ESTIMATING THE LAND TRANSITION MATRIX BASED ON ERRONEOUS MAPS

Robert Gilmore Pontius Jr[1] and Xiaoxiao Li[2]

[1]Clark University, Department of International Development, Community and Environment
[2]Purdue University, Department of Forestry and Natural Resources

Abstract

This article presents methods to estimate a land cover transition matrix based on maps from two points in time. The land cover transition matrix indicates the amount of land that transitions from each category at time 1 to each category at time 2. Observed differences between the two land cover maps can be due to change on the ground or error in the maps. If the maps were perfectly correct, then the observed differences would indicate true land transition on the ground. This paper considers the situation when the maps are not correct, and a confusion matrix indicates the structure of the errors in each map. For situations where formal confusion matrices are not available, we perform sensitivity analysis to show how the suspected error in the maps influences the estimates of the land cover transitions. We illustrate the technique using land cover data from 1971 and 1999 in the Plum Island Ecosystems of Northeastern Massachusetts, which is a Long Term Ecological Research site of the National Science Foundation. If the maps were perfectly correct, then the transition from forest to residential would account for 6% of the study area. Our method shows that if each category were to have a user's accuracy of 85 percent, then this transition would account for 7% of the study area, and could range from 4% to 8% depending on the assumptions concerning the distribution of errors in the maps. The method also produces maps that show the probability of any particular land cover transition, given the observed data and the confusion matrix.

Keywords

accuracy, land-use, land-cover, change, matrix, uncertainty.

1 Introduction

Land change and map error are two factors that lead to observed differences between two maps of the same extent for two different points in time. Researchers are becoming increasingly interested in assessing the accuracy of maps, because maps that are integrated into a GIS database may show a large number of erroneous changes, since error at either date can give a false impression of change (Khorram 1999, Lunetta and Elvidge 1999, Foody 2002, Yang and Lo 2002, Liu and Zhou 2004, Mas 2005). If we simply ignore the error, then we risk having inaccurate estimates for the amount of change on the land as expressed by a land transition matrix. A confusion matrix is a common method to express map accuracy, but such information does not exist in many situations. Therefore, this paper addresses two questions: 1) How do we estimate the land transition matrix when we have two maps and their confusion matrices? 2) How do we estimate the land transition matrix when we have two maps but do not have their confusion matrices?

2 Methods

This article illustrates the methods with a case study of the Plum Island Ecosystems (PIE) site in northeastern Massachusetts, USA, which is part of the National Science Foundation's Long Term Ecological Research program. There exist raster maps for 1971 and 1999 for four categories:

Forest, Residential, Open and Other. Each pixel at the 30-meter resolution has full membership to exactly one of the four categories. There does not exist information concerning the accuracy of these maps.

The top number in each cell of table 1 shows the observed differences between the maps of 1971 and 1999 summarized in the form of a transition matrix. This remainder of this methods section describes how to compute the other three entries in each cell of table 1 based on various assumptions about possible errors in the maps.

*Table 1. Transition matrices for the PIE case study. Matrix **D** is the top row in each cell, **U** is the second row, **M** is the third row, **L** is the bottom row. All entries express percent of the study area.*

		Map 1999				
		Forest	Residential	Open	Others	Total (1971)
Map 1971	Forest	**39** **28** **29** **39**	**6** **9** **7** **5**	**0** **2** **2** **1**	**2** **5** **5** **2**	**48** **44** **43** **46**
	Residential	**0** **3** **3** **0**	**21** **15** **16** **21**	**0** **1** **1** **0**	**0** **2** **2** **0**	**21** **21** **22** **22**
	Open	**1** **2** **2** **1**	**1** **2** **2** **1**	**2** **2** **2** **2**	**1** **2** **2** **1**	**4** **8** **8** **4**
	Others	**0** **3** **3** **1**	**2** **4** **4** **2**	**1** **2** **2** **1**	**25** **18** **18** **25**	**27** **27** **27** **28**
	Total (1999)	**40** **37** **37** **40**	**30** **30** **29** **29**	**3** **7** **8** **4**	**27** **27** **27** **27**	**100** **100** **100** **100**

Figure 1 illustrates the logic of the method to estimate the transition matrix. "Map Time 1" and "Map Time 2" are the maps of 1971 and 1999 respectively for the case study. Matrix **D** shows the observed correspondence between the maps in the format of table 1, where agreement is on the diagonal and difference is off the diagonal. If the maps were perfectly accurate, then matrix **D** would express the transitions on the ground, since matrix **D** summarizes the direct overlay of the map from time 1 on the map from time 2. However, the purpose of this paper is to estimate the transition matrix for the case where the maps have errors. Therefore, we need to know the structure of the errors in the maps. **C1** and **C2** express these errors in the form of confusion matrices where the rows show the categories in the maps and the columns show the same categories for some type of validation information, such as ground information. The entries in matrices **C1** and **C2** are the user's conditional probabilities, meaning that the entry in row *i* and column *k* of matrix **Ct** is the probability that a pixel is category *k* in truth at time *t*, given that the map shows it as category *i* at time *t*. Our method combines information from matrices **D**, **C1** and **C2** to estimate the land transitions, expressed by the matrix at the bottom of figure 1.

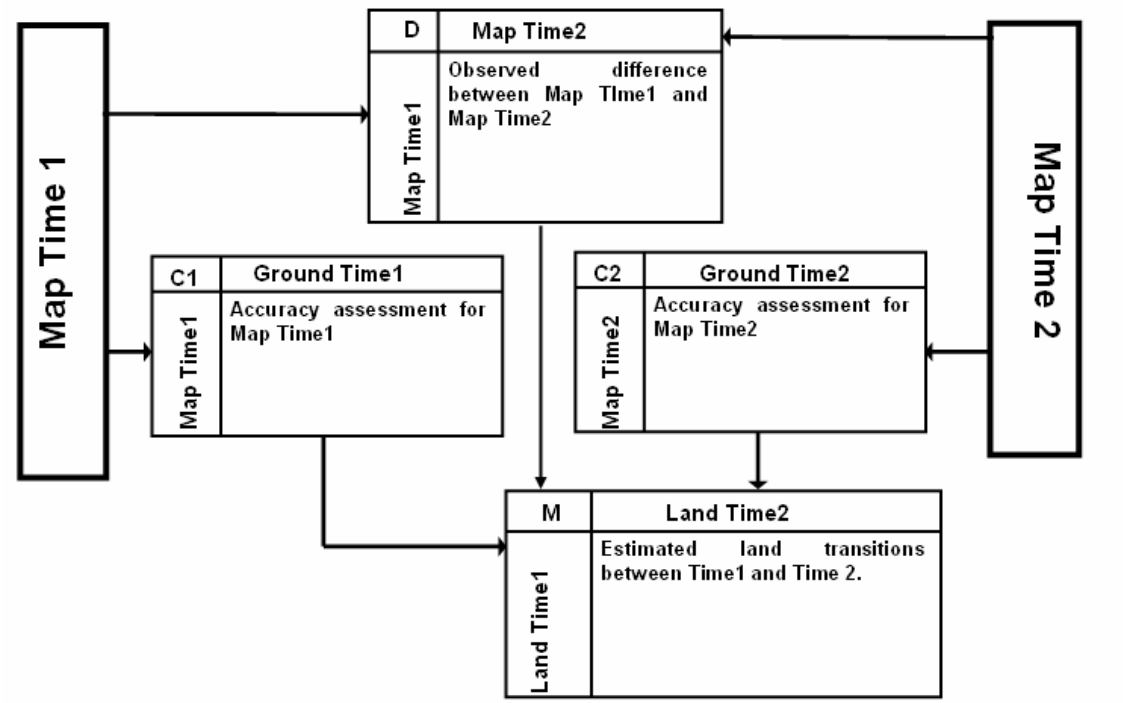

*Figure 1. Logic of methods showing flows of information from matrices **D, C1**,and **C2** to **M**.*

We consider three different methods to estimate the land transition matrix based on the overlay of the maps from times 1 and 2. All three methods are based in part on an assumption that the accuracy of the resulting overlaid land change classification is equal to the accuracy obtained by multiplying the accuracies of each individual classification (Mas 1999, Stow 1999, Fuller et al. 2003). The first method applies this assumption to all of the pixels in the study area and produces matrix **M**. The second method applies the multiplication assumption to only the pixels that show disagreement between times 1 and 2, while it assumes no error in the pixels that show persistence over time. This second method estimates the land transitions in a matrix denoted as **L**, which is designed to offer a lower bound on the estimated change. The third method applies the multiplication assumption to only the pixels that show agreement between times 1 and 2, while it assumes no error in the pixels that show difference over time. This third method produces matrix **U**, which is designed to offer an upper bound on the estimated land change.

Ideally, the confusion matrices would derive from an empirically-based independent accuracy assessment, however it is common that information about map accuracy does not exist. For this situation, we apply sensitivity analysis concerning a range for plausible accuracies to consider a variety of possibilities for matrices **C1** and **C2**. In our case study, this range is from 0.85 to 1.00 in terms of proportion correct. For any particular selection of the overall accuracy, we generate an entire confusion matrix. We do this by setting the assumed overall accuracy equal to the user's accuracy for each category, meaning that each of the diagonal entries in the confusion matrix is set equal to the assumed overall proportion correct. Then the remaining overall proportion incorrect is distributed equally among the off-diagonal entries within each row of the confusion matrix to give the same commission error for each category (Pontius and Lippitt 2006). For example, if the assumed overall accuracy of the map is 0.85, then 0.85 is the value for each diagonal entry in the confusion matrix. The resulting implied proportion incorrect of 0.15 is

distributed equally among the other categories in each row of the confusion matrix. In our case study that has three other categories, all off-diagonal entries in the confusion matrices are 0.05.

All of the mathematical notation and equations are given and explained below:

$t \equiv$ time such that $t = 1$ or $t = 2$,

$i \equiv$ index for a category in a map,

$j \equiv$ index for a category in a map,

$k \equiv$ index for a category in truth,

$J \equiv$ number of categories in the study area,

$d_{ij} \equiv$ entry in row i and column j of matrix **D** that gives the percent of the study area that is classified as category i in the time 1 map and classified as category j in the time 2 map,

$c_{tik} \equiv$ entry in row i and column k of the confusion matrix for time t denoted **Ct**, which gives the conditional probability that a pixel in truth at time t is category k, given that the map at time t shows it as category i,

$At \equiv$ assumed user's accuracy of all categories in the confusion matrix, which appears on all diagonal entries in matrix **Ct** for cases where ground information is not available,

$Ft \equiv$ value of off-diagonal entries in matrix **Ct** when assumed user's accuracy is At,

D ≡ J-by-J difference matrix that shows categories of the time 1 map in the rows and categories of the time 2 map in the columns in terms of percent of the study area,

Ct ≡ J-by-J confusion matrix that shows categories of the map at time t in the rows and categories of truth at time t in the columns in terms of probabilities,

Vti ≡ 1-by-J row vector that is row i of the confusion matrix **Ct**, which gives is the probability of a pixel being category k at time t in truth, given that it is category i at time t in the map,

Hij ≡ J-by-J matrix where each entry is the probability that a pixel transitions in truth from the category in the row to the category in the column, given that the empirical maps show category i at time 1 and category j at time 2,

Eij ≡ J-by-J matrix where the entry in row i column j is one and all other entries are zero.

M ≡ J-by-J transition matrix that gives the middle estimate for the percent of the study area that transitions from the category in the row to the category in the column,

L ≡ J-by-J transition matrix that gives the lower estimate for the percent of the study area that transitions from the category in the row to the category in the column,

U ≡ J-by-J transition matrix that gives the upper estimate for the percent of the study area that transitions from the category in the row to the category in the column,

The following equations hold:

$$\sum_{k=1}^{J} c_{tik} = 1 \qquad \text{equation 1}$$

$$Ft = \frac{1 - At}{J - 1} \qquad \text{equation 2}$$

$$\mathbf{Hij} = \mathbf{V1i}^{\mathrm{T}} \times \mathbf{V2j} \qquad \text{equation 3}$$

$$\mathbf{M} = \sum_{i=1}^{J}\sum_{j=1}^{J}\left(d_{ij} \times \mathbf{Hij}\right) \qquad \text{equation 4}$$

$$\mathbf{L} = \sum_{i=1}^{J}\left(d_{ii} \times \mathbf{Eii}\right) + \sum_{i=1}^{J}\sum_{j=1}^{J}\left(d_{ij} \times \mathbf{Hij}\right) \quad \text{for } i \neq j \qquad \text{equation 5}$$

$$\mathbf{U} = \sum_{i=1}^{J}\left(d_{ii} \times \mathbf{Hii}\right) + \sum_{i=1}^{J}\sum_{j=1}^{J}\left(d_{ij} \times \mathbf{Eij}\right) \quad \text{for } i \neq j \qquad \text{equation 6}$$

Matrix **Ct** is the *J*-by-*J* confusion matrix for time *t*, where its *J* rows are *J* vectors, each denoted as **Vti**. The superscript T on **Vti** means the transpose of vector **V1i**, which converts it from a row vector to a column vector. Therefore matrix algebra produces matrix **Hij** as a *J*-by-*J* matrix, and there are J^2 such matrices. The design of matrix **Hij** indicates that if a pixel is category *i* in the map of time 1 and category *j* in the map of time 2, then c_{1ix} multiplied by c_{2jy} is the probability that the pixel transitioned in truth from category *x* at time 1 to category *y* at time 2. Therefore, the entries of matrix **Hij** give the conditional probability that a pixel transitioned in truth from the category in its row to the category in its column, given that the maps show that the pixel transitioned from category *i* to category *j*. Consequently, all the entries in each matrix **Hij** sum to one.

Matrix **M** is a weighted average of all such J^2 matrices, where each **Hij** is weighted by the percent of the particular observed transition in the study area, given by entries d_{ij} in **D**. Matrices **L** and **U** are similar to matrix **M**, in the respect that they estimate the land transitions, albeit with different assumptions concerning which pixels have errors. Matrix **L** assumes that error exists only in the pixels that show difference between the maps of times 1 and 2, whereas matrix **U** assumes that error exists only in the pixels that show persistence over time. Equation 5 computes **L** as a weighted sum of **Eii** to reflect the pixels that show agreement, plus a weighted sum of **Hij** to reflect the pixels that show disagreement for which $i \neq j$. Equation 6 computes **U** as a weighted sum of **Hii** to reflect the pixels that show agreement, plus a weighted sum of **Eij** to reflect the pixels that show disagreement.

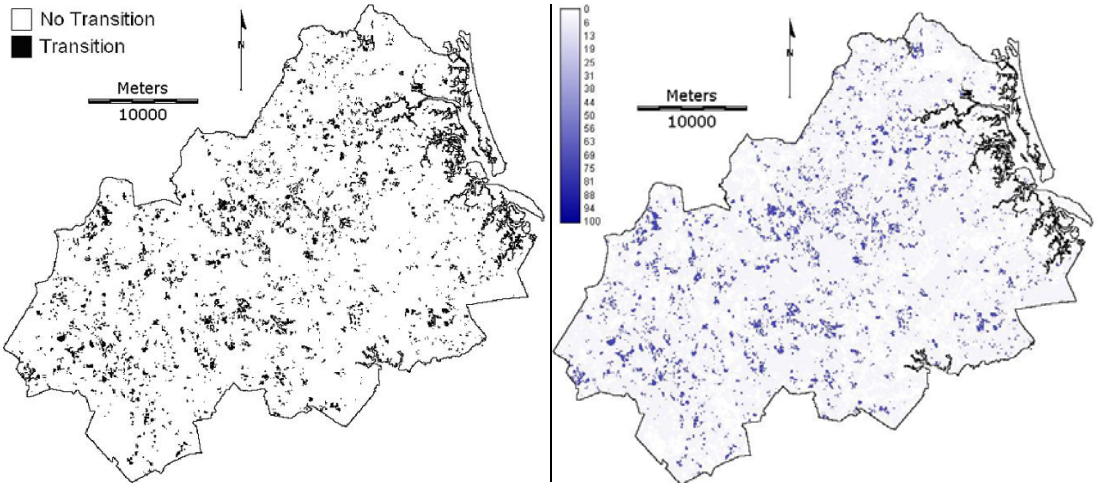

Figure 2. PIE maps showing: (a) observed Boolean transition from Forest to Residential on the left, and (b) probability of transition from Forest to Residential on the right.

Matrix **Hij** can be used to produce a map that shows the probability of any particular transition, given the overlay of the maps from times 1 and 2. Figure 2a shows the apparent transition from Forest to Residential in black and all other areas in white, based on a simple overlay. We use the J^2 entries in matrix **Hij** to convert each pixel in the overlaid map into a probability of transition from Forest to Residential. Figure 2b shows the spatial distribution of the probability of this transition expressed as a percent from 0 to 100. Figure 2b reflects the information in **Hij** that was based on an assumed user's accuracy of $At = 0.85$, where i = the index for Forest and j = the index for Residential.

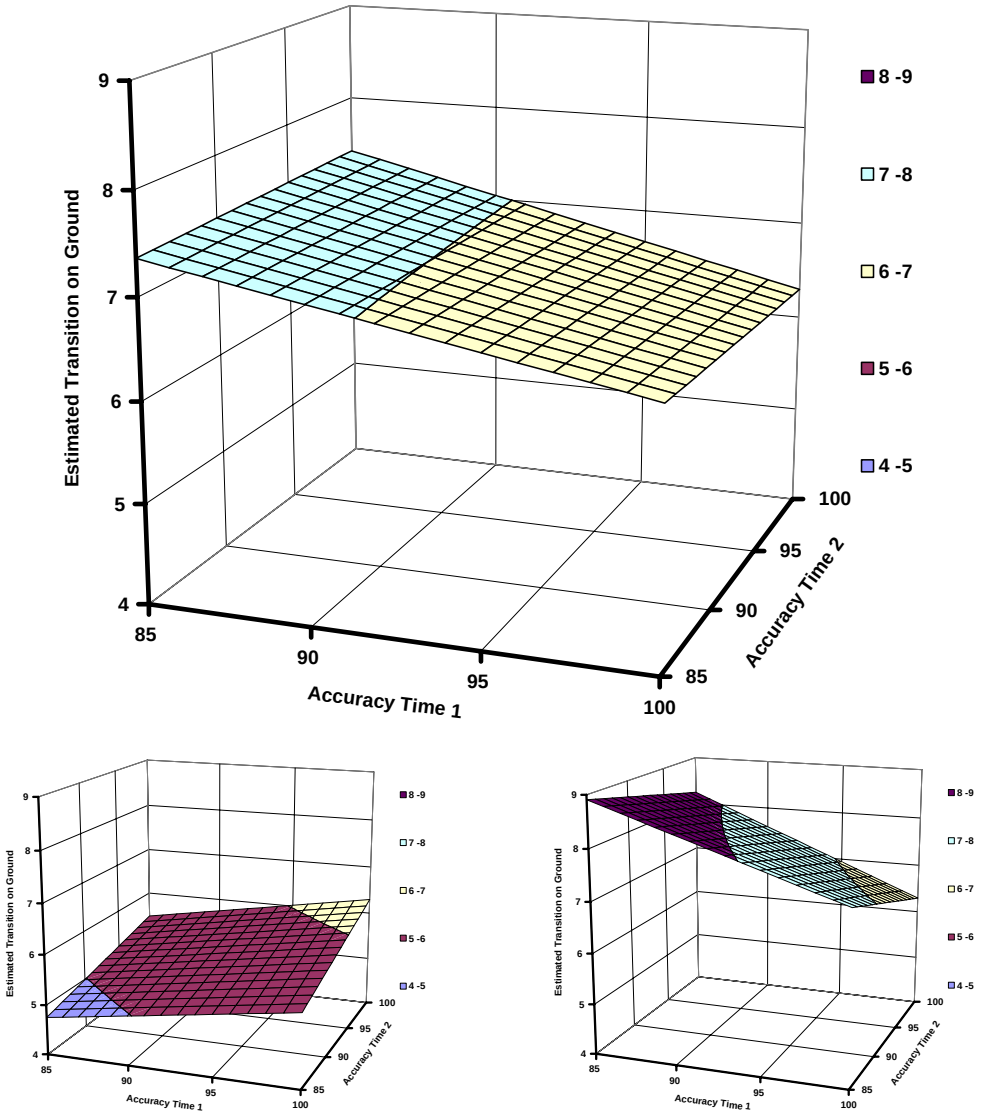

Figure 3. Sensitivity of estimated percent of study area that transitions from Forest to Residential as a function of percent correct in maps of time 1 and 2. Figure 3a at top gives middle estimate; figure 3b at bottom left gives lower estimate; figure 3c at bottom right gives upper estimate.

Figure 3 illustrates the type of information we can obtain by applying sensitivity analysis to the parameters *A1* and *A2*. All three plots have the same axes. The z-axis is the estimated transition from Forest to Residential expressed as percent of the study area. The x-axis is the assumed accuracy of the map of time 1, i.e. *A1*, ranging from 85 to 100 percent correct. Similarly, the y-axis is the assumed accuracy of the map of time 2, i.e. *A2*. Matrix **M** is the basis of the top figure 3a; matrix **L** is the basis of the bottom left figure 3b; matrix **U** is the basis of the bottom right figure 3c. When accuracy is 100 percent, the surfaces for matrices **M, L,** and **U** intersect at a single point, which is the amount of the transition given in matrix **D.** The surface for matrix **M** is between the lower surface of matrix **L** and the upper surface of matrix **U**.

3 Results

Table 1 gives the four types of estimated transition matrices assuming 85 percent accuracy for both maps. Figure 2 shows the implications visually for the transition from Forest to Residential. Darker shading represents higher probability of a transition from Forest to Residential. Figure 3 shows the results of the sensitivity analysis for the transition from Forest to Residential for accuracies ranging from 85 to 100 percent for both maps. If the maps were perfectly correct, then they would indicate that 6 percent of the study area transitioned from Forest to Residential between 1971 and 1999. If the user's accuracy were 0.85 for each category in both maps, then the transition from Forest to Residential would be 7 percent based on the middle estimate, 5 percent based on the lower estimate, and 9 percent based on the upper estimate.

4 Discussion

The proposed method is based on simplifying assumptions, just as all types of analysis are. First, it describes the map error in the form of a confusion matrix, which lacks information about possible spatial dependencies in the errors. Next, the technique applies a single user's accuracy to all categories in each confusion matrix. Then, equation 2 applies an identical probability of confusion with each of the other categories within each row of the confusion matrices. Furthermore, equation 3 assumes that the errors in the map of time 1 are distributed independently from the errors in the map of time 2, which implies no temporal dependence. These assumptions about the distribution of error are unlikely to match exactly the real error structure in the maps in at least four respects. First, it is likely that there is some spatial dependence in the errors, since some regions of the study area are likely to be more difficult to classify than others. Second, it is common for some categories to be more accurate than other categories. Third, some categories are more likely to be more confused with similar categories than with dissimilar categories (Rogan and Chen 2003). Fourth, if spatial dependence among the errors persists over time, then it is likely to cause temporal dependence in the errors.

Other investigators have approached this problem with methods that rely on either detailed information about the error structure (van Oort 2005, 2007) or computationally intensive simulation methods (Burnicki et al. 2007). These other approaches can be helpful to illustrate the possible range of effects that various assumptions can have on the estimates of land change, but they require either more information than is typically available or additional assumptions concerning the details of the distributions of the errors in the maps. In this paper, we have intentionally taken an approach that is mathematically and conceptually simpler than other proposed approaches, because we intend for our approach to be as intellectually accessible as possible, in spite of the fact that proper interpretation of the results still requires careful attention.

5 Conclusions

This article proposes methods: 1) to produce three types of land transition matrices based on the comparison of two possibly erroneous maps over time, 2) to generate maps of the probability of any particular land transition, and 3) to perform sensitivity analysis concerning a range of plausible levels of map accuracies. The methods are based on simplifying assumptions that allow us to express the calculations in six equations. The procedure has been designed for either the case where there exist confusion matrices that express the accuracy of the maps or the case where there does not exist information concerning map accuracy. This procedure offers scientists an alternative to ignoring potential error in maps when comparing maps that share a categorical variable.

Acknowledgements

The National Science Foundation (NSF) supported this work via three of its programs: 1) Human-Environment Regional Observatory program via grant 9978052, 2) Long Term Ecological Research via grant OCE-0423565, and 3) Coupled Natural-Human Systems via grant BCS-0709685. Opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect those of the NSF. The Massachusetts Geographic Information System supplied the original data. Clark Labs facilitated this work by creating the GIS software Idrisi®.

Literature

Burnicki, A., Brown, D.G. and Goovaerts, P., (2007) Simulating error propagation in land-cover change analysis: The implications of temporal dependence, *Computers, Environment and Urban Systems*, 31, pp. 282-302.

Foody, G.M., (2002) Status of Land Cover Classification Accuracy Assessment, *Remote Sensing of Environment*, 80(1), pp. 185-201.

Fuller, R.M., Smith, G.M. and Devereux, B.J., (2003) The characterization and measurement of land cover change through remote sensing: problems in operational applications?, *International Journal of Applied Earth Observation and Geoinformation*, 4, pp. 243-253.

Khorram, S. (ed.), (1999) *Accuracy assessment of remote sensing-derived change detection*, Bethesda MD, American Society for Photogrammetry and Remote Sensing.

Liu, H. and Zhou, Q., (2004) Accuracy analysis of remote sensing change detection by rule-based rationality evaluation with post-classification comparison, *International Journal of Remote Sensing*, 5(25), pp. 1037-1050.

Lunetta, R.L. and Elvidge, C.D. (eds.), (1999) *Remote Sensing change detection: environmental monitoring methods and applications*, London: Taylor and Francis.

Mas, J.F., (1999) Monitoring land-cover changes: a comparison of change detection techniques, *International Journal of Remote Sensing*, 20(1), pp. 139-152.

Mas, J.F., (2005) Change Estimates by Map Comparison: A Method to Reduce Erroneous Changes Due to Positional Error, *Transactions in GIS*, 9(4), pp. 619-629.

Pontius Jr, R.G. and Lippitt, C.D., (2006) Can Error Explain Map Differences Over Time?, *Cartography and Geographic Information Science*, 33(2), pp. 159-171.

Rogan, J. and Chen, D., (2003) Remote sensing technology for mapping and monitoring land-cover and land-use change, *Progress in Planning*, 61(4), pp. 301-325.

Stow, D.A., (1999) Reducing misregistration effects for pixel-level change detection, *International Journal of Remote Sensing*, 20(12), pp. 2477-2483.

Van Oort, P.A.J., (2005) Improving land cover change estimates by accounting for classification errors, *International Journal of Remote Sensing*, 26(14), pp. 3009-3024.

Van Oort, P.A.J., (2007) Interpreting the change detection error matrix, *Remote Sensing of Environment*, 108, pp. 1-8.

Yang, X. and Lo, C.P., (2002) Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area, *International Journal of Remote Sensing*, 23(9), pp. 1775-1798.