Assessing the Robustness of Predictions in Spatially Explicit Models of Land Use[1]

Alessandro De Pinto and Gerald C. Nelson

June 1, 2006

Abstract:

We propose an information-theoretic approach to assess the performance of a discrete choice model used to analyze land use and land use change. We show that our disaggregated measure can be used to compare robustness of predictions across land use categories and across models. Furthermore, a proper reformulation of the problem shows that a disaggregated (observation by observation) log-likelihood lends itself to an information theoretic interpretation, which allows comparisons performance across models.

Corresponding author:

Alessandro De Pinto, Assistant Professor, Department of Economics, University of Redlands.

---

[1] Selected paper prepared for the American Agricultural Economics Association annual meeting, Long Beach, California, July 23 - 26, 2006.

The authors are solely responsible for the analysis and conclusions presented here.

Introduction

Spatially explicit models of land use and land use change focus on the statistical estimation of the determinants of land use or the estimation of transition probabilities of land use units. One of the challenges is testing performance. The variable to be explained, land cover or land use, is discrete and therefore the usual regression measure, R^2, with its useful interpretation of accounting for unexplained variation, does not apply.

Arguably the most common test used to assess goodness-of-fit of a model is the likelihood ratio index (McFadden, 1974; Ben-Akiva and Lerman, 1985). This index is often used as a pseudo R^2 since its value is one if the model predicts perfectly and zero if it has the predictive power of the null model. The problem with this measure is that there is no direct correspondence between the value of the index and accuracy of prediction: a likelihood ratio index value close to one this does not necessarily mean that the model predicts well. But it is precisely with predictions that we are most concerned because predicting land use change is one of the most appealing features of discrete choice models that use spatially explicit data.

Other commonly used measures of performance indices are derived from the map accuracy literature. These include user and producer accuracy, the kappa statistic (Nelson and Hellerstein (1997), Walker (2003)) and the confusion matrix. The rationale behind the use of these indices is that a "good" model is one that has a high ratio of correctly predicted choices.

The central problem with these measures of performance is that they are based on predictions, typically using a winner-take-all approach. However, the output of a discrete choice model is a series of probability values, one for each category, and, as Train (2003) points out, the researcher can only make claims on the probability of a choice to be made. When a prediction is made, as is necessary to construct all prediction-based performance measures, the probabilistic nature of the

discrete choice model output is severely distorted because the probabilities of the other possible choices collapse to zero. An observation can be assigned to a particular choice category even when the predicted probability value for that category is low, as long as the probability values for the other categories are lower. As a consequence, the information regarding how strong a prediction is, the uncertainty present in a prediction, is lost. Moreover, the prediction – the assignment to one choice category – depends entirely on how probabilities are translated into predictions. Each of the possible applicable rules to map probabilities into predictions is essentially arbitrary.

Assessing the strength of predictions: a first attempt and shortcomings

A researcher with perfect knowledge of the underlying decision process that relates exogenous variables to the land use choice could construct a model that predicts each of the choices observed with probability 1 and the choices that are not made with probability 0. With this "perfect" model as a benchmark, it is intuitive to consider one model superior to an alternative if it predicts observed choices with a higher probability than the alternative model.

This idea is the rationale behind a comparison of the performance of three discrete choice model specifications by Nelson et al. (2004). The authors compare what they call the "strength" of predictions by computing the average probability value of the predicted land use across models. However, this procedure is not completely accurate because the term "strength" is misleading. The problem is better framed in terms of the uncertainty present in a prediction. In that case, the uncertainty is dependent on the whole probability output and cannot be assessed by looking only at the category that obtains the highest probability value.

As an example, consider a location with seven possible land use categories. For a particular location, a model generates the following probabilities: [0.7, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05]. When a winner-take-all prediction rule is used, the first category is chosen and information regarding the 0.3 probability of making a mistake is lost. This particular probability outcome is also interesting because for the researcher interested in the consequences of making a mistake, there is not much more information that can be gained. The remaining six categories all have equal predicted probability. This is rather different from a situation where the predicted probabilities are [0.7, 0.299, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002]. There is still a 0.3 probability of making a mistake with the winner-take-all rule. However, if a mistake is made it is most likely at the expense of the second category. In other words there is a greater uncertainty in the first prediction than in the second.

Entropy as a quantitative measure of uncertainty

In an attempt to exploit the information contained in *all* the predicted probability values, we turn to the concept of entropy. This rather widely used concept has its roots in both thermodynamics and information theory. We use an interpretation from the information theory literature. Define X as the set of *n* possible events and $P = \{p_1 ..... p_n\}$ the probability of occurrence of each event in x_n, where $\sum_{i=1}^{n} p_i = 1$. The occurrence of each event x_n is said to carry an amount of information $I(x_n)$ that is dependent on the initial probability scheme. Suppose that a certain event occurs with probability .999. One would be hardly surprised when receiving a message that this event has indeed occurred. This message has very little information content. In contrast, if an event has an occurrence probability of .001 a message that tells us that that event occurred has high information content. It is evident that the information content of the message is a decreasing

function of the probability *p*: the more unlikely the probability of the event before the message of its occurrence, the larger the information content of the message.

In principle, any decreasing function in *p* could be chosen to describe this relationship between prior probability and information, but traditionally the function used is the logarithm of the reciprocal of the probability *p*: $I(x_n) = \log \frac{1}{p_n} = -\log p_n$.

Using this function it also possible to measure the uncertainty associated with a probability scheme $\{p_1 ..... p_n\}$ *before* any of the possible events takes place $H(X) = -\sum_{i=1}^{n} p_i \log p_i$, where the function $H(X)$ is called entropy and can be interpreted as the average amount of information carried by events with a certain probability scheme. In this paper we normalize our entropy measure by using

$$H(X) = \frac{-\sum_{i=1}^{n} p_i \log p_i}{\log n}, \qquad (1)$$

thereby ensuring that the value of entropy is bounded below by zero and above by unity.

If one of the possible events occurs with probability 1 there is no uncertainty in the probability scheme ($H(X) = 0$). On the other hand, if each event is equally probable the uncertainty reaches its highest value ($H(X) = 1$)[2].

---

[2] In information theory the base of the logarithm is traditionally 2. So that the information associated with the occurrence of one of two equally probable events x_1 and x_2 is 1 (1 bit):

$I(x_1) = I(x_2) = -\log 1/2 = 1$.

If we go back now to the example provided in the previous section and measure the entropy of the two probability schemes we can see how the difference in uncertainty is fully captured by the entropy measure. The entropy is 0.59 for the first set of probabilities and 0.32 for the second set. In the next section we will use the concept of entropy as an exploratory tool. We will compare the performance of each model by looking at the uncertainty present in those predictions that a winner-take-all rule assigned to the correct category. A model with little or no uncertainty in those correct predictions is preferable to a model that has a greater uncertainty.

An empirical application

As an application of this new proposed method we recreated the results obtained by Nelson et al. (2004). The objective of that study was to simulate the overall land use change and its location caused by a proposed road improvement project in Panama's Darién province. The authors compared the performance of three specifications: Multinomial Logit (MLogit), Nested Logit (NLogit), and Random Parameters Logit (RPLogit).

The authors used 2,555 observations to estimate the parameters of interest. These estimated parameters are then used to generate probability values and winner-take-all predictions for the remaining 63,894 observations in the data set. The diverse landscape present in the area is aggregated into seven categories of land use: three different types of forest, two categories of human intervention, and two of natural vegetation. The authors use a variety of measures of predictive power in the attempt to select the best model:

- a pseudo-R^2
- a probability-based measure of the power of prediction for each category and
- prediction matrices for the different models.

The pseudo R^2 values were 0.627 for the MLogit model, 0.639 for the NLogit model and 0.595 for the RPLogit model. The log-likelihood values (not reported in the paper) were: -1434.253, -1389.352, and -1557.168[3]. These results suggest the NLogit specification is best.

However, when the category-specific probability-based measure is used, the results are less clear. Table 1 shows the average probability value for all locations that were correctly predicted to be in a certain land use.

The NLogit makes stronger predictions in the first four categories and the MLogit in the remaining three. The RPLogit lags behind the other two specifications.

Table 1: From Nelson et al. (2004), Table 4. Average probability value for predicted categories

	MLogit	NLogit	RPLogit
Forest without cuipo (0)	0.937	**0.952**	0.936
Forest with cuipo (1)	0.822	**0.851**	0.797
Forest with cativo (2)	0.460	**0.463**	0.397
Agriculture (3)	0.386	**0.435**	0.392
Pasture (4)	**0.626**	0.621	0.578
Brush (5)	**0.488**	0.461	0.433
Marshes (6)	**0.707**	0.695	0.665

Note: The bold values indicate the model with the highest average probability value for the category in that row.

Table 2 reports the model comparisons using prediction matrices. Using the winner-take-all assignment rule, the MLogit predicts better than the other models in six out of seven categories, the MLogit predicts better than the others in category 5. The ratio of correct to total is also

[3] These are the log-likelihood values for the various models estimated with the sample of 2,555 observations.

slightly higher for the NLogit (0.799) than the other two models; 0.792 for the MLogit and 0.769 for the RPLogit. Since it is impossible to perform statistical inference on these numbers, we cannot really say that 0.799 is better than 0.792.

Table 2: -Prediction matrices by estimation method, (Rows are predictions, columns are actual)

Category Id	Total	0	1	2	3	4	5	6	Ratio, correct to total
Logit									
Forest without cuipo (0)	17,970	**16,887**	970	0	26	51	36	0	0.919
Forest with cuipo (1)	27,562	1,427	**23,533**	553	620	122	948	359	0.893
Forest with cativo (2)	217	1	32	**84**	48	18	31	3	0.092
Agriculture (3)	1,380	1	125	28	**775**	38	409	4	0.289
Pasture (4)	7,126	8	149	2	352	**4,323**	2,264	28	0.734
Brush (5)	7,067	9	1,330	132	798	1,234	**3,228**	336	0.450
Marshes (6)	2,572	36	209	111	61	100	257	**1,798**	0.711
Total	63,894	18,369	26,348	910	2,680	5,886	7,173	2,528	0.792
Nested Logit									
0	18,233								0.928
		17,044	1,095	-	26	32	36	-	
1	27,452	1,280	**23,713**	465	505	145	968	376	0.900
2	341	0	82	**148**	28	5	72	6	0.163
3	1,943	2	200	31	**1,052**	64	588	6	0.393
4	7,874	9	170	5	499	**4,581**	2,576	34	0.778
5	5,387	2	833	127	504	966	**2,674**	281	0.373
6	2,664	32	255	134	66	93	259	**1,825**	0.722
Total	63,894	18,369	26,348	910	2,680	5,886	7,173	2,528	0.799
Random parameters logit									
0	18,031	**16,890**	1,016	0	26	62	37	0	0.919
1	28,183	1,404	**23,243**	591	920	356	1,347	322	0.882
2	168	0	41	**40**	46	26	12	3	0.044
3	1,333	0	98	16	**784**	57	374	4	0.293
4	6,090	21	754	0	48	**3,276**	1,924	67	0.557
5	7,382	21	1,008	158	764	1,922	**3,142**	367	0.438
6	2,707	33	188	105	92	187	337	**1,765**	0.698
Total	63,894	18,369	26,348	910	2,680	5,886	7,173	2,528	0.769

Note: Diagonal (bold) cells indicate correct predictions. Source: Nelson et al. (2004), Table 4.

We now revisit these results using the concept of entropy introduced earlier. We computed the normalized entropy (equation 1) for each of the correctly-predicted observations (winner-take-all

assignment rule). Table 3 reports the average values of the normalized entropy for each category. Lower values indicate that predictions have a lower level of uncertainty. We focus our attention on the NLogit, which seemed to perform better than the other two models, and contrast its performance with the MLogit and RPLogit models. To do this, we used a *t*-test to determine whether the average entropy values for MLogit and RPLogit are significantly different from the values obtained for the NLogit.

Table 3: Share predicted correctly and average entropy values for each correctly-predicted land use category

	NLogit		MLogit		RPLogit	
	Share predicted correctly	Ave. Entropy	Share predicted correctly	Ave. Entropy	Share predicted correctly	Ave. Entropy
Land use 0	0.928	**0.047**	0.919	0.062***	0.919	0.064***
Land use 1	0.900	**0.181**	0.893	0.211***	0.882	0.242***
Land use 2	0.163	0.626	0.092	**0.594*** **	0.044	0.629
Land use 3	0.393	**0.669**	0.289	0.696***	0.293	0.692***
Land use 4	0.778	**0.448**	0.734	0.464***	0.557	0.490***
Land use 5	0.373	0.663	0.450	**0.650*** **	0.438	0.685***
Land use 6	0.722	0.379	0.711	**0.363*** **	0.698	0.425***

Note: The lower the average entropy number the less uncertainty, on average, that the average location is predicted correctly. Bold figures indicate lowest values for the category in that row. ***, **, and * indicate levels of significance at the 1, 5, and 10 percent level respectively that the value differs from the Nlogit value in the same row.

We start with a closer look at the results for the NLogit model. There is little correspondence between the number of observations correctly predicted and average uncertainty present in predictions. Consider category 2; the share of correct predictions is only 0.163, the worst ratio of any other category. Yet, the average entropy value is lower (better) than the value for category 3 and 5 and, although not indicated in the table, the difference is statistically significant. One interpretation is that the average correct prediction for land use 2 is more robust than land uses 3 and 5. Similarly, category 6 has a lower correct-to-total ratio than category 4 but the average entropy value is lower. On the other hand, average entropy values and share of correct predictions are much better for land uses 1 and 2 than for the other land uses.

Moving on to a comparison across model specifications, our results confirm that the RPLogit has the worst performance, a result with the earlier findings. The average entropy values are higher for all categories indicating a higher uncertainty in the predictions. One small but interesting exception is category 2. The NLogit model correctly predicts 148 pixels, more than three times the 40 predicted by the RPLogit. However, there is no statistically significant difference between the two measures of entropy.

The NLogit model seemed to outperform the MLogit when looking at the number of correct predictions: six categories out of seven exhibited a higher correct-to-total ratio. Our analysis tells a less clear story and is somehow more similar to the alternative analysis proposed by Nelson et al. shown in Table 1. When looking at the uncertainty present in the models predictions, the NLogit performs better in four categories (0, 1, 3, 4) while the MLogit does better in the remaining three (2, 5, 6).

An information-theoretic measure of goodness of fit, Kullback-Leibler divergence

The analysis we performed above has several limits but most importantly it provides information only on the uncertainty present in the observations that are correctly predicted in a certain category. It does not tell us anything about the observations that were predicted incorrectly in the same category and those that are in that category in reality but are predicted as one of the alternatives. A model that predicts *wrongly* with very little uncertainty is a poor model. Therefore, in order to truly compare the performance of various models we need a measure that rewards a model if predicts correctly with little or no uncertainty and penalizes it if predicts wrongly with little uncertainty.

In probability theory and information theory, Kullback-Leibler divergence (Kullback and Leibler, 1951) is a natural distance measure of an arbitrary probability distribution q from a "true" probability distribution p.

For probability distributions of a discrete variable the K-L divergence is defined to be:

$$KL(p;q) = \sum_i p_i \log \frac{p_i}{q_i} \qquad (2)$$

Equation 2 can be rewritten as:

$$KL(p;q) = \sum_i p_i \log p_i - \sum_i p_i \log q_i \qquad (3)$$

In information theory, Kullback-Leibler divergence can be interpreted as the loss of information for an individual when she uses distribution q instead of p. It is always nonnegative – with values that range from 0 to $+\infty$ –and zero only if $p = q$.

Each observed choice can be translated into a probability scheme where one of the choices has probability one of occurring and the others zero; following our example with seven categories

the probability values are [1, 0, 0, 0, 0, 0, 0]. The original probability scheme is transformed by the model into a probability scheme in which each choice has a probability of occurring greater than zero. We use the Kullback-Leibler divergence to measure, for each observation, the distance between the true probability distribution and the model output.

Note that the first term on the right hand side of equation 3 is the negative of the entropy for probability distribution *p* and that the entropy of the true probability distribution is always equal to zero. Therefore, equation 3 is reduced to:

$$KL(p;q) = -\sum_i p_i \log q_i \ . \qquad (4)$$

Table 4 shows the average value for the Kullback-Leibler divergence for each category. As before, we contrast the performance of the NLogit model with those of the MLogit and RPLogit models. We use a paired *t*-test to see whether the averages for MLogit and RPLogit differ significantly from the values obtained for the NLogit model.

Results show that the NLogit performs statistically significantly better than then the MLogit in three categories 2, 4, and 6, and that the MLogit performs statistically significantly better than the NLogit in category 0. There is no significant difference for categories 1, 3, and 5. These results confirm that the RPLogit performs worse than the other two in all categories.

Table 4: Average values of Kullback-Leibler divergence for each land use category

	NLogit	MLogit	RPLogit
Category 0	0.284	**0.268***	0.274
Category 1	**0.468**	0.469	0.541***
Category 2	**3.151**	3.334**	3.511***
Category 3	3.708	**3.533**	3.740
Category 4	**1.204**	1.361***	1.660***
Category 5	1.988	**1.986**	2.033
Category 6	**1.143**	1.303*	1.499***

Bold figures indicate lowest values.

***, **, and * indicate levels of significance at the 1, 5, and 10 percent level respectively. The test is for significant difference between the MLogit or RPlogit result and the Nlogit result for a category.

Finally, we computed the average Kullback-Leibler divergence across ALL categories for each model. This measure indicates that there is no significant difference between the performance of Nested and MLogit, and it shows that they are both superior to the RPLogit model.

Table 5: Average values of Kullback-Leibler divergence across land use category

	NLogit	MLogit	RPLogit
Overall K-L divergence	**0.854**	0.866	0.949***

Bold figure indicate lowest values.

***, **, and * indicate levels of significance at the 1, 5, and 10 percent level respectively.

General Discussion and Conclusions

There is a relatively wide range of discrete choice models that can be used to analyze issues related to land use and land use change. In some cases, theory or model limitations such in the case of the assumption of irrelevant alternatives for the multinomial logit, might guide the researcher through model selection. In other cases, for instance choosing a distribution for some of the parameters in a random parameters logit, the researcher is left without guidance. When theory or model limitations cannot be of assistance, researchers have used measures of predictive power as a tool for model selection. The use of and attention to predictions is also justified by the fact that the ability to make location-specific predictions of land use change is an important feature of discrete choice models that use spatial data. However, the use of predictions as in the case of confusion matrices has severe practical and theoretical limitations.

In this paper we have proposed an information theoretic measure for the comparison of model performances that acknowledges the probabilistic nature of discrete choice model outputs and accounts for the uncertainty present in a prediction. The concept of entropy can be used to compare across models the uncertainty present in those observations that were correctly predicted.

Applying this new method of assessment of predictive power to a previous study changed somewhat the identification of the best model. The analysis of log-likelihood values, pseudo-R^2, and correct predictions suggested that the nested logit performed better than the multinomial logit. The information theoretic approach shows no significant difference between the two models. Our analysis confirms that the random parameters logit had the worst performance among the three specifications.

Although the assignment rule remains arbitrary (winner-take-all), the researcher has now some information regarding how robust a prediction in a certain category is. Land uses are not equally important. If the researcher is concerned with agriculture encroaching on forested land, she should be particularly interested in having low uncertainty in the prediction of agriculture.

The analysis of those observations that are correctly predicted does not tell us much about the overall performance of each model. To achieve this goal we used another measure based on information theory, the Kullback-Leibler divergence. This measure allows us to assess model performance considering all observations, correctly and incorrectly predicted. For the analysis we reviewed, the Kullback-Leibler divergence measure confirms the results obtained using the entropy measure: there is no statistically significant difference in the performance of multinomial logit and nested logit and both are superior to the random parameters logit. This is in contrast to what the pseudo R^2 and prediction matrices would suggest, that the nested logit is the best model. We believe that the these new measures should become part of the standard toolkit of land use researchers with spatial data.

References

Ben-Akiva, M. and S. R. Lerman (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, The MIT Press.

Kullback, S. and R. A. Leibler (1951). "On information and sufficiency." Annals of Mathematical Statistics **22**(1): 79–86.

Nelson, G. C. and D. Hellerstein (1997). "Do Roads Cause Deforestation? Using Satellite Images in Econometric Analysis of Land Use." American Journal of Agricultural Economics **79**(1): 80-88.

Nelson, G. C., A. De Pinto, et al. (2004). "Land Use and Road Improvements: a Spatial Perspective." International Regional Science Review **27**(3): 297-325.

Walker, R. T. (2003). "Evaluating the Performance of Spatially Explicit Models." Photogrammetric Engineering and Remote Sensing **6**(11): 1271-1278.

Hauser, J. R. (1978). "Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach." Operation Research **26**: 406 - 421.

Wear, D. N. (1998). "Orginal Article: Land-Use Changes in Southern Appalachian Landscapes: Spatial Analysis and Forecast Evaluation." Ecosystems **1**(6): 575-594.