Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Florianopolis-SC, Brazil, July 10-13, 2012.

Canasat Project accuracy assessment of sugarcane thematic maps

Marcio Pupin Mello, Marcos Adami, Bernardo F. T. Rudorff and Daniel Alves Aguiar

National Institute for Space Research (INPE), Remote Sensing Division (DSR)
Av. dos Astronautas, 1758 – Jd. Granja – São José dos Campos – SP, 12227-010, Brazil
mello@ieee.org, adami@dsr.inpe.br, bernardo@dsr.inpe.br, daniel@dsr.inpe.br

Abstract

In order to perform the thematic accuracy assessment of Canasat Project mapping, we used a novel approach to construct the reference dataset which consists of a web platform used to integrate remote sensing images, time series, and ancillary data. Results showed that, although Canasat Project mapping of sugarcane areas has a mean overall accuracy of 98%, the mean thematic error associated with sugarcane area estimates was of –0.7%, since part of the omission errors was compensated by the inclusion errors. Despite assessing only one crop year (2010/2011) we would also expect to obtain very similar results for other crop years due to the consistent method and the careful visual interpretation carried out by the Canasat Project experienced team.

Keywords: Remote sensing, web platform, sampling, Monte Carlo method.

1. Introduction

Brazil is the world's largest sugarcane producer and, with the advent of *flex fuel* cars in 2003, it has experienced a rapid growth of sugarcane cultivated areas for ethanol production in its South-central region (Mello *et al*., 2009). Besides energy concerns, the cultivation of sugarcane in Brazil plays an important role in the development of its agriculture, economy, and environment (Goldemberg, 2007; Nass *et al*., 2007).

As sugarcane cultivation covers millions of hectares in the South-central region of Brazil, remote sensing images acquired in specific periods of the agricultural calendar have been used by Canasat Project (www.dsr.inpe.br/laf/canasat/en) since 2003 to elaborate annual thematic maps of cultivated sugarcane areas in this region (Rudorff *et al*., 2010). These maps have been used as basis for sugarcane harvest monitoring (Aguiar *et al*., 2011), for assessment of changes in land use and cover due to the expansion of sugarcane cultivation (Adami *et al*., 2012), and for crop yield analyses (Sugawara, 2010). Therefore, it is still necessary to assess the quality of Canasat Project sugarcane thematic maps.

Thus, this work aims at assessing the thematic accuracy of Canasat Project sugarcane mapping.

2. Material and methods

The thematic accuracy assessment of Canasat Project mapping was based on the thematic map of sugarcane areas in the 2010/2011 crop year in the South-Central region, which comprises the states of São Paulo, Minas Gerais, Paraná, Mato Grosso, Mato Grosso do Sul and Goiás. We used: *(i)* 396 images from the TM/Landsat-5 collected between January 2009 and September 2010; *(ii)* times series of MODIS data (MOD09 product) from February 2000 to December 2009; *(iii)* a partial sugarcane map for São Paulo State available at its Secretary of the Environment (SMA-SP); and *(iv)* information on cultivated sugarcane areas by municipality available at IBGE (2011) – the official agency responsible for agricultural statistics in Brazil. All TM/Landsat-5 images were registered based on the orthorectified mosaics from the ETM+/Landsat-7 (Tucker *et al.*, 2004) archiving squared mean errors of less than 0.5 pixels. All the data were integrated in a web platform, using the remote sensing time series virtual laboratory, as described by Freitas *et al.* (2011).

2.1. Statistical design

A stratified random sampling was conducted with the strata chosen as a function of the percentage of the municipality area covered by sugarcane, given by

$$\varphi_i = \frac{S_i}{A_i} \tag{1}$$

where: S_i is the IBGE's estimated sugarcane area within the i^{th} municipality; and A_i is the total area of the i^{th} municipality. Euclidean distances were computed considering the values of $\boldsymbol{\varphi}$ in the grouping analysis using the Ward method (Ward Jr, 1963) as the clustering method, resulting in a dendogram (see Figure 1) which enables the selection of four strata. The municipalities where S=0 were not included in the analyses. The minimum sample size, considering all the study area, was defined by the binomial function as

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 pq}{E^2} \tag{2}$$

where: n is the sample size; $Z_{\alpha/2}$ is the tabulated value for the standard normal distribution with a significance level of α (adopted as 1%) for the bilateral analysis; p is the probability of occurrence of the *sugarcane* class, given by the mean of the values calculated by Equation (1) for all M municipalities considered ($p = \bar{\varphi}$); q is the probability of occurrence of the *non-sugarcane* class, given by $q = p - 1$; and E is the permitted sample error (adopted as 2.5%). In agreement with Cochran (1977), the minimum sample size in each stratum (n_h), utilizing the optimal allocation (Stehman, 2012) is calculated as

$$n_h = n\frac{P_h . sd(\boldsymbol{\varphi}_h)}{\sum P_h . sd(\boldsymbol{\varphi}_h)} \tag{3}$$

where: n is the previous minimum sample size considering all the study area, calculated from Equation (2); P_h is the number of elements (pixels) of the stratum h and sd($\boldsymbol{\varphi}_h$) is the standard deviation of $\boldsymbol{\varphi}$ in the stratum h.

Since the web platform facilitated the analysis of points, we sampled an excessive number of points (*N;* ~50% greater than *n*) to investigate, through the Monte Carlo method (Robert and Casella, 2010), the distribution of the accuracy indices used rather than having a unique analysis using *n* points. The total sampled points within each stratum (N_h) was arbitrarily defined with the condition $N_h > n_h$. It is worth mentioning that sampled points located on the edge of the analyzed classes were relocated to fit into the polygon to which they belonged, since the focus is on thematic and not positional accuracy.

2.2. Reference dataset

Using the data in the web platform, the *N* sampled points were individually visually interpreted by four skilled interpreters and labeled as either *sugarcane* or *non-sugarcane*. One of the four interpreters is specialized in sugarcane mapping so his classification prevailed over the other three in case of disagreement.

In order to confirm the suitability of the web platform in the process of constructing the reference dataset, 362 points (from *N*) were field checked along 2,620 km traveled in the states of São Paulo, Minas Gerais and Paraná, the largest national producers (IBGE, 2011). All these points agreed with the label given by the interpreters. The reference dataset for the thematic accuracy assessment was then composed by the results of the visual interpretation in the web platform.

We conducted 5,000 random drawings, sampling *n* from *N* points in each drawing and constructing a confusion matrix which allowed the computation of the overall Accuracy (*O*), sensitivity (*S*) and specificity (*E*) indices. These indices were calculated both by considering all *n* points and by stratum. The latter was calculated using the weight of each stratum (w_h), given by $P_h / \sum_{i=1}^{4} P_i$ (Cochran, 1977), where the denominator represents the sum of pixels in all strata.

3. Results and discussion

Considering the four strata defined by the grouping analysis, Table 1 summarizes the following parameters: min(φ_h) and max(φ_h) – corresponding to minimum and maximum values of φ in the stratum; $\bar{\varphi}_h$ and sd(φ_h) – mean and standard deviation of φ in the stratum *h*; M_h – number of municipalities in the stratum *h*; P_h – number of TM/Landsat-5 corresponding pixels (30 × 30 m) contained in the stratum *h*; n_h – minimum sample size in each stratum *h*, defined by Equation (3); N_h – total sampled points within each stratum; and w_h – the weight of the stratum *h*.

Table 1: The four strata and their limits, parameters, sample sizes and metrics.

Stratum	**A**	**B**	**C**	**D**
Limits (φ in %)	**(0 ; 5.5]**	**(5.5 ; 27]**	**(27 ; 53]**	**(53 ; 100]**
min(φ_h)	0.003%	5.555%	27.332%	53.961%
max(φ_h)	5.472%	26.926%	52.832%	83.299%
$\bar{\varphi}_h$	2.076%	15.067%	37.839%	65.945%
sd(φ_h)	504.008	1,894.428	2,407.126	2,398.604
M_h	286	343	199	74
P_h	12,495,627	28,040,236	24,634,031	25,620,349
n_h	36	307	342	355
N_h	104	396	504	500
w_h	0.1376	0.3089	0.2713	0.2822

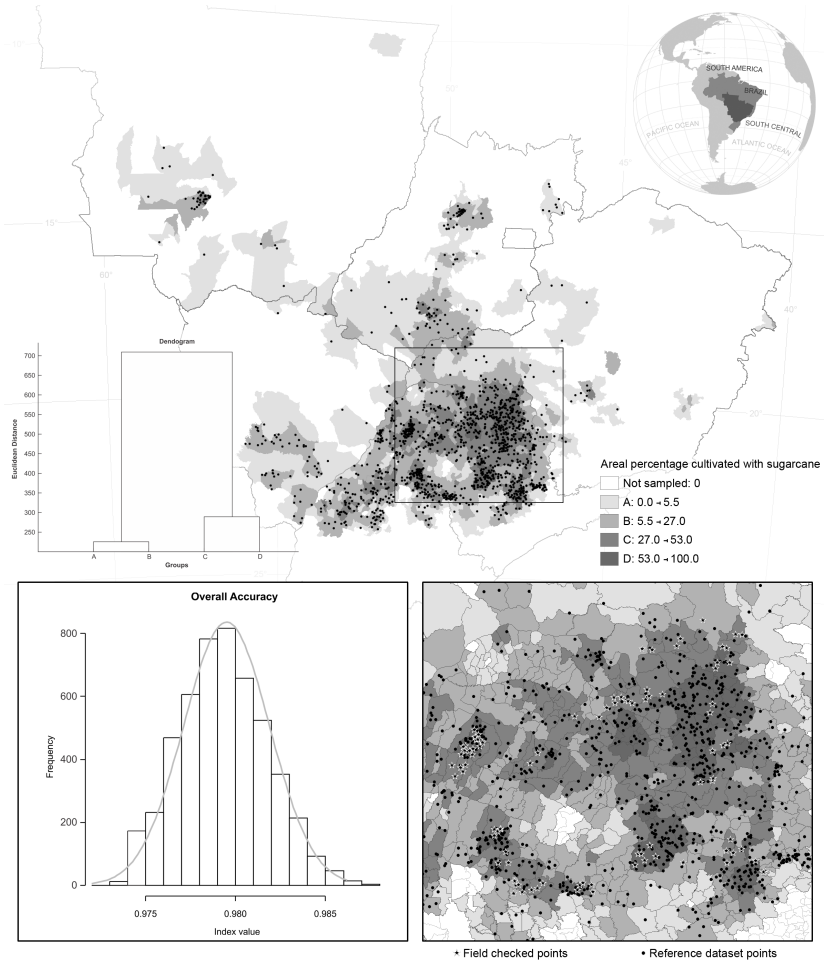

Figure 1: Study area, highlighting the four strata (defined in the dendogram); and histogram with the overall accuracy index (*O*) distribution based on 5,000 random drawings.

The web platform permitted the interpreters to assess more than the minimum of 1040 points (*n*) defined by Equation (2). Thus, 104, 396, 504 and 500 points in the A, B, C and D strata, respectively (N_h related with each stratum), were evaluated, totaling 1504 points (*N*). The field campaign assessed 24.1% (362) of the 1504 points in the reference dataset. Of a total of 1504 interpreted points in the reference dataset, 1040 points, obeying the values of n_h (as described in Table 1), were randomly drawn. Thus, by comparing, for each point, the label attributed in the Canasat Project mapping with that attributed in the reference dataset, it was possible to calculate the values of *O*, *S* and *E* considering all *n* points and each stratum. The random drawing and the calculation of the indices were conducted 5,000 times.

Figure 1 shows the dendogram, the spatial distribution of the four strata along the study area, the sampled points and the histogram with the 5,000 values of *O* computed considering all *n* sampled points in each draw. The *O* values were normally distributed around a mean close to 98% (97.96%; Table 2). Moreover, 95% of the 5,000 drawings presented *O* between 97.5% and 98.5%. The *S* and *E* indices also presented similar behavior to that of *O*. Table 2 summarizes the mean and

standard deviation values for the accuracy indices based on the 5,000 drawings, computed both for all (n) points and for each stratum.

Table 2: Descriptive statistic, considering 5,000 random drawings, for the accuracy indices, computed both for all (n) points and for each stratum.

Index	Statistic	Stratum				All points
		A	B	C	D	
O	mean	97.10%	97.72%	97.62%	98.60%	97.96%
	sd (x 10^{-2})	2.2747	0.4040	0.4532	0.3823	0.2387
S	mean	≈ 100%	98.96%	97.61%	97.64%	98.09%
	sd (x 10^{-2})	≈ 0	0.3901	0.6413	0.6148	0.3252
E	mean	94.50%	96.54%	97.63%	99.59%	97.84%
	sd (x 10^{-2})	4.3251	0.6970	0.6498	0.6257	0.3525

O – overall accuracy; *S* – sensitivity; and *E* – specificity indices.

All of the index means were greater than 96.5% (Table 2), except E for stratum A ($E_A = 94.50\%$). The reduced number of samples ($N_A = 104$; Table 1) together with the fact that the municipalities in this stratum presented a reduced percentage of sugarcane (at most 5.5% of their areas) contributed to the fact that the omission errors of the *sugarcane* class in this stratum were not observed in the sampling ($S_A \approx 100\%$). Thus, the omission errors of the *non-sugarcane* class were directly responsible for the lower accuracies compared with the other strata. In short, Canasat Project mapping overestimated in about 5.50% and 2.15% the sugarcane areas in the municipalities of strata A and B, respectively. That might be associated to sugarcane expansion in the vicinity of pasture lands which can cause interpretation errors especially for well cultivated pasture (Adami *et al*., 2012).

In stratum C, the sugarcane area estimate did not present either overestimation or underestimation since the mean omission error was compensated by the mean inclusion error. On the other hand, in municipalities with large areal percentages covered with sugarcane cultivation (stratum D) Canasat Project mapping underestimated the sugarcane area in 1.95%. In these municipalities, agricultural aptitude is principally due to climatic conditions and infrastructure, rendering greater agricultural diversity. Thus, among the cultivated crops, some can present similar characteristics in Landsat images to those of sugarcane in certain cultivation stages, such as corn, which before senescence has a similar color and texture to those of sugarcane crops and could generate confusion leading to omission errors for the *sugarcane* class (Rudorff *et al*., 2010).

Although the overall mean error associated with the thematic mapping in Canasat Project was around 2% (i.e. $O \approx 98\%$; Table 2), the process of estimating sugarcane area should present a smaller error value, since omission errors are supposed to be compensated by inclusion errors, and vice-versa. In fact, the mean thematic error associated with sugarcane area estimates was close to –0.7%, which was calculated using a weighted mean of the strata, where the individual weights were computed by multiplying the area of the stratum (obtained from P_h; Table 1) by the average sugarcane proportion within the stratum ($\bar{\varphi}_h$; Table 1). In other words, Canasat Project mapping underestimated in 0.7% the total sugarcane area in the South-central region of Brazil for the 2010/2011 crop year, which represented an omission of about 57 thousand ha in a total of more than 8.3 million ha. It is worth mentioning that this error was only in regard to the thematic accuracy assessment, since this work did not comprise positional accuracy assessment.

4. Final considerations

Based on field observations, we could state that the web platform integrating Landsat images, high spatial resolution images, time series of MODIS sensor data and ancillary data allowed skilled interpreters to build an accurate reference dataset based on visual interpretation. Results showed that Canasat Project mapping overestimated the sugarcane area in municipalities where less than 27% of their areas were covered by sugarcane cultivation and underestimated the sugarcane area in municipalities where more than 53% of their areas were covered by this crop.

Despite a mean overall accuracy index of about 98%, the mean thematic error associated with the estimate of sugarcane areas was of –0.7%, since part of the omission errors were compensated by the inclusion errors. Indeed, the thematic accuracy assessment conducted stated that the procedure used to generate the thematic sugarcane maps of Canasat Project is accurate and provides maps with high confidence levels. Although the assessment was performed for only one crop year, very similar results can also be expected for the other crop years of Canasat Project owing to the consistent method (see Rudorff *et al*. (2010)) and the careful visual interpretation of an experienced team.

References

Adami, M., Rudorff, B.F.T., Freitas, R.M., Aguiar, D.A., Mello, M.P. (2012), "Remote sensing time series to evaluate direct land use change of recent expanded sugarcane crop in Brazil". *Sustainability*, Vol. 4(4):574-585.

Aguiar, D.A., Rudorff, B.F.T., Silva, W.F., Adami, M., Mello, M.P. (2011), "Remote sensing images in support of environmental protocol: Monitoring the sugarcane harvest in São Paulo State, Brazil". *Remote Sensing*, Vol. 3(12):2682-2703.

Cochran, W.G. (1977), *Sampling techniques*, 3rd ed, John Wiley & Sons, New York, NY, USA, 428p.

Freitas, R.M., Arai, E., Adami, M., Ferreira, A.S., Sato, F.Y., Shimabukuro, Y.E., Rosa, R.R., Anderson, L.O., Rudorff, B.F.T. (2011), "Virtual laboratory of remote sensing time series: visualization of MODIS EVI2 data set over South America". *Journal of Computational Interdisciplinary Sciences*, Vol. 2(1):57-68.

Goldemberg, J. (2007), "Ethanol for a sustainable energy future". *Science*, Vol. 315(5813):808-810.

IBGE – Brazilian Institute for Geography and Statistics (2011), "Sistema IBGE de Recuperação Automática (SIDRA) – Produção Agrícola Municipal (PAM)", Rio de Janeiro, RJ, Brazil, Available at http://www.sidra.ibge.gov.br/ (assessed on May 2nd, 2011).

Mello, A.Y.I., Espindola, G.M., Alves, D.S. (2009), "Perspectives on ethanol use by the transportation sector in Brazil". *In*: Dias, P.L.S., Ribeiro, W.C., Neto, J.L.S., Zullo Jr, J. (eds.). *Public Policy, Mitigation and Adaptation to Climate Change in South América*, IEA-USP, São Paulo, SP, Brazil, pp. 103-116.

Nass, L.L., Pereira, P.A.A., Ellis, D. (2007). "Biofuels in Brazil: An overview", *Crop Science*, Vol. 47(6):2228-2237.

Robert, C., Casella, G. (2010), *Introducing Monte Carlo methods with R*, Springer, New York, NY, USA, 284p.

Rudorff, B.F.T., Aguiar, D.A., Silva, W.F., Sugawara, L.M., Adami, M., Moreira, M.A. (2010). "Studies on the rapid expansion of sugarcane for ethanol production in São Paulo State (Brazil) using Landsat data", *Remote Sensing*, Vol. 2(4):1057-1076.

Stehman, S.V. (2012), "Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change". *Remote Sensing Letters*, Vol. 3(2):111-120.

Sugawara, L.M. (2010), *Variação interanual da produtividade agrícola da cana-de-açucar por meio de um modelo agronômico*. PhD Thesis, National Institute for Space Research (INPE), São José dos Campos, SP, Brazil.

Tucker, C.J., Grant, D.M., Dykstra, J.D. (2004), "NASA's global orthorectified Landsat data set". *Photogrammetric Engineering and Remote Sensing*, Vol. 70(3):313–322.

Ward Jr, J.H. (1963), "Hierarchical grouping to optimize an objective function". *Journal of the American Statistical Association*, Vol. 58(301):236-244.