LANDSCAPE INDICES BASED TWO-RANK SAMPLING FOR LAND COVER VALIDATION

F. Chen [a, *], J. Chen [b], H. Wu [c]

[a] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China – cfei0618@163.com
[b] National Geomatics Center of China, Beijing 100036,China - chenjun@nsdi.gov.cn
[c] National Geomatics Center of China, Beijing 100036,China - wuhao@nsdi.gov.cn

Commission VI, WG VI/4

KEY WORDS: Validation, Globalland 30, Sampling, Landscape heterogeneity, Landscape indices, Web-based

ABSTRACT:

A web-based validation is necessary for assessing the accuracy of Globalland30. As one of the obstacles in validation of global land cover data, it is a challenge to effectively select a reasonable sample dataset. Global land cover is heterogeneous and complex. However, some sampling plans based on probability and mathematical statistics don't consider the landscape heterogeneity. To address this disadvantage of the two-rank acceptance sampling plan(TRASP) for validation of land cover data, landscape indices are used to improving this sampling plan. Landscape indicator is an typical method of quantifying landscape heterogeneity. The landscape indices based sampling combined landscape indicators and TRASP, with an innovation on the two side of computing samples' size and distributing samples. Firstly, landscape shape index(*LSI*) is introduced in the equation of TRASP for the sample size. Further, the validation area is divided into some small grids, and *LSI* is used to find more valuable hotspots from these grids to distributing samples.The theory and formulas are presented in this paper, and an example application is provided in which the sample of a land-cover map is chosen.

1. INTRODUCTION

At the beginning of 2010, China launched the project of global land cover products at 30m resolution (Globalland30) (Chen, 2011). Now, The data sets has be finished for the two base line years (i.e., 2000 and 2010). As products of higher resolution, Globalland30 would be better to characterize the interactions between human activities and land system at global scale(Giri et al. 2013), and support a number of societal benefit areas, such as sustainable development, environmental change studies, etc(Chen et al. 2013). However, the accuracy of land cover data is necessary to allow users to evaluate the utility of a map for their particular applications (Olofsson et al. 2012). So, accuracy assessment of Globalland30 is an urgent and imperative work. Now, web-based applications that use land cover information are evolving at a rapid pace. Many geo-web applications blend geographic and non-spatial information for improving global land cover products. An example is the Geo-Wiki Project, seeking crowd sourced review of areas where three global land-cover maps disagree in terms of forest and agriculture (Fritz et al. 2012). With online validation mode, there is advantage in many ways, such as sharing the product with co-operators, integrating multiple information as the reference classification, comparing accuracy among areas and so on.

A web-based validation is used to assess the accuracy for Globalland30, with the fundamental components an automatic and optimized sampling, integration of multiple reference data and achieving to documentation of accuracy. It could provide users with Globalland30 products of one class and all classes, original images, Google earth high resolution imagery, Tianditu, Corine land cover, Geo-wiki, USGS, Globcover maps and so on. Users can define interested area to validation, get optimized sample points automatically, and edit these samples. Compare the sample with other land cover products (for example, Corine land cover) or some reference information from "crowdsourcing" (for example, Geo-wiki). In addition, some experts can validate these samples from Google Earth high resolution imagery (sub-meter to 4-m resolution). However, although rigorous accuracy assessment of maps at high resolutions has been the norm for decades (see the book by Congalton and Green (1999)), issues persist related to the global land cover data. It is illustrated that validation of land cover data at global scale has always been a difficult challenge from the history of global land-cover assessment (Olofsson et al. 2012). One critical reason is design a sampling approach applicable to global land cover.

How to choose the sample is always complicated, and become more complicated especially when the goal is a sampling plan applicable to global land-cover map on the web. The first difficulty is how to compute the sample size based on probability and mathematical statistics. Typical plan is percent sampling by subjective demarcation, with dependence on prior knowledge of experts. There is a sampling plan (TRASP) for the inspection of geospatial data outputs based on the acceptance quality level(AQL) (Tong., 2011). But it don't accommodate the multiple geographical landscape arising from validation of several maps in different areas. According to a excellent sampling design, it should focuses sample sites in the areas most difficult for land-cover mapping. Because the interpretation of imagery by regional experts for mapping in the large homogeneous class cover area is easier than those mosaic landscapes with a mix of different land-cover classes landscapes.

* Corresponding author. Email: cfei0618@gmail.com
863 Program: 2013AA122802

In some areas, large homogeneous class is the major geomorphologic shape, such as desert in northern Africa, which would have an unnecessarily large sample size. Heterogeneous landscapes with a mix of different land-cover classes (such as mosaics of cropland and natural vegetation and built-up areas on the north shore of the Mediterranean) were considered complex, which have a larger sample size than the former. The typical sampling designs may not be effective when different geographical landscape must be assessed from a common sampling design. Another challenge is the automatically distribute the sample size to land cover. Some sampling designs is that researchers stratify their sampling regimes along subjectively chosen landscape features or transect, for example, an accuracy assessment of the Global Land Cover 2000 (GLC 2000, Ledwith 2000) map (Mayaux et al. 2006). For web-based validation, it is necessary for an automatic and statistical analysis based sampling.

This paper discusses the obstacles that remain in the TRASP sampling plan for global land cover data validation. It is focused on the one issue: the challenge of sampling plan to validate the area with landscapes heterogeneity. A robust and sound sample is the foundation of a rigorous accuracy assessment. This is, a automatic sampling plan in a statistically robust and sound manner is the fundamental components of a web-based validation for Globalland30. This paper is developing a sampling approach based on landscape indices for assessing the accuracy of 30m GLC products on the web.

2. LIMITATION OF TRASP

The general-purpose two-rank acceptance sampling plan (TRASP) is on the basis of the acceptable quality level (AQL), with nonconformities being modeled by a hypergeometric distribution function for lots with small sizes, and by a Poisson distribution function for lots with larger sizes, and using the interval estimation method to determine optimal sample size by controlling the probability of the relative difference between the proportion of nonconforming items in the lot and the observed sample value. It address the shortcomings of existing classical sampling plans (100% inspection or total inspection, and percent sampling inspection) for geospatial data products. The first rank sampling plan is for the inspection of the lot consisting of map sheets, and the second for inspection of the lot consisting of features in an individual map sheet. It can be expressed in Eq.(1) (the sample size is $\hat{n}$, AQL is p_0 , confidence interval is μ) (Tong ,2011)

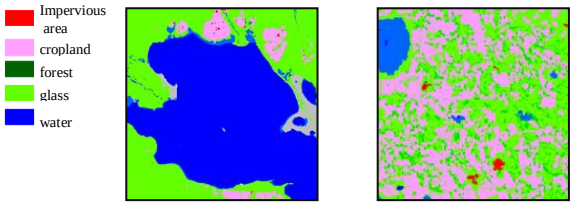
$$\hat{n} = \frac{(\mu^2_{1-(\alpha/2)}(1-p_0))/(r^2 p_0)}{1+\frac{1}{N}(\frac{\mu^2_{1-(\alpha/2)}(1-p_0)}{r^2 p_0}-1)} \quad (1)$$

Where $\hat{n}$ =the sample size
p_0=1-AQL
α=confidence interval
N=lot size
r=a limit value of the relative difference
$\mu_{1-(\alpha/2)}$ = the critical value of the standard normal distribution at the confidence level of $1-\alpha/2$.

This methodology is for the inspection of geospatial data outputs. Two-rank sampling plan is applicative for land cover validation on a global scale or a large scale. Further, this sampling plan has an obvious improvement in the obvious drawback of the traditional percent sampling plan -'strictness for larger lot size, toleration for smaller lot size' .

However, with analysis of equation (1), it is can be discovered that the sampling plan has nothing to do with the landscape pattern. When this method is used to calculate the sample size of glass land (green) in following figure1, the area of glass is the same in figure (a) and figure (b). It be judged that sample size is the same. But the sample size of figure (b) should be larger than figure (a). Because the interpretation of figure (a) is easier. Thus it can be seen equation (1) does not conform to the requirements of the sampling and so it also does not adapt with in homogeneous data.

The sampling design for land cover data should achieve several criteria: (1) it satisfies definition of a probability sampling design; (2) it provides adequate sample sizes for rare land-cover classes; (3) it allows flexibility to change sample in response to unpredictable funding or revised accuracy assessment objectives; (4) it focuses sample sites in the areas most difficult for land-cover mapping. That means that calculation of sample size and distribution of the sample are connected not only with definition of the probability, quantity of land-cover classes, but also landscape pattern(Olofsson et al. 2012). For land cover validation, TRASP has the shortcomings of un-accommodated approach to landscape diversity over extensive and heterogeneous land surface.



(a) Homogeneous landscapes (b) Mosaic landscapes

Figure 1. Example of different landscapes

3. LANDSCAPE INDICES BASED SAMPLING

Ecologists have embraced the view that the world is heterogeneous and complex (Chesson., 1986). Landscapes exhibit various degrees of spatial heterogeneity due to the interactions of natural and anthropogenic processes (De Bie C A J M., 2012.). Heterogeneity and vegetation complexity is the chief obstacles that remain in the development of sampling for land cover data.

3.1 Introduction of landscape indices

An existing method of quantifying landscape heterogeneity is through landscape indices. Landscape indices are measures of the size, shape, and spatial juxtaposition of particular land types as well as the complexity and configuration of all land types within an area(Wu., 2007). The diversity of landscape indicators has been discussed in many papers. Landscape metrics were developed to quantify spatial heterogeneity in the composition and configuration of landscape elements and to describe changes in landscape character (such as forest fragmentation) and functionality (Dale V H., 2013).

An indicator was defines in landscape ecology to describe the differences as shown in figure (a) and figure (b)-landscape

fragmentation index. Landscape fragmentation index is the ratio between average area of all pattern spots and quantity of all pattern spots. However, Globalland 30 is raster data, so if we use this index it is need to convert raster data into vector data. This transformation process is time-consuming. Therefore, other indicator applicative based pixels is better. There are another landscape indice - landscape shape index. It can be expressed in equation (2).

$$LSI = \frac{0.25E}{\sqrt{A}} \quad (2)$$

Where LSI =landscape shape index

E= the size of pixels on the boundaries and edges between one legend class and other classes

A=the size of pixels classed as one legend

3.2 Landscape indices based sampling

The purpose of sampling is in order to solve two problems: the sample size and the distribution of samples. Both are relational to landscape indices. Essentially, landscape indices serve to organize environmental and biotic heterogeneity in a logical way. Such sampling approaches could be further improved if they were guided by an initial idea of landscape indices.

The first, it is proved that statistic analysis based sampling cannot adapt with heterogeneity areas. If a landscape index is introduced in the equation of TRASP, the sampling plan could be regarded as a robust methodology. We try to use landscape shape index to modify the equation (1). According to the graphical analysis, if the value of *LSI* is higher, the complexity of land cover is greater and it is need more samples to validate the heterogeneity region. Therefore, based on equation (1) and equation (2), the sample size n can be derived using the number of outline pixels w by

$$n = \frac{(\mu^2_{1-(\frac{\alpha}{2})}(1-p_0))/(r^2 p_0)}{1+0.25*w*\sqrt{N}(\frac{\mu^2_{1-(\frac{\alpha}{2})}(1-p_0)}{r^2 p_0}-1)} \quad (3)$$

Where n =the sample size

p_0=1-AQL

α=confidence interval

N=lot size

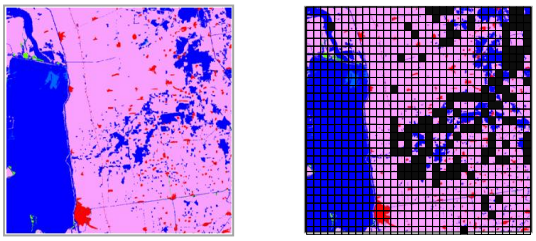
r=a limit value of the relative difference

$\mu_{1-(\alpha/2)}$ = the critical value of the standard normal distribution at the confidence level of $1-\alpha/2$.

E= the size of pixels on the boundaries and edges between one legend class and other classes

A=the size of pixels classed as one legend

On the other hand, landscape indices can be used to optimize the distribution of samples. In the validation area, there are different landscapes due to the different classes and different spatial location. It is illustrated by figure 2(a). The water (blue) can be divided into the left and the right in the area (figure 2). On the left side, the water body is large homogeneous area, while it is a composite of broken figure spots on the right. When the water land cover is validated, the sample should be a hotspot on the right side. In order to achieve the concentration of samples on the right, smaller grids are used to split this area. The grid is the same size and small enough that there is only one landscapes in them. Prioritizing every grid according to *LSI*, these grids with higher *LSI* has more chances to be selected.



(a) An example (b) The most valuable grids (black) based on *LSI*

Figure 2. The overall design of landscape indices based sampling

3.3 Methods

We can use two rank sampling plan to validate Globalland30. Sampling for the first time could be the equal probability sampling. The map would be divided into a series of sheet according to 1°*1°. The sample sheet can be chosen from those sheets in the way of simple random sampling. The amount can be achieved in equation (1).

Within each virtual chosen map, processing steps are:

（1） We can get the sample size combined landscape shape index and the area of one legend class(equation (2)). The quantity would increases if the proportion of one class increases, and the quantity would also increases as well if the landscape fragmentation degree increases.

（2） The map would be divided with 10 km * 10 km grid. The grid would be sorted from large to small based on the landscape shape index.

（3） Distributing sample points should prioritize the mosaic landscapes. This is, the sampling plan is sampling with unequal probabilities. Some grids would be filtered out that the value of *LSI* is under threshold ψ. Sample points could be chosen from other grids.

The sampling design proposed in this paper is shown schematically in figure 3.

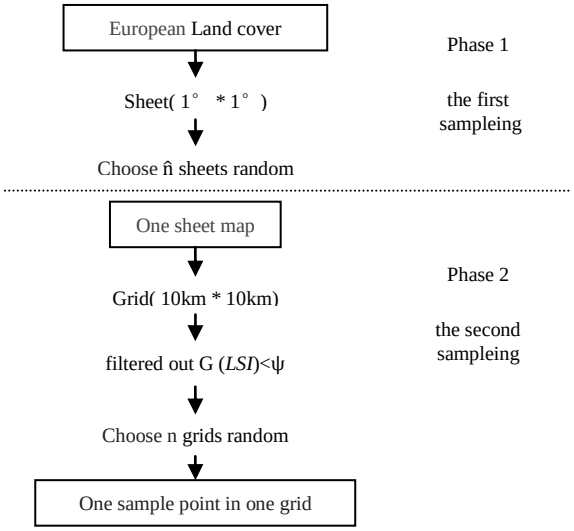
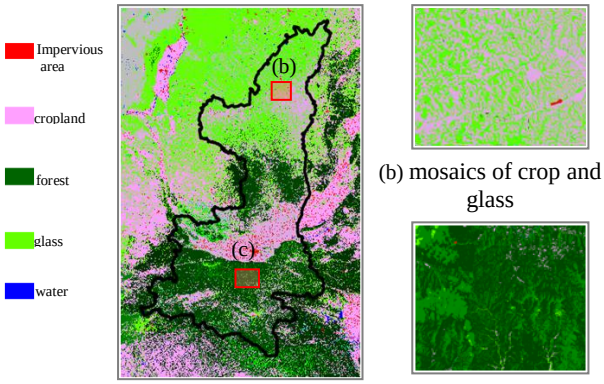


Figure 3.The method of landscape Indices based sampling

4. EXPERIMENT AND DISCUSSION

Results presented in this study are based on land cover products at 30m resolution of Shanxi province. The predominant land-cover type in the region is crop land in the central of Shanxi, forest in southern and central area, grass land of northern Shanxi province, and impervious areas throughout this province. Through statistics analysis of this data, some rare classes is found, such as water bodies, in minority areas of Shanxi.

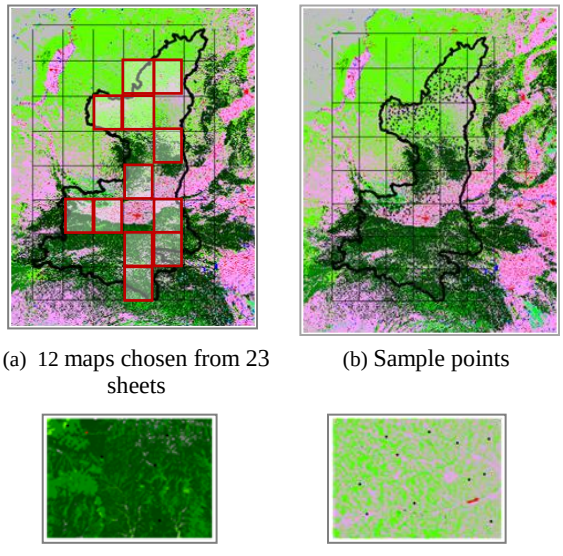
In figure 4, forest is divided into five subclasses. However, they are seen as one class in this experiment. It is obvious that forest and cropland of southern Shanxi province are major geomorphologic shape(figure 4(a)). Conversely, the area in the north and midland is mosaics of cropland and natural vegetation(glass and forest) and built-up areas(figure 4(b)).



(a) Shanxi province (b) mosaics of crop and glass (c) homogeneous forest

Figure 4. The land cover data in Shanxi province

The First Step, the map is be divided into 23 sheets (figure 5(a)). The amount of selected sheets is 12 according to equation (1) (the limit value of the relative difference is 0.2; confidence interval is 0.95; AQL is 0.2;). Theses 12 maps in red are chosen with simple random sampling. Within each sheet, sample points can be selected with the proposed design method in 3.3 (figure 5(b)).



(a) 12 maps chosen from 23 sheets (b) Sample points

(c) homogeneous forest (d) mosaics of crop and glass

Figure 5. The results of sampling

Landscape indices based sampling could be adaptive to landscape diversity. In figure 5(c) and (d), the distribution of samples have significant difference. We can use this approach as the fundament of web-based validation. These sample points must be determined the reference classification. It is high cost to gather reference classification from conducting interviews and fieldwork. Now Google Earth can provide high resolution imagery (sub-meter to 4-m resolution), which can be used for low-cost and reasonably accurate reference data for both producing land-cover maps and testing their accuracy. And some web-based land cover information is another way to get reference classification.

5. CONCLUSIONS AND FUTURE WORK

There are some barriers to realize a web-based validation for Globalland 30 . One of major challenges of online validation is an automatic sampling with consideration of landscape diversity over extensive and heterogeneous land surface. Landscape indices based sampling combines the probability and mathematical statistics and landscape indices, improving the general-purpose two-rank acceptance sampling plan (TRASP) through computing the sample size using landscape shape index. Further, *LSI* is used to prioritize small area and discover hotspots to distributing samples at those sites with higher probability. The result of a experiment in Shanxi province proved the applicability of landscapes diversity.

The further work is to solve another major challenges of online validation. That is how to collect the reference classification. One way is through comparative analysis with other land cover maps. The comparison of land cover data sets provides insight for both data producers and users (Giri., 2005). The multiplicity of existing products differ in scale, nomenclature, minimum mapping unit (MMU) and data format, among other factors. Efficient integration of these multisource data on the web would facilitate data access management Globalland 30 maps.

REFERENCE

Chen J., Chen, J., Gong, P., et al. 2011. Higher Resolution GLC Mapping. *Geomatics World*, 4 (2), pp. 12-14.

Chen J., Chen, J., Cao X., et al. 2013. A Precise Classification Methodology for China's Global Land Cover Mapping at 30 Meters Resolution. *ISPRS Journal of Photogrammetry and Remote Sensing*, in review.

Chesson P L., Case T J. 1986. Overview: nonequilibrium community theories: chance, variability, history and coexistence. In: J. Diamond and T. Case, eds. Community ecology, Harper and Row, New York, pp. 229–239.

Clark M L., Aide T M. 2011. Virtual interpretation of Earth Web-Interface Tool (VIEW-IT) for collecting land-use/land-cover reference data. *Remote Sensing*, 3(3), pp. 601-620.

Congalton R G, Green K. 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices . CRC Press, Boca Raton, FL.

Dale V H, Kline K L. 2013. Issues in using landscape indicators to assess land changes. *Ecological Indicators*, 28, pp. 91-99.

de Bie C, Nguyen T T H, Ali A, et al. 2012. LaHMa: a landscape heterogeneity mapping method using hyper-temporal datasets. *International Journal of Geographical Information Science*, 26(11), pp. 2177-2192.

Fritz S, McCallum I, Schill C, et al. 2012. Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, 31, pp. 110-123.

Giri C, Zhu Z, Reed B. 2005. A comparative analysis of the Global Land Cover 2000 and MODIS. *Remote Sensing of Environment*, 94, pp. 123–132.

Giri C, Pengra B, Long J, et al. 2013. Next generation of global land cover characterization, mapping, and monitoring. *International Journal of Applied Earth Observation and Geoinformation*, 25, pp. 30-37.

Jiang C S, Wang J F, Cao Z D. 2009. A Review of Geo-Spatial Sampling Theory. *Acta Geographica Sinica*, 64, pp. 368-380

Mayaux P, Eva H, Gallego J, et al. 2006. Validation of the global land cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, 44, pp. 1728–1739.

Olofsson P, Stehman S V, Woodcock C E, et al. 2012. A global land-cover validation data set, part I: fundamental design principles. *International Journal of Remote Sensing*, 33(18), pp. 5768-5788.

Tong X, Wang Z, Xie H, et al. 2011. Designing a two-rank acceptance sampling plan for quality inspection of geospatial data products. *Computers & Geosciences*, 37(10), pp. 1570-1583.

Wu J G., 2007. Landscape Ecology – Pattern, Process, Scale and Hierarchy. Heiher Education Press, Beijing.