Verification of robust regression approach in land use classification without ground data: a case of terraced paddy development in Sapa, Vietnam

Yuzuru ISODA, HOANG Thi Thu Huong, NGUYEN An Thinh and Doo-Chul KIM

Abstract: For past land use classification of satellite images where ground data does not exist, we propose the application of robust logistic regression using land use data of irrelevant date as training data. To verify this approach, we conducted field survey at Sapa, Vietnam asking farmers when and where they developed their terraced paddy field, and the result is used as test data for verification.

Keywords: remote sensing, land use classification, robust regression, accuracy assessment

1. Introduction

Satellite images having accumulated over 40 years are invaluable source in monitoring land use changes. However, without past ground data that approximately matches with the date images were taken, it is challenging to classify the past land use and to verify that classification, due to differences in available bands, seasons, solar effect, atmospheric effects and so forth (Coppin et al., 2004). Unavailability of past ground data, in forms of aerial photo or land use map, is especially common in developing countries where rapid land use changes have been occurring in the last 40 years or so.

For this reason, we are proposing the use of robust regression approach in doing supervised classification with ground data of irrelevant date, i.e. using the recent ground data in identifying the past land use (Isoda et al., 2010). By applying the robust method, it should be possible to restore past land use with the available ground data provided that the majority of land use did not change.

Another challenge to past land use classification without relevant ground data is accuracy assessment of the classification result. How can we verify if the ground data did not exist in the first place? However difficult it is, estimated land use change cannot be used in further scientific research without accuracy assessment, at least until the method is established. This paper does just that. We conducted interview survey to the people of our study area and obtained witnesses to where and when the land use changed based on their memory. The challenges with such data source are that our test data would have margins of error in time and location.

2. Study area and background

Ethnic minorities such as H’Mong have developed terraced paddy on steep mountain ranges exceeding 1,000m asl. in Sapa District, Lao Cai Province in Northern Vietnam (Fig.1), however the history of terraced paddy may not be as old. According to interview in Summer 2009, terraced paddy development has become prominent since the 1980s. H’Mong population in Vietnam is increasing at 3% per annum (1989-99, Population Census), suggesting they have

Yuzuru ISODA
6-3 Aramaki-aza-Aoba Sendai, 980-8578 Japan
Graduate School of Science Tohoku University
E-mail: isoda@m.tohoku.ac.jp Phone: 022-795-6674

shifted their traditional semi-sedentary swidden agriculture to wet rice cultivation on terraced paddy, to respond to the rapid population increase. Our final goal is to assess whether terraced paddy cultivation is sustainable environmentally and socially, but the first step was to verify whether the aforementioned parole evidence is true, and to find out how much and where the recent terraced paddy development have occurred.

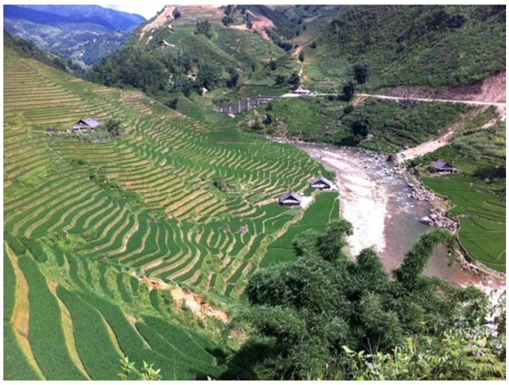


Fig 1. Terraced paddy in Lao Chai Commune

Our study area, Lao Chai Commune in Sapa (554 households, 3585 population) is primarily occupied by H'Mong people, predominantly subsistence farmers, and population is still increasing at 2.7% (1999-2009, Population Census).

3. Robust logistic regression

In order to identify terraced paddy expansion, we used Landsat satellite images of dates Nov. 1973 (MSS, 4 bands, 60 m resolution), Feb. 1993 and Sep. 2007 (TM, 6 bands, 30 m resolution), from the USGS Global Land Survey collection, however, we lacked ground data for the past. Thus we took the approach of using the recent land use map 2004 as training data to all dates, with the use of robust logistic regression to do supervised classification and then do *post hoc* comparison to identify expansion in paddy fields. Robust method was necessary because the training data would have more misclassifications as discrepancy in the dates between the image and the training data widens.

We used a simple large residual cut-off algorithm in applying robust logistic regression. A regression curve determining the probability of a pixel being paddy field is estimated using only the pixels that have small residuals, removing pixels with large residual, so to remove the misclassified pixels due to discrepancy in the dates taking effect on parameter estimates. Obviously, large residuals cannot be known until the true curve is estimated, so the process is an iterative one; an preliminary curve is estimated and then pixels with large residuals from that preliminary curve are removed from estimating the parameters in the next iteration; and the iteration is repeated until the curve converges.

The predicted paddy fields for 2007 was tested with ALOS pan-sharpened 2.5m resolution color image through visual interpretation and yielded overall accuracy of 87%. Classification for older years were done in the same way with the same training data, however accuracy test for two older years had not been possible (Isoda et al., 2010). The three predicted maps were overlaid to produce a map of paddy field development (Fig 3) .

4. Test data acquisition

Field work is conducted to obtain data on where and when paddy fields were developed to use that for accuracy assessment, along with the aim of exploring productivity difference depending on location and developed year. It was conducted on September 2010 and July 2011, and we visited 50 farms guided by the village heads. For the question regarding the test data, we brought pan-sharpened ALOS 2.5 m resolution color image annotated for major roads and public facilities, printed out at about 1:10,000 scale. We interviewed primarily the household head, and he was asked to indicate the locations of his paddy fields on the map, and then the year of development, previous land use and the amount of seed sown. The amount of seed is asked to estimate the size of the plot, as farmers in the region do not know the sizes. Local government uses 30 kg/ha conversion constant, and we followed that as well.

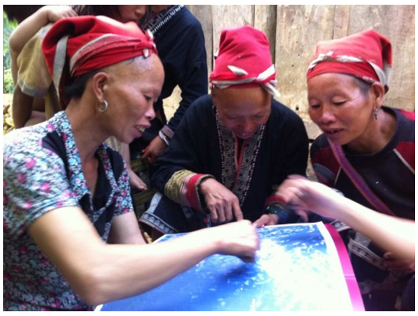
.


Fig 2. Interview using maps
Dao women in Trung Chai Commune

It may have been the first time for farmers to see the map at large scale or the satellite image, but after explaining the location of his own house and other local features, most respondent recognized the match with the image and the reality, and were able to indicate the locations. In other cases, when the respondent had difficulty understanding the map, neighbors and by-standers helped interpreting the map. Because of such situation, we should expect substantial location-wise errors. Option to actually visit the paddy to measure GPS coordinates was not practical, as typical farm has two or more plots scattered in the mountain slopes. Farmers replied the year of paddy development by either the year or years since developed, but for older paddies, many could not specify the length of time and in such cases we had to record such as ‘old’.

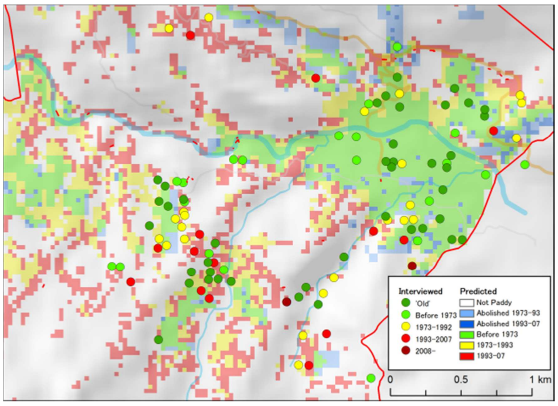


Fig 3. Sample of paddy fields collected on predicted development map

Sample paddies are plotted on the predicted map (Fig 3), and the patterns of the two seem to roughly match. Development of paddy fields in the sample is plotted in figure 4, showing rapid increase during the 1980s and stabilizing in the 2000s.

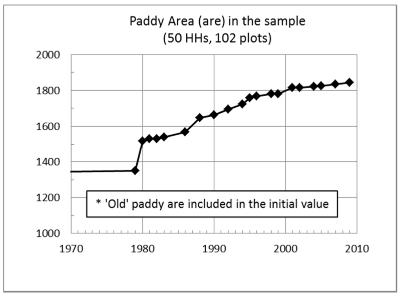


Fig 4. Paddy fields in the sample by year of development

Table 1 tabulates paddy fields in our sample by development year categorized in accordance with the satellite images we used to allow comparisons of growth rates in size. There is difficulty in how to treat ‘old’ paddies, but if they are totally neglected, an average annual increase during 1973-2007 would be 1.7%, and if they are assumed to have existed before 1973, the figure would be 0.8%. Since the two figures are an overestimate and an underestimate, respectively, remote sensing result of 1.4% seem to be a reasonable estimate, although remote sensing result of 2.3% per annum growth during 1993-2007 may be too large..

Table 1. Paddy fields by development year in the sample

			Neglecting 'Old' paddies		Assuming 'old' before 1973		cf. Remote sensing result	
Developed Year	N	Developed area (are)	Accum. Area	Annual growth	Accum. Area	Annual growth	Area (ha)	Annual growth
'Old'	42	967	–	–	–	–	–	–
Before 1973	21	373	373	–	1340	–	239	–
1973–1992	23	357	730	2.5%	1697	1.1%	312	1.2%
1993–2007	15	138	868	0.8%	1835	0.4%	453	2.3%
2008–2011	2	10	878	0.0%	1845	0.0%	–	–
Total	103	1845						
1973–2007	59	868		1.7%		0.8%		1.4%

5. Accuracy assessment

Accuracy of the paddy field development estimated by remote sensing and robust logistic regression can be assessed by measuring the match between the interview data. Confusion matrix in table 2 shows the result, examining (1) producer accuracy: how an actual class was correctly predicted (in row margins) and (2) User accuracy: how a predicted class was correct (in column margins).

Table 2. Confusion matrix

	Remote sensing result					Producer Accuracy	
Interview data	Before 1973	1973–1992	1993–2007	Not Paddy	Total	Valid Total	Correct rate
'Old'	19	10	5	8	42	–	–
Before 1973	6	4	3	7	20	20	30%
1973–1992	8	4	5	6	23	23	17%
1993–2007	3	2	4	6	15	15	27%
2008–	0	1	0	1	2	–	–
Total	36	21	17	28	102	58	24%
User Accuracy							
Valid Total	17	10	12	–	39		
Correct rate	35%	40%	33%	–	36%		
		$\kappa = 0.05$		$\chi^2 = 1.26$			

The result is disappointing, firstly of invalid data. Because of 42 'old' paddies and 2 paddies newer than our study period, there are only 58 valid samples for examining producer accuracy. Furthermore, 28 plots of paddies were predicted as 'not paddy' and while these indicate serious undercount of paddies in the prediction, these cannot be used for examining user accuracy and thus leaving mere 39 valid samples. With these small numbers of valid samples, overall producer accuracy was 24%, and overall user accuracy was 36%; only slightly more accurate than random assignment.

Since the test data was obtained by asking respondents to point on a map, there seem to be considerable locational errors. Thus we created a buffer around the sample plots and counted the pixels with the predicted class. We tried with 30m, 60m, 90m and 120m, and the 60m buffer yielded the highest overall accuracy; producer accuracy at 25% and user accuracy at 40% (table 3). Still, the prediction is only 10% better than random assignment (based on Kappa value), and we can only say that the classes in the interview data and remote sensing data is not independent at statistical significance level of 1% (based on chi-square test of independence). Another thing to note is that when 60m buffer is applied, the highest overall accuracy that can be reached is 80%, with given land use mixture around our sample points.

Table 3. Accuracy based on pixels in 60m buffer

	Remote sensing result					Producer Accuracy	
Interview data	Before 1973	1973–1992	1993–2007	Not Paddy	Total	Valid Total	Correct rate
'Old'	253	79	39	148	519	–	–
Before 1973	95	29	39	90	253	253	38%
1973–1992	88	39	52	105	284	284	14%
1993–2007	38	19	43	85	185	185	23%
2008–	1	7	1	15	24	–	–
Total	475	173	174	443	1265	722	25%
User Accuracy							
Valid Total	221	87	134	–	442		
Correct rate	43%	45%	32%	–	40%		
		$\kappa = 0.10$		$\chi^2 = 13.46^{**}$			

There is little to say about accuracies given these low accuracies, but one thing to note is that producer accuracy is not less for paddies developed 'Before 1973' compared to more recent ones, despite the fact that we used recent land use map as training data to identify paddy fields in the 1973 satellite images; supporting the idea of using robust method in analyzing land use change.

There are more to discuss about invalid data. First, majority of 'old' paddies are predicted as paddies before 1973, so it may be reasonable to assume that most 'old' paddies were developed before 1973. Second, major reason for very low producer accuracy is that robust logistic regression incorrectly assigned paddies as 'not paddy', probably because robust logistic regression misclassifies atypical paddies that are mixed with other land use within a pixel or in steep slopes. Some logical reasoning to draw a line between atypical paddies and non-paddy has to be sought. However, pixels (incorrectly) assigned as 'not paddy' is at similar ratio at all periods, thus the predicted rate of growth shown in table 1 roughly matches with the interview result.

6. Conclusion

Classifying past land use without past ground data is challenging but assessing accuracy of such classification is equally troublesome. In our case, we used robust logistic regression so to use the recent land use map as training data to identify paddy fields of the past, and then to identify paddy field expansion. For accuracy assessment, we interviewed farmers in the study area asking where and when they developed their paddy fields, to use that as test data. The accuracy assessment turned out to be a failure. The major difficulty was that we encountered so many invalid data. For old land use changes, respondents could not specify how old the change happened. About half of our sample were was removed from this reason. Another reason was that there were locational errors, as we asked to respondents to indicate the locations on the map, rather than actually going there. Applying buffers around the sample points improved the accuracy slightly, but then was difficult to interpret the accuracy values.

One good news were that even though we used recent ground data to classify land use of the past, the accuracy for the past was no less inaccurate than that for the recent one. So the use of training data of irrelevant date seem to work, and our challenge is to improve the classification accuracy for each period, possibly by using a more elaborate robust method, incorporating auxiliary data such as hill shade and slope, and finding a way to adjust for general undercount.

Acknowledgment

Our gratitude goes to the people and officers of Lao Chai Commune, and to Prof. VU Kim Chi who provided substantial help. Collaboration with the Faculty of Geography, Hanoi University of Science materialized of the Official Development Assistance Grants for UNESCO Activities granted to Doo-Chul KIM (Okayama University). This work was supported by KAKENHI Grant 23520972.

References

Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B. and Lambin, E., 2004, Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, **25**, 1565-1596.

Isoda, Y., Kobayashi, S., Sanga-Ngoie, K., Kanda, T., Nguyen, N.H., Kim, D.C., 2010. Identifying long-term land use changes using robust regression method: a study of sustainability of terraced paddy development in Northern Vietnam. *Papers and Proceedings of the Geographic Information Systems Association* 19, (CD-ROM), (In Japanese).