Sensing*, **2**(5): 1249–1272.

Klug, W., Graziani, G., Grippa, G., Pierce, D., and Tassone, C. (eds). 1992. *Evaluation of long range atmospheric transport models using environmental radioactivity data from the Chernobyl accident: The ATMES Report*. Elsevier, London, 366 p.

Perica, S., and Foufoula-Georgiou, E. 1996. Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions. *Journal of Geophysical Research*, **101**(D21): 26347–26361.

Pontius Jr., R. G., 2000. Quantification error versus location error in the comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, **66**(8): 1011–1016.

Pontius Jr., R. G., Boersma, W., Castella, J. C., Clarke, K., de Nijs, T., Dietzel, C., Duan, Z., *et al*. 2008. Comparing the input, output, and validation maps for several models of land change. *Annals of Regional Science*, **42**(1): 11–47.

Pontius Jr., R. G., and Malanson, J. 2005. Comparison of the structure and accuracy of two land change models. *International Journal of Geographical Information Science*, **19**(2): 243–265.

Pontius Jr., R. G., and Millones, M. 2011. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, **32**(15): 4407–4429.

Pontius Jr., R. G., Walker, R., Yao-Kumah, R., Arima, E., Aldrich, S., Caldas, M., and Vergara, D. 2007. Accuracy assessment for a simulation model of Amazonian deforestation. *Annals of the Association of the American Geographers*, **97**(4): 677–695.

Sloan, S., and Pelletier, J. 2012. How accurately may we project tropical forest-cover change? A validation of a forward-looking baseline for REDD. *Global Environmental Change-Human and Policy Dimensions*, **22**(2): 440–453.

Thapa, R. B., and Murayama, Y. 2011. Urban growth modelling of Kathmandu metropolitan region, Nepal. *Computers, Environment and Urban Systems*, **35**(1): 25–34.

Vliet, J. V., Bregt, A. K., and Hagen-Zanker, A. 2011. Re-

visiting Kappa to account for change in the accuracy assessment of land-use change models. *Ecological Modelling*, **222**(8): 1367–1375.

Received 2 August 2012
Accepted 24 October 2012