African Journal of Agricultural Research Vol. 5(13), pp. 1647-1654, 4 July, 2010
Available online at http://www.academicjournals.org/AJAR
DOI: 10.5897/AJAR09.011
ISSN 1991-637X ©2010 Academic Journals

*Full Length Research Paper*

# Monte Carlo simulation and remote sensing applied to agricultural survey sampling strategy in Taita Hills, Kenya

**Eduardo Eiji Maeda*, Petri Pellikka and Barnaby J. F. Clark**

Department of Geosciences and Geograph, University of Helsinki, Gustaf Hällströmin katu 2, 00014, Helsinki, Finland.

Accepted 22 June, 2010

**Remote sensing and Geographical Information Systems (GIS) are important tools used for assisting agricultural surveys. Such tools can be used to stratify the population samples in a study area, optimizing and reducing the costs of field work. Nevertheless, defining the number of samples to be visited in the field is a challenging task. In the presented research, the sampling strategy for agricultural survey was addressed by integrating GIS, remote sensing techniques and a Monte Carlo simulation. A study case was carried out in the Taita Hills, Kenya to test the operational viability of the method. The applied approach allowed the estimation of crop areas with reduced uncertainties and management of the errors.**

**Key words:** Agricultural survey, Taita Hills, Monte Carlo simulation, remote sensing.

## INTRODUCTION

Crop area estimation is an essential procedure in supporting policy decisions on land use allocation, food security and environmental issues. Nevertheless, producing agricultural statistics in regions with limited financial resources and restricted access is a challenging task.

Agriculture is currently the main economic activity in most Sub-Saharan African countries, representing around 40% of their gross domestic product (Barrios et al., 2008). Despite the clear importance of this sector for the economy, a great part of the agricultural activities in this region are characterized by low yields, subsistence practices and low technological development. Due to these facts, together with concomitant problems such as weeds, pests, and diseases, this is the only region of the world where per capita food production has remained stagnant over the past 40 years (Sanchez, 2002).

In such areas, a detailed knowledge of the ongoing agricultural activities is the first step in the development of strategies towards adequate and sustainable food production. In Kenya, a periodic agricultural survey is carried out by the Ministry of Agriculture. However, crop areas are currently estimated using a subjective approach, which is mostly based on interviews carried out with local producers. Although such an approach can sometimes retrieve relatively accurate figures, it is highly subject to biases and uncertainties. Moreover, it is costly and slow, given that it requires a large number of agents and vehicles to carry out the interviews.

During the last years, remote sensing techniques and Geographical Information Systems (GIS) have proven to be efficient tools to monitor agricultural activities. Among the remote sensing and GIS applications for agriculture, it is worth mentioning precision agriculture (Seelan et al., 2003; Beeria and Peled, 2009), water resources management (Folhes et al., 2009; Karatas et al., 2009) and yield forecast (Ferencz et al., 2004; Pan et al., 2009).

However, although remote sensing has proven to be useful in different agricultural applications, many limitations are still faced in the operational usage of this tool to estimate crop areas. For instance, the spatial resolution of many operational satellite sensors is not adequate to discriminate crop types in areas with intensive agriculture and small properties. On the other hand, orbital sensors with very high spatial resolution are costly and typically have low temporal resolution, which is inadequate to monitor dynamic activities such as agriculture. Moreover,

*Corresponding author. E-mail: eduardo.maeda@helsinki.fi. Tel: +358-44-2082876.

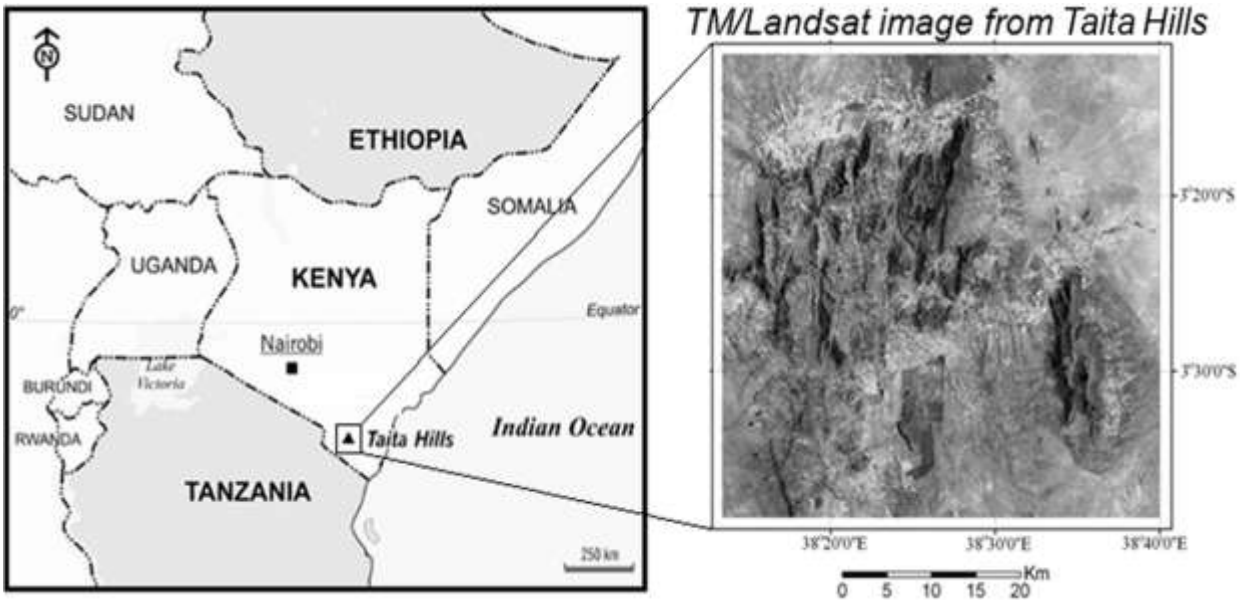


**Figure 1.** Geographic location of the study area.

in tropical areas cloud cover may preclude the use of satellite images in most periods of the year (Asner, 2001). As an alternative, remote sensing can be used to assist agricultural surveys by defining sampling units, optimizing the sample allocation and size of sampling units (Carfagna and Gallego, 2005). Consequently, the integration of remote sensing techniques with ground surveys has been the focus of much research in past years (Pradhan, 2001; Epiphanio et al., 2002; Gallego, 2004).

One of the main issues that remains unsolved is regarding the sampling strategy to be carried out during the ground surveys. Namely, an oversized number of samples will result in unnecessary costs with regards to field work and increase the time necessary to perform the survey. On the other hand, if too few samples are collected, the accuracy and reliability of the survey may be compromised.

The objective of this research was to develop a sampling scheme methodology to estimate crop areas in the Taita Hills, Kenya, by integrating a Monte Carlo simulation, GIS and remote sensing techniques.

## Study area

The Taita Hills (Figure 1) are the northernmost part of the Eastern Arc Mountains, which is a biodiversity hotspot containing many endemic animals and plant species. The indigenous cloud forests have suffered substantial loss and degradation for several centuries as abundant rainfall (annual 1100 mm) and rich soils (cambisols and humic nitosols) have created good conditions for agriculture. The agriculture in the hills is intensive small-scale subsistence farming. In the lower highland zone and in the upper midland zone, the typical crops are maize, beans, peas, potatoes, cabbages, tomatoes, cassava and banana (Soini, 2005). In the slopes and lower parts of the hills with average annual rainfall between 600 and 900 mm, early maturing maize, sorghum and millet species are cultivated. In the lower midland zones with average rainfall between 500 and 700 mm, dryland maize types and onions are cultivated, among others. With a high population growth of 2%, the decreasing available lands in the hills have led to clearance of new fields in the foothills and lowlands.

## MATERIALS AND METHODS

The method applied in this study uses an adaptation of the approach proposed by Epiphanio et al. (2002), which associates statistical methods with GIS and remote sensing techniques to assist surveys aiming at crop areas estimation. The first step of the applied approach consisted of using remote sensing techniques to identify the areas where agricultural activities are taking place within the study area. Next, a stratified random sampling scheme was applied using GIS. In order to define the number of samples and to estimate the errors involved in the analysis, a Monte Carlo simulation was performed. After defining the sample strategy, field work was carried out to visit the areas indicated by the point samples. The crop types observed in each sample were used in a statistical analysis to estimate the area planted with each crop type in the study area.

A flowchart illustrating the steps taken in this research is presented in Figure 2. Some of the steps are:

### Stratification

In order to generate the stratification for the statistical analysis, a land cover map of Taita Hills was created from a 15th October 2003 dated SPOT 4 HRVIR 1 satellite image (path and row, 143 - 357),

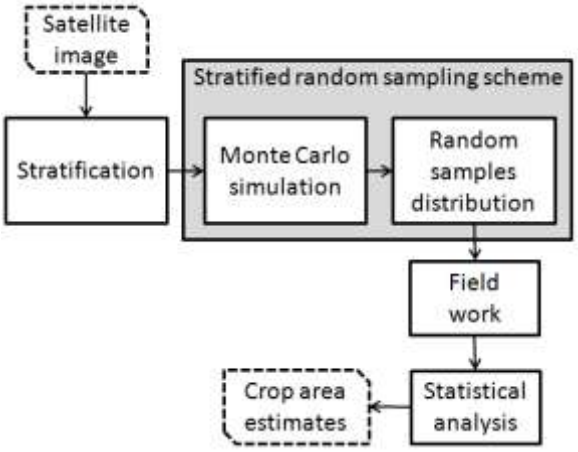


**Figure 2.** Flowchart illustrating the steps carried out during this research.

with a 20 m spatial resolution and green, red and NIR spectral bands. Because of the rugged terrain in the area, the image was orthorectified using a 20 m planimetric resolution digital elevation model (DEM), interpolated from 50-feet interval contours captured from 1:50 000 scale topographic maps. Atmospheric correction was implemented utilizing the historical empirical line method (HELM) (Clark and Pellikka 2005, 2009). Removal of slope-aspect effects from the atmospherically corrected imagery utilized a topographic correction method based on the cosine function described by Teillet et al. (1982), with band specific 'c' correction factors calculated for general vegetation classes identified by clustering Normalized Difference Vegetation Indexes (NDVI).

The SPOT image was classified into land use and land cover types according to a nomenclature which was derived using the Land Cover Classification System (LCCS) of the Food and Agriculture Organization of the United Nations (FAO) and the United Nations Environment Programme (UNEP) (di Gregorio, 2005). The classification methodology utilized a multi-scale segmentation/object relationship modeling (MSS/ORM) approach implemented with the Definiens software tool (Clark and Pellikka, 2009; Burnett and Blaschke, 2003; Baatz et al., 2004). Central to this methodology is the generation of meaningful image objects, relating to homogeneous land cover patches, by multi-scale segmentation based on both spectral and textural characteristics of the imagery. Such discrete regions of a remotely sensed image are made up of aggregations of pixels that are homogenous with regard to spectral and spatial characteristics. Homogenous in this instance refers to the fact that the within-object variance is less than the between-object variance. In a multi-scale segmentation, a so called "scale" parameter is used to determine the average size of the image segments at each level in the hierarchy.

The first segmentation level is critical because the borders defined at this stage will be adhered to by any subsequent segmentation, either subdividing the image object primitives or combining them into larger objects. In the small-scale cultivation areas in the Taita Hills a very detailed initial segmentation with a scale parameter of 2 was used, while in the shrublands in the foothills and lowlands an aggregation with a scale parameter of 4 was used. These segmentation levels were merged to the final mapping level 2.

Finally, the classification map was subjected to final visual inspection and manual editing of any noted errors. An accuracy assessment was undertaken based on ground reference test data, independent of the training data, collected during field visits to the Taita Hills in January 2005 and 2006.

**Stratified random sampling and statistical analysis**

In a stratified random sampling, the population is first divided into a number of parts or 'strata' according to characteristics that are considered to be associated with the main variables being studied. In the particular case of this study, the stratification was performed by separating the agricultural areas from the remaining land use classes, as described in the previous section. Hence, the pixels of the image classified as agricultural area were inserted in a subpopulation with N members. After the stratification was defined, a random sampling algorithm was applied to collect n samples inside the subpopulation. The stratified random sampling was likely to achieve better results than the simple random sampling, provided that the strata was chosen in such a way that members of the same stratum are as similar as possible in respect to the characteristic of interest.

In order to identify the type of crop cultivated in the area represented by each sample, field work was carried out assisted by GIS and GPS. After the crop type of each sample was defined, it follows that the proportion in which a determined crop type occurs in n is equivalent to the proportion of this respective crop type in N (Cochran, 1977). The method can be elucidated using the following notation (Epiphanio et al., 2002):

$x$ = crop type being considered;
$A$ = number of elements in N classified as $x$;
$a$ = number of elements in n classified as $x$;
$P = A/N$ = proportion of $x$ in N;
$p = a/n$ = proportion of $x$ in n;
$f = n/N$ = sampling fraction ($1/f = N/n$ = expansion factor);

Hence, the estimation of P is directly given by p, and the estimation of A is calculated using the expansion factor, that is:

$$A = a \times \left( \frac{1}{f} \right) \quad (1)$$

Attributing for each member of n a variable y, which assumes the value 1 if the crop type belong to x, or 0 if the crop type is other than x, the average proportion of the crop type x in n and N are given by the following equations, respectively:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{a}{n} = p \quad (2)$$

$$\bar{Y} = \frac{\sum_{i=1}^{N} Yi}{N} = \frac{A}{N} = P \quad (3)$$

**Monte Carlo simulation**

Defining the number of samples for statistical analysis in natural resources surveys has always been an important issue. To overcome this problem, the presented research carried out a Monte Carlo simulation (Metropolis and Ulam, 1949) prior to the field work. The results of the simulation were used to define the most suitable sampling strategy taking in account the errors inherent in the analysis and the time and resources available for the field work. The Monte Carlo method uses random numbers and probability to solve problems by directly simulating the process. It may be used to iteratively evaluate a deterministic model using sets of random numbers as inputs.

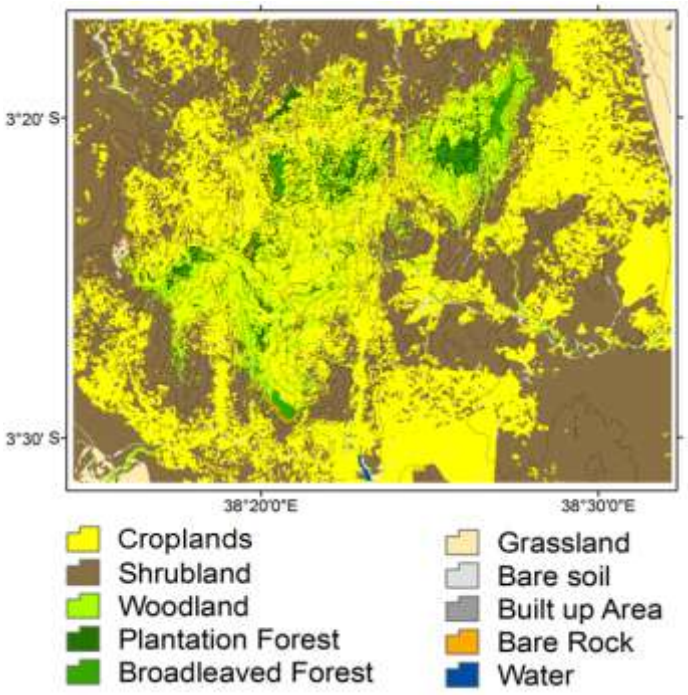


**Figure 3.** Land use and land cover classification used to perform the stratification of the study.

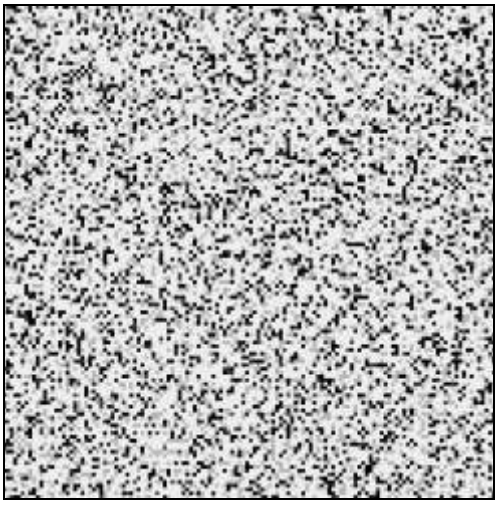

**Figure 4.** Synthetic crop field

The first step in performing the analysis was to assemble a simple model to simulate the field work activities. An image representing a Synthetic Crop Field (SCF) was generated by creating a matrix with the same number of pixels (N) as observed in the stratum classified as agricultural areas in Taita Hills, resulting in an image with dimensions $N^{1/2}$ x $N^{1/2}$ pixels. A crop type class was randomly attributed to each element of the SCF. In order to create a consistent proportion, the number of classes and the percentage of each class in the SCF were based on numbers acquired from previous surveys carried out by the Kenya's Ministry of Agriculture.

After setting the SCF, random samples were collected within the matrix, and used to estimate the proportion of crop types using equations 2 and 3. The number of samples (n) used to estimate the crop type proportion in the SCF ranged from 10 to 1000.The simulation was repeated 100 times for each n value. The error in estimating the crop type proportion was monitored in every iteration, and the average Root Mean Squared Error (RMSE) for a determined n was calculated using the following equation:

$$RMSE_n = \left( \frac{1}{R} \sum_{1}^{R} \left( p_n - P \right)^2 \right)^{0.5} \quad (4)$$

Where; R represents the number of iterations using a determined number of samples (n). In this study, R is equal to 100. And $p_n$ is the proportion of a determined crop type estimated using n samples. The calculations and Monte Carlo simulation described in this study were carried out using the software MATLAB™ R2008a. Given the small size of the study area, the simulations were not computational intensive, and could be performed quickly and efficiently using a single general-purpose computer with a processor of 2.53 GHz and 3 GB of RAM memory.

## RESULTS AND DISCUSSION

The result of the land use and land cover classification used to perform the stratification of the study area is presented in Figure 3. The overall accuracy of the classification was 89%, with a Kappa index for agreement of 0.87. The croplands class, which is the main target in the present research, achieved a producer's accuracy of 96% and a user's accuracy of 82%. The producer's accuracy indicates the probability of a reference pixel for a particular category on the map being correctly classified, and is a measure of the omission error. Alternatively, the user's accuracy indicates the probability that a pixel classified on the map actually represents that category on the ground, and is a measure of the commission error. Provided the fact that the rate of land cover change was not high between 2003 and the time when the study was being carried out, the use of an image prior to the ground survey can be used without great loss in the accuracy (Gonzáles-Alonso et al., 1997).

In total, 902500 pixels in the SPOT image were classified as croplands, representing an area of approximately 360 $km^2$. Once the total area occupied by agricultural activities was known, the SCF was assembled by creating a matrix with the same number of pixels (Figure 4). Five different types of crops were randomly assigned for each element of the matrix, assuming that each 20 m x 20 m pixel was used to cultivate just one type of crop. A hypothetic proportion for each crop type was defined based on figures provided by previous agricultural surveys carried out by Kenya's Ministry of Agriculture, as showed in Table 1.

The proportions defined in the simulation suggest the existence of a predominant crop, represented by Crop 1, which is cultivated in 60% of the agricultural areas. Another important crop type, that is cultivated in 20% of the SCF, is represented by Crop 2. The remaining crops

**Table 1.** Crop type distribution in the synthetic crop field (SCF).

| Crop type | Percentage (%) | Number of pixels | Total area (km$^2$) |
|---|---|---|---|
| Crop 1 | 60 | 541500 | 216.6 |
| Crop 2 | 20 | 180500 | 72.2 |
| Crop 3 | 10 | 90250 | 18.05 |
| Crop 4 | 5 | 45125 | 18.05 |
| Crop 5 | 5 | 45125 | 36.1 |

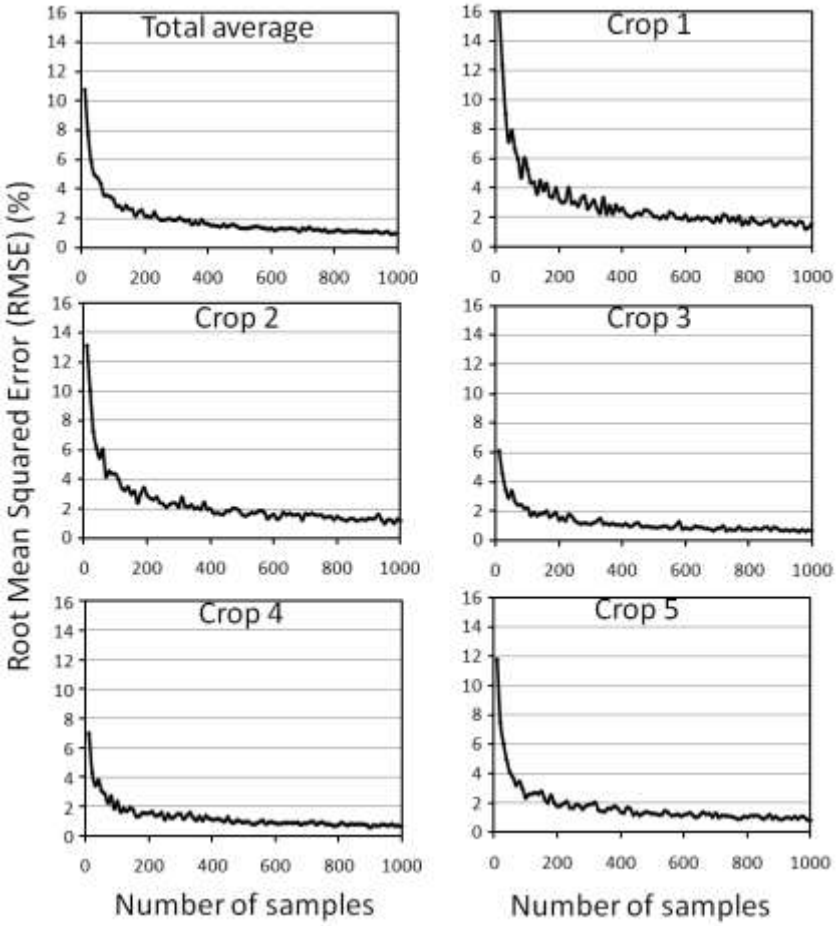


**Figure 5.** RMSE retrieved from the Monte Carlo simulation.

(3, 4 and 5) stand for secondary agricultural activities that occupy less than 10% of the croplands. In comparison with the real figures, Crops 1 and 2 represent maize and beans, which are, according to Kenya's Ministry of Agriculture, the main crops planted in the Taita Hills. Crops 3, 4 and 5 represent crops such as cassava, edible seeds (e.g. pigeon peas, cowpeas) and vegetables (e.g. kales, cabbage, tomato, onion).

The results of the Monte Carlo simulation performed to estimate the proportion of each crop type in the SCF are presented in Figure 5.

The total average RMSE ranged from around 10%, using 10 samples, to less than 1%, when 1000 samples were used to estimate the crop type's proportion in the SCF. In general, it is noted that the predominant Crops (1 and 2) retrieved higher RMSE, although the error curve in both cases followed the same trend observed in the other crops. The trend line that best fits all regressions can be described by an exponential function. The exponential regression between the total average RMSE and the number of samples achieved a coefficient of determination ($R^2$) of 0,985, and is described by the following

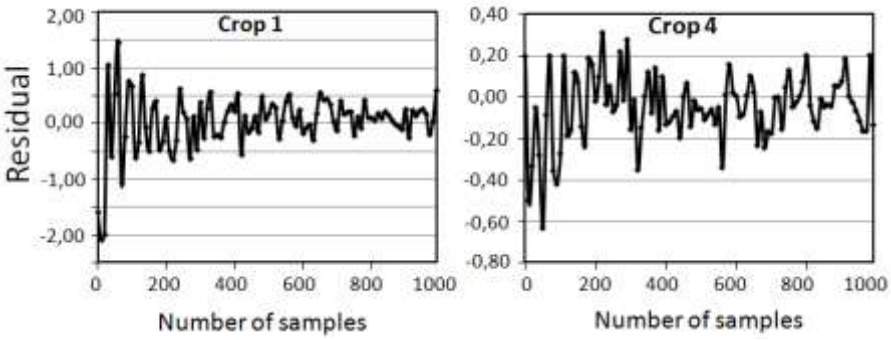


**Figure 6.** Residuals obtained during the simulation using different number of samples.

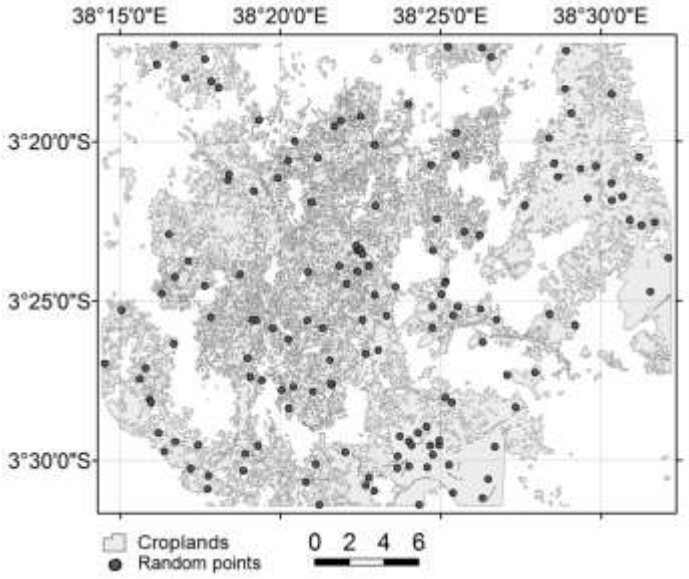


**Figure 7.** Random sample points distributed in the study areas.

equation:

$$RMSE_n = 33{,}17\, n^{-0{,}506} \quad (5)$$

In general, no clear tendencies could be observed during the simulations in overestimating or underestimating the crop areas proportion. That is to say, the residuals fluctuated between positive and negative throughout the simulation (Figure 6). It was observed that when using a small number of samples (less than 100), the proportion of crops less predominant, such as crop 4, were likely to be slightly underestimated. However, there are no reasons to believe that these results were caused by negative biases.

After the simulations, a field work campaign was organized in order to test the operational viability of the method. Using GIS, 300 random points were distributed in the areas classified as croplands (Figure 7). This number was chosen taking into account the error curves showed in Figure 5 and the time and resources available for the field work. In total, six days were spent for the field work. A GPS was used to locate the points in the field and the crop types being cultivated in the exact location of each point were annotated. By the end of the field work, the gathered information was processed in a spreadsheet, and the proportion of each crop type was calculated using equations 1, 2 and 3.

The field work was successful in visiting 225 point out of the 300 initially set. The remaining points were not reached due to a variety of factors, such as closed roads or time constraints. As a result, using equation 5, it was possible to estimate that the calculation of the crop type's proportion in the study area using the 225 samples involved the possibility of a RMSE of approximately 2.14%. The results of the crops proportion and cultivated area estimations are presented in Table 2.

The proportion of the crop types estimated by the

**Table 2.** Crops proportion and cultivated area estimated using the stratified random sampling scheme.

| Crop | Percentage | Cultivated area (km$^2$) |
|---|---|---|
| Maize | 38.02 | 137.27 |
| Beans | 16.79 | 60.61 |
| Peas | 9.88 | 35.65 |
| Cassava | 8.15 | 29.41 |
| Sisal | 5.93 | 21.39 |
| Vegetables | 5.43 | 19.61 |
| Pasture | 3.70 | 13.37 |
| Sugar cane | 3.46 | 12.48 |
| Other | 8.64 | 31.20 |

presented methodology was consistent with previous agricultural surveys carried out by the Kenya's Ministry of Agriculture. Moreover, the results properly reflect the current trends observed in the field and described by officers from the Ministry of Agriculture. Namely, there is an increasing interest in diversifying the crop types cultivated in the region. Due to this fact, crops such as cassava, cowpeas and pigeon peas have started to be grown in locations previously dominated by maze.

Nevertheless, despite the clear indications that the method retrieved appropriated results, further investigations are necessary to perform a detailed validation. For this, the results obtained in the presented research must be compared with upcoming reports from Kenya's Ministry of Agriculture and the possible sources of errors and uncertainties investigated.

## Conclusions

This paper presented the application of a simple and effective approach to improve the sampling strategy for an agricultural survey in the Taita Hills, Kenya. By integrating GIS, remote sensing techniques and a Monte Carlo simulation, the proposed method decreases the uncertainties and costs involved in the agricultural survey. It was shown that the average RMSE used in estimating the crop types proportion can vary from around 10%, using 10 samples, to less than 1%, when 1000 samples are used.

The methodology was tested in field work, where 225 points were visited and used to estimate the crop areas in the Taita Hills. The results obtained agree with figures provided by Kenya's Ministry of Agriculture. Namely, maize and beans continue to be the predominant crops in the region, although current trends show that farmers are heading for a more diversified agriculture. Crops more resistant to drought, such as cassava, cow peas and pigeon peas, are starting to be cultivated as an alternative to maize.

Moreover, the present research illustrates a successful connection between science and practice, where theoretical knowledge is used to reduce costs and obtain reliable information necessary for the society.

## ACKNOWLEDGEMENTS

The authors kindly thank Mr. Henry Mazera from Kenya's Ministry of Agriculture for the support provided during the field work campaign in September 2009. The authors also wish to thank Mika Siljander for providing part of the geospatial dataset used in this research, and Tapio Kallio for his assistance during the field work activities. The research was funded by the Centre of International Mobility (CIMO), University of Helsinki and Academy of Finland for TAITATOO project.

### REFERENCES

Asner GP (2001). Cloud cover in Landsat observations of the Brazilian Amazon. Int. J. Remote Sensing, 22(18): 3855-3862.

Baatz M, Benz U, Dehghani S, Heynen M, Höltje A, Hofmann P, Lingenfelder I, Mimler M, Sohlbach M, Weber M, Willhauck G (2004). eCognition Professional: User Guide 4 (Munich: Definiens-Imaging).

Barrios S, Salvador, Ouattara B, Strobl E (2008). The impact of climatic change on agricultural production: Is it different for Africa? Food Policy, 33: 287 - 298.

Beeria O, Peled A (2009). Geographical model for precise agriculture monitoring with real-time remote sensing. ISPRS J. Photogr. Remote Sensing, 64(1): 47-54.

Burnett C, Blaschke T (2003). A multi-scale segmentation/object relationship modelling methodology for landscape analysis. Ecol. Modelling, 168: 233-249.

Carfagna E, Gallego FJ (2005). Using remote sensing for agricultural statistics. Int. Stat. Rev., 73 (3): 389-404.

Cochran WG (1977). Sampling Techniques. 3rd ed. New York: John Wiley & Sons, p. 428.

Clark BJF, Pellikka PKE (2005) The development of a land use change detection methodology for mapping the Taita Hills, South-East Kenya. Proceedings of the 31st International Symposium of Remote Sensing of the Environment, 20-24 June 2005, St. Petersburg, Russia. CD-Rom publication.

Clark BJF, Pellikka PKE (2009). Landscape analysis using multiscale segmentation and object orientated classification. In: A. Roeder & J. Hill (Eds.), Recent Advances in Remote Sensing and Geoinformation Processing for Land Degradation Assessment. Taylor and Francis.

Di Gregorio A (2005). Land Cover Classification System (LCCS), version 2: Classification Concepts and User Manual. FAO Environmental and Natural Resources Series 8 FAO Rome, p. 208.

Epiphanio JCN, Luiz AJB, Formaggio AR (2002). Crop area estimates using simple sampling scheme on satellite images. Bragantia, 61: 187-197.

Ferencz C, Bognár P, Lichtenberger J (2004). Crop yield estimation by satellite remote sensing. Int. J. Remote Sensing, 25(20): 4113-4149.

Folhes MT, Rennó CD, Soares JV (2009). Remote sensing for irrigation water management in the semi-arid Northeast of Brazil. Agric. Water Manage., 96 (10): 1398-1408.

Gallego FJ (2004). Remote sensing and land cover area estimation. Int. J. Remote Sensing, 25: 3019-3047.

Gonzáles-Alonso F, Cuevas JM, Arbiol R, Baulies X (1997). Remote sensing and agricultural statistics: crop area estimation in north-eastern Spain through diachronic Landsat TM and ground sample data. Int. J. Remote Sensing, 18(2): 467-470.

Karatas BS, Akkuzu E, Unal HB, Asik S, Avci M (2009). Using satellite remote sensing to assess irrigation performance in Water User Associations in the Lower Gediz Basin, Turkey. Agric. Water Manage., 96(6): 982-990.

Metropolis N, Ulam S (1949). The Monte Carlo Method. J. Am. Stat.

Assoc., 44 (247): 335-341.
Pan G, Sun GJ, Li FM (2009). Using QuickBird imagery and a production efficiency model to improve crop yield estimation in the semi-arid hilly Loess Plateau, China. Environmental Modelling and Software 24 (4): 510-516.
Pradhan S (2001). Crop area estimation using GIS, remote sensing and area frame sampling. Int. J. Appl. Earth Observation and Geoinformation 3 (1): 86-92.
Sanchez PA (2002). Soil Fertility and Hunger in Africa. Sci., 295: 2019-2020.
Seelan SK, Laguette S, Casady GM, Seielstad GA (2003). Remote sensing applications for precision agriculture: A learning community approach. Remote Sensing Environ., 88(1-2): 157-169.
Soini E (2005). Livelihood capital, strategies and outcomes in the Taita hills of Kenya. ICRAF Working Paper no. 8. Nairobi, Kenya: World Agrofor. Centre, p. 48.
Teillet PM, Guindon B, Goodenough DG (1982). On the slope-aspect correction of multispectral scanner data. Canadian J. Remote Sensing, 8: 84-106.