Sugarcane yield estimation in São Paulo State - Brazil

William Foschiera[1],
Marcio Pupin Mello[2],
Clement Atzberger[3],
Antonio R. Formaggio[1]

[1]National Institute for Space Research (INPE)
Remote Sensing Division (DSR)
Avenida dos Astronautas 1758 – 12227-010 – São José dos Campos, SP – Brazil
wfoschiera@gmail.com, formag@dsr.inpe.br

[2]The Boeing Company
Boeing Research & Technology – Brazil (BR&TB)
Estrada Dr Altino Bondesan 500 – 12247-016 – São José dos Campos, SP – Brazil
marcio.p.mello@boeing.com

[3]University of Natural Resources and Life Sciences (BOKU)
Institute of Surveying, Remote Sensing and Land Information (IVFL)
Peter Jordan Straße 82, Vienna, Austria 1190.
clement.atzberger@boku.ac.at

Abstract. *This paper presents a method for crop yield forecast based on remote sensing and official data. The method uses a statistical approach to extract different pixels of smoothed NDVI data derived from MODIS sensor to be used as proxies for sugarcane yield estimation at a municipal scale. From 368 municipalities with yield's historical data from 2003 to 2012, three groups were created based on acreage percentile and then 30 municipalities was randomly selected, 10 for each group. Two municipalities with extreme acreage values (minimum and maximum) were discarded from each group and two different approaches were tested to normalize yield data and NDVI: Zscore and Rscore. In addition, two methods were used as selection criteria: RMSE and Pearson's correlation. Results showed that municipalities with large sugarcane acreage tended to present better agreement between observed and estimated yield, which reinforce the potential of this method to be operationally used for sugarcane yield forecast over large areas. Moreover, the proposed method may estimate yield early in crop season, whereas official statistics are usually published late after harvest.*

Keywords: remote sensing, image processing, python, statistical yield estimation.

1. Introduction

Crop yield information provided early in the crop season is essential to appropriately plan storage, price, food security and stock, also preventing excessive market speculation (Naylor, 2011). When supported by accurate and spatialized information, decision makers are able to decide quickly and identify geographically regions with large variations in production and productivity, which may significantly impact public policies (Atzberger, 2013).

According to Atzberger (2013), an agricultural monitoring system should be able to provide information on crop production, status and yield in a standardized and regular manner. Estimates should be provided as early as possible during the growing season and updated periodically throughout the crop season. According to Roughgarden et al. (1991), sophisticated modeling tools combined to remotely sensed data is the more suitable way to provide such information over large areas with reasonable costs. However, covering wide areas with remotely sensed data demands heavy computing processing and large storage capacity.

In Brazil, official agricultural statistics have been conducted using subjective methods, based on interviews with supply chain and producers. The Brazilian Institute of Geography and Statistics (IBGE) and the National Supply Company (Conab) are the official agencies for such agricultural statistics in Brazil. One constraint of this approach is the time required to publish the information, which may take up to two years after harvest. Nonetheless, this model do not provide any statistical error associated. However, despite these issues, this is by far the best survey of the country's agricultural production.

In order to improve agricultural statistics in Brazil, the yield monitoring by remotely sensed data has been investigated over the last decades. Models based on weather, spectral data or combination of both has been widely investigated, but few were able to achieve good matching between estimated yield and the yield published by official agencies in Brazil (Rudorff and Batista, 1990; Rudorff, 1985; Sugawara, 2002; Victoria et al., 2012). Part of this problem is attributed to either the difficulty in obtaining good input data or to current inaccuracy in the data published by different official agencies (Sugawara, 2010). Moreover, weather based methodologies lays on lack of data available for the Brazilian territory (Assad et al., 2007) , while spectral based methodologies still faces cloud cover limitations (Sano et al., 2007; Sugawara et al., 2008).

Daily data from Moderate Resolution Imaging Spectroradiometer (MODIS) based products has been successfully used to minimize cloud cover limitations due to an almost daily temporal resolution. The Institute of Surveying, Remote Sensing and Land Information (IVFL) of the University of Natural Resources and Life Sciences (BOKU) created a seven-day Normalized Difference Vegetation Index (NDVI) composite data derived from MODIS filtered by Smooth Whittaker algorithm (Vuolo and Mattiuzzi, 2012)

This paper proposes a replicable method to estimate crop yield in near real time based on BOKU's remote sensing data and official agricultural data in Brazil, without using a cropland mask. The proposed method draws on earlier research described by Kastens et al. (2005), Atzberger (2013) and Mello et al. (2014).

2. Study Area

368 municipalities of São Paulo State were previously investigated for the period of 2003 to 2012 to select three groups of municipalities that represent the sugarcane planted conditions. These 368 municipalities were ordered by acreage and divided into three groups based on percentiles. These groups were named A for municipalities with the low sugarcane acreage (0-33 percentile), B for municipalities with averaged acreage (percentile 34-66) and C for municipalities with the high acreage (percentile 67-100). Ten municipalities were then randomly selected from each group (Table 1). Official sugarcane acreage published by IBGE (2014a) was used as reference for both training (proxy selection – to be detailed further – and accuracy assessment). Study area map and geographical location of selected municipalities are presented in Figure 1.

Table 1. Municipalities selected for this study, harvested sugarcane acreage, in hectares, for 2012, and group assigned to each municipality.

Municipality	Group	Area (Ha)	Municipality	Group	Area (Ha)	Municipality	Group	Area (Ha)
Capela do Alto	A	600	Cajobi	B	8006	Bariri	C	23800
Indaiatuba	A	2100	Dumont	B	8500	Barretos	C	64554
Ipiguá	A	2500	Itapeva	B	5500	Guará	C	23000
Paulistânia	A	1361	Nipoã	B	6800	Jaborandi	C	20000
Pedreira	A	30	Pacaembu	B	8500	Nhandeara	C	13718
Piracaia	A	40	Rio Claro	B	10400	Orindiúva	C	22100
Pracinha	A	363	Santa Maria da Serra	B	5280	Paulo de Faria	C	27900
Ribeirão Corrente	A	2500	São José do Rio Preto	B	6700	Penápolis	C	37869
Santópolis do Aguapeí	A	3997	Severínia	B	8000	Rio das Pedras	C	16340
Timburi	A	156	Torrinha	B	9100	Santa Adélia	C	24500

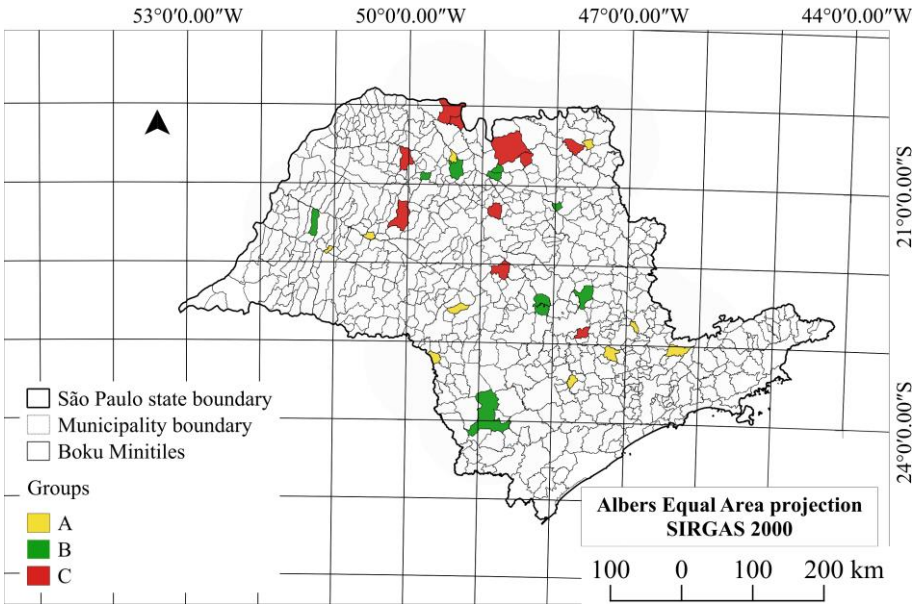

Figure 1. Sao Paulo State with selected municipalities detached.

To avoid extreme values influencing our analyses, we excluded municipalities with both minimum and maximum acreage values from each group.

3. Input Data and Proxies Selection

The remotely sensed data used is part of a composed 7-day smoothed real-time MODIS NDVI time series (MOD13) and available in tiles of 1 x 1 degree, called minitiles (Vuolo and Mattiuzzi, 2012). This data set has been weekly updated (every Monday), and each update contains around 1 Gb, with a total storage requirement of approximately 1.3 Tb. The data coming from BOKU contains 5 filtered products: N (Near real time), which regards to the last observed value from MODIS, A (1 week lag), B (2 weeks lag), C (3 weeks lag) and D (4 weeks lag).

In addition official yearly production and acreage data obtained from IBGE (2014a) were stored in tabular format for later reference. Vector data, obtained from BOKU were used to delimit the minitiles scenes, also official municipalities boundaries was obtained from IBGE (2014b) and used to calculate municipalities centroids.

The data processing was performed on a desktop PC (Intel Core I7-4770 CPU 3.40GHz, 16 gigabyte RAM and 4 Tb of available storage) running Ubuntu 14.04. Weekly download of BOKU data has being performed through a Python script, which is the programming language of all processing algorithms. Python language was chosen because of its simplicity to program. Moreover, it is the native language for some Geographic Information Systems (GIS), such as ArcGis and Qgis. In Python, raster data can be process entirely through the numpy package, widely known for efficient array processing. On the other hand, vector data can be processed via GDAL package.

The script allows one to set some parameters. Using a shapefile with municipalities' centroid coordinates, the script can process different buffer sizes, according to user's needs. In this paper, we adopted a buffer of 100 km. A mosaic was created to cover the whole area (Figure 1), including the buffers (see Figure 2), that were used to restrict the pixels to be considered for each municipality. The script also allows setting a Lag Time (LT) and an Integration Time (IT). This means that with a LT of zero, the script will consider only the N filtered product, while with a LT set as four, the script will consider the D filtered product from BOKU. IT allows to set weekly data to be integrated, i.e. an IT of equal to one will just use the scene of base date (2003083, a Monday), whereas an IT set as four will use the averaged values for dates 2003083 to 2003062 (four Mondays). To understand the influence of this smoothed data, two scenarios were chosen: Scenario 1, with LT=0 and IT=1 (product N only); and Scenario 2, with a LT=3 and IT=5. (average considering five products: one C product with three weeks leg from the base date and four D products, considering the 4th week before the basedate until the 7th week).

Relied in previous research performed by Mello et al. (2014), we defined the start year (2003), the final year (2012), and the Day of Year (DOY – 83). This DOY was chosen as the processing basedate due to its suitability to predict with good accuracy, as pointed out by Mello et al. (2014).

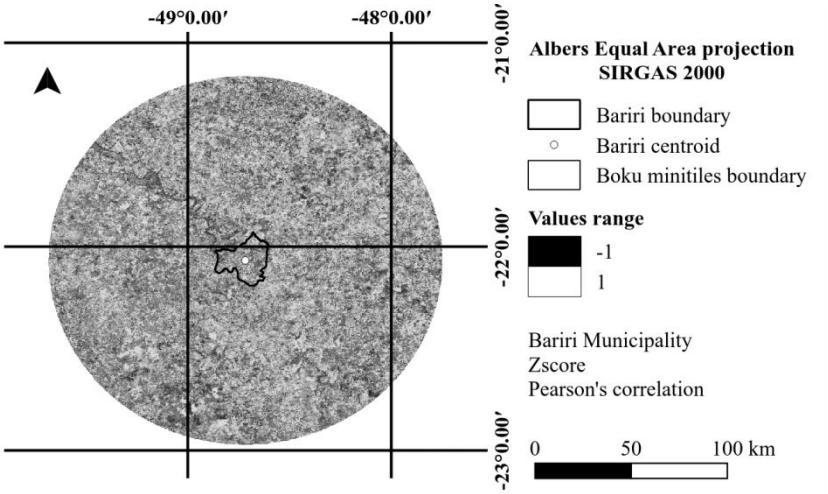

Figure 2. Example of a raster file with Pearson's correlation calculated between yield data and NDVI values for Bariri municipality in São Paulo State.

The proxy selection involves two steps: first, a normalization method shall be used to transform yield and NDVI to a common scale; second, a metric is used to compare yield and NDVI time series aiming to select the pixels that best represent yield variation for the municipality. In this paper, 200 pixels for each municipality were selected as proxies. We used two normalization methods: Zscore (Equation 1) and Rscore (Equation 2).

$$Z_i = \frac{v_i - \mu(v)}{\sigma(v)} \quad (1)$$

where:Z_i is the i-th Z-scored value; V is a vector with the observed values (either for official yield in a given municipality or for NDVI in a given pixel); v_i is the i-th value of V; $\mu(v)$ is the average value of v; and $\sigma(v)$ is the standard deviation value of V.

$$R_i = \frac{v_i - min(v)}{max(v) - \min(v)} \quad (2)$$

whereR_i: is the *i-th* R-scored value; V is a vector with the observed values (either for official yield in a given municipality or for NDVI in a given pixel); v_i is the *i-th* value of V; $\min(v)$ is the minimum value of V; and $max(v)$ is the maximum value of V.

Original NDVI values were normalized using both Zscore and Rscore. To compare NDVI values and yield data derived from IBGE, yearly values was extracted in order to create a list of values for each pixel, as shown in Figure 3.

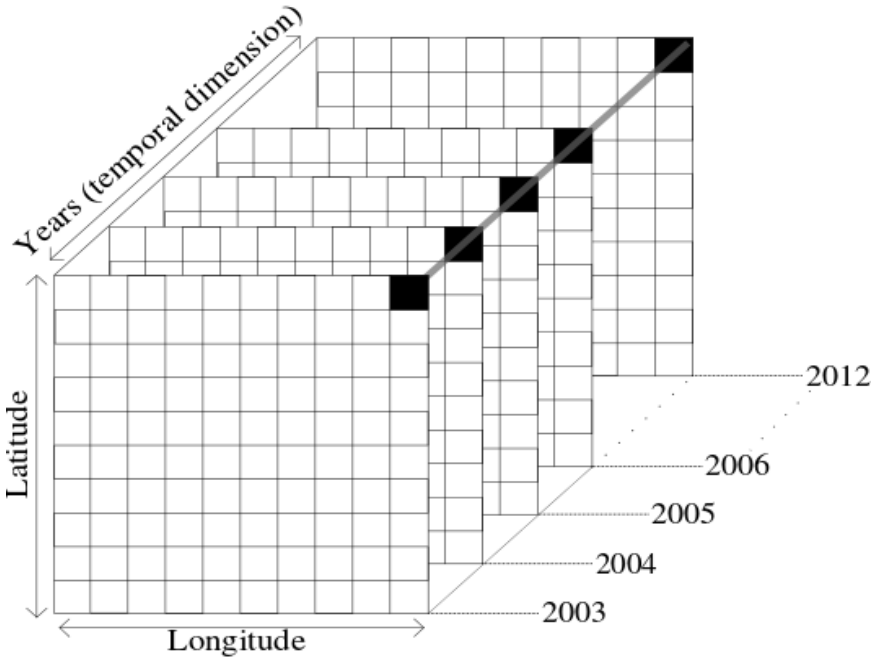

Figure 3. Example of a stacked raster to extract NDVI values arrays

After normalization, the Root-Mean-Squared Error (RMSE) and Pearson's correlation was computed to select the proxies. For Pearson's correlation 200 pixels which values were close to +1 were extracted. 200 pixels were also selected based on the RMSE, but considering values close to 0. For each year (2003 to 2012) a cross-validation was performed using the leave-one-out approach, as shown in Figure 4. In this approach, NDVI values for the year under analysis were set aside from the proxy selection procedure to be used later for accuracy assessment (validation).

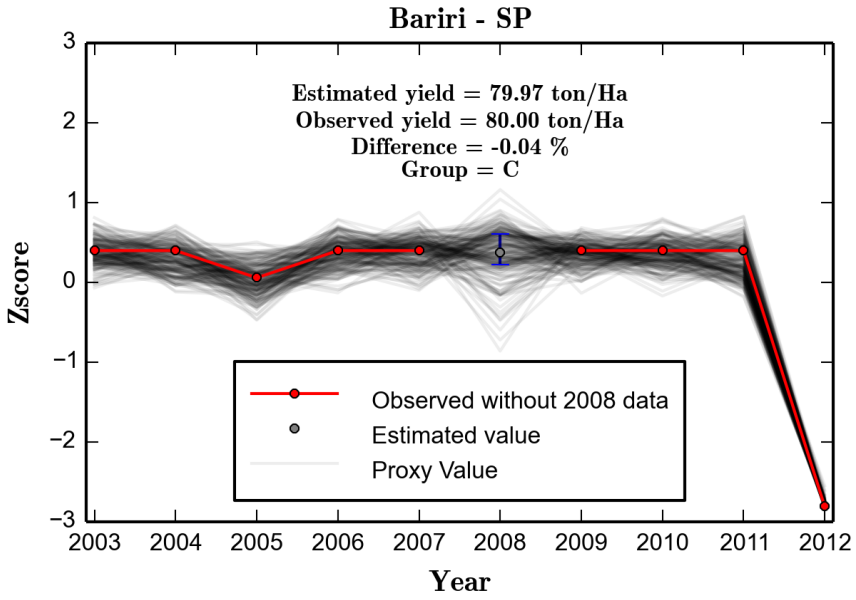

Figure 4. Leave-one-out example for Bariri municipality, highlighting: (gray lines) proxies values for 2008 year; (red line) Zscore normalized yield observations; (gray dot) estimated yield value for 2008.

Images for each municipality with Zscore and Rscore normalizations as RMSE and Pearson's correlation were generated as shown in Figure 2. Each of these images was temporarily stored in and then the 2000 best fitted values were saved as tabular data containing latitude and longitude of the selected pixels, leave-one-out year, values of RMSE or Pearson's correlation and normalized NDVI data.

4. Results, discussion and concluding points

The mean value of 200 proxies of each year and municipality were used for yield estimation. Estimated and observed yield was evaluated and mean error and RMSE were computed. Each scenario involving the 30 selected municipalities toke about 20 hours to be processed. Most of the processing time was spent on the pixel-to-pixel calculation of Pearson's correlation. Summary information of scenario 1 is presented in Table 2 and for scenario 2 in Table 3.

Table 2. Summary of scenario 1 (IT=1, LT=0). RMSE[1] means root-mean-squared errors between yield data and NDVI while RMSE[2] means root-mean-squared errors of relative estimated and observed yield values and Pearson's correlation.

Scenario 1					
Agreement measures	Groups	Zscore		Rscore	
		Mean Error (%)	RMSE[2] (%)	Mean Error (%)	RMSE[2] (%)
Pearson's correlation	A	1.37	42.50	7.68	66.54
	B	0.64	29.40	2.38	30.38
	C	0.20	22.59	0.25	22.51
RMSE[1]	A	1.65	40.03	4.99	54.92
	B	1.26	30.04	2.00	30.85
	C	0.70	22.30	0.90	22.40

Although we have not performed a statistical test, mean errors indicated that Rscored normalization tended to overestimate yield more than Zscore. Mean errors and RMSE¹ also indicated that municipalities of group C apparently showed better results than group A, and also B, which may indicate that the proxy methodology works better for municipalities with high acreage.

Table 3. Summary of scenario 2 (IT = 5, LT = 3). RMSE¹ means root-mean-squared errors between yield data and NDVI while RMSE² means root-mean-squared errors of relative estimated and observed yield values and Pearson's correlation.

Scenario 2					
Agreement measures	Groups	Zscore		Rscore	
		Mean Error (%)	RMSE² (%)	Mean Error (%)	RMSE² (%)
Pearson's correlation	A	1.00	42.91	8.88	70.90
	B	0.41	30.12	2.50	31.20
	C	0.02	21.83	0.40	23.14
RMSE¹	A	1.96	41.18	4.26	57.66
	B	1.38	30.78	2.08	31.07
	C	0.76	21.28	1.26	22.86

Comparing agreement measures (Table 2), the mean error values showed small differences between Pearson's correlation and RMSE¹ when Zscore normalization was used. Although Pearson's correlation seemed to lead to small differences between observed and estimated yield when compared to RMSE¹, the value of RMSE² between observed and estimated data showed that when using RMSE¹ the method fits better than Pearson's correlation.

Based on this preliminary analysis, Zscore normalization might be preferred upon Rscore. In addition, Pearson's correlation had slightly better results with Zscore normalization, however when Rscore was used, RMSE¹ show better results than Pearson's correlation for group A, but quite discreet for groups B and C, when exceeded.

The proposed approach achieved good results, particularly for municipalities of group C. The next steps for this research will involve improvements to minimize processing time and test different scenarios for a number of combination for IT, LT and DOY basedate.

Acknowledgements

The authors would like to thank Coordination for the Improvement of Higher Education Personnel (CAPES) for their financial support

References

Assad, E. D.; Marin, F. R.; Evangelista, S. R.; Pilau, F. G.; Farias, J. R. B.; Pinto, H. S.; Júnior, J. Z. Sistema de previsão da safra de soja para o Brasil. **Pesquisa agropecuária brasileira**, v. 42, n. 1, p. 615–625, 2007.

Atzberger, C. Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs. **Remote sensing**, v. 5, n. 2, p. 949–981, 2013.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Produção Agrícola Municipal 2012**. Disponível em: <http://sidra.ibge.gov.br/bda/tabela/listabl.asp?z=t&c=1612>. Acesso em: 15 jan. 2014.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Estados@**. Disponível em: <http://www.ibge.gov.br/estadosat/perfil.php?sigla=sp>. Acesso em: 26 jan. 2014.

Kastens, J.; Kastens, T.; Kastens, D.; Price, K.; Martinko, E.; Lee, R. Image masking for crop yield forecasting using AVHRR NDVI time series imagery. **Remote sensing of environment**, v. 99, n. 3, p. 341–356, 2005.

Mello, M. P.; Atzberger, C.; Formaggio, A. R. Near real time yield estimation for sugarcane in Brazil combining remote senting and official statistical data. In: Proceedings of the 34rd IEEE International Geoscience and Remote Sensing Symposium (IGARSS). **Proc**..., Quebec, Canada: 2014

Naylor, R. Expanding the boundaries of agricultural development. **Food security**, v. 3, n. 2, p. 233–251, 2011.

Roughgarden, J.; Running, S.; Matson, P. A. What does remote sensing do for ecology? **Ecology**, v. 72, n. 6, p. 1918–1922, 1991.

Rudorff, B. F. T. **Dados landsat na estimativa da produtividade agrícola da cana-de-açúcar**. [s.l.] Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1985.

Rudorff, B. F. T.; Batista, G. Yield estimation of sugarcane based on agrometeorological-spectral models. **Remote sensing of environment**, v. 33, n. 3, p. 183–192, 1990.

Sano, E. E.; Ferreira, L. G.; Asner, G. P.; Steinke, E. T. Spatial and temporal probabilities of obtaining cloud free Landsat images over the Brazilian tropical savanna. **International journal of remote sensing**, v. 28, n. 12, p. 2739–2752, 2007.

Sugawara, L. M. **Avaliação de modelo agrometeorológico e imagens noaa/avhrr no acompanhamento e estimativa de produtividade de soja no estado do paraná**. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2002.

Sugawara, L. M. **Variação interanual da produtividade agrícola da cana-de-açucar por meio de um modelo agronômico**. Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2010.

Victoria, D. D. C.; Rolim, A.; Coutinho, A. C.; Kastens, J. Cropland area estimates using Modis NDVI time series in the state of Mato Grosso, Brazil. **Pesquisa agropecuária brasileira**, v. 47, n. 1, p. 1270–1278, 2012.

Vuolo, F.; Mattiuzzi, M.; Klisch, A.; Atzberger, C. Data service platform for MODIS Vegetation Indices time series processing at BOKU Vienna: current status and future perspectives. **Proc...** SPIE 8538, Earth Resources and Environmental Remote Sensing/GIS Applications III, 85380A (October 25, 2012); doi:10.1117/12.974857.