generar una alerta temprana de los predios que requieren asesoría técnica para mejorar la productividad. Al incluir otras variables como clima, prácticas agrícolas y características de suelo en la elaboración del modelo, este puede mejorar su desempeño. En el futuro este modelo puede adaptarse a otros cultivos de importancia para el país.

Palabras clave: aprendizaje automático, NDVI, sensoramiento proximal, sensoramiento remoto, regresión.

I. INTRODUCTION

In Ecuador, maize (*Zea mays L.*) is one of the most important crops, as it is worldwide. Production of "hard maize" is destined for animal feed, while "soft maize" production serves for human consumption. Recently, the Ecuadorian government has prioritized the elaboration of yield predictive models, in order to develop decision making strategies for imports and commercialization purposes. Demand in Ecuador is around 1.2 Mt of maize grain, while national production reaches only 0.5 Mt. [4] During 2012, the national harvesting was severely affected due to imports at lower price. Imports of maize grain attempted to satisfy the national demand. These imports should have ended before February 15th, but they continued entering the country and overlapping the national production.

Research on maize crop in Ecuador has been focused on genetic enhancement, agricultural management and crop protection practices. [9] However, little has been done on site adaptability and yield responses of the distributed varieties to a specific zone. Furthermore, there have been no attempts to develop yield estimation models to predict the production at those specific sites. Estimation of maize production is done through field data assessment and extrapolation to a county and province level, which is costly and time consuming. Conversely, in the literature many crop yield prediction models were created using remote sensing data and deriving vegetation indices (e.g. NDVI, LAI, REIP), which is much more time effective. Additionally, crop state variables and climate variables from the crop/soil/atmosphere interfaces were included in the model development to predict the crop production before harvest in different crops. [1, 13, 17, 6, 10, 11, 12] However, most of these models are confined to particular regions and/or periods, thus they cannot be applied directly. Therefore, it is required reliable yield prediction models for Ecuador.

Remote sensing data have been proven to be an effective tool for yield prediction in different crops. Vegetation indices extracted from spectral data have been employed for constructing the predictive models. [8]The Normalized Difference Vegetation Index (NDVI) has a wide application in vegetative studies as it has been used to estimate crop yields, pasture performance and rangeland carrying capacities among others. [8] NDVI is directly related to other ground parameters, such as percent of ground cover, photosynthetic activity of the plant, surface water, leaf area index (LAI) and the amount of biomass. [15] NDVI is a suitable index to timate crop production before harvesting, because it is the optical representation of vegetation canopy "greenness". NDVI gives a direct measure of photosynthetic potential resulting from the composite property of total leaf chlorophyll, leaf area, canopy cover and structure. [13, 17, 2, 12, 5, 7]

Rice yield in Egypt was predicted with the use of satellite remote sensing data. [14] They used two multi-regression models of LAI, as one input factor, and NDVI or any other vegetation index. These indices were calculated from visible and near-infrared spectral reflectance, under normal environmental conditions and common agricultural practices during the period of the maximum vegetative growth. The result was the best practice for rice yield forecasting using satellite imagery. A model to predict crop growth and yield variability using airborne multispectral and hyperspectral imagery and high-resolution satellite imagery, taken during the growing season, was proposed by Yang et al. [19] Their model can be used to monitor crop growing conditions and identify potential production problems, which could be addressed within the growing season. A yield estimation algorithm of corn and soybean in Midwestern USA, which did not require retrospective analysis to construct the empirical relationships between reported yields and remotely sensed data, was developed and proposed by Xin et al. [18] Those authors recommended that development of future yield estimation methods, based on production efficiency models, should consider the sub-pixel spatial heterogeneity and irrigation effects. Another yield estimation method using the relationship between LAI and yield was proposed by Zhang *et al*. [22] These authors developed a relationship between climate variability impact index (CVII) and crop production using historical data. The CVII-based model can provide near real-time, global coverage of the percent change in the climatological crop yield.

This study aimed to develop a yield prediction model for maize, based on spectro-radiometer readings and satellite imagery, using machine learning algorithms and fuzzy logic. It was hypothesized that vegetation indices obtained from remote sensing data fairly represented the crop productive characteristics in the field, under the variable conditions of the two regions of Ecuador where the study took place: coast and highland regions. Thus, predicted models could give an early warning for

decision making in agricultural policy to schedule imports.

II. METHODOLOGY

2.1 Experimental site

Assessments of maize cropping systems across the four major producing provinces in Ecuador, were used to develop the model. Surveyed provinces were: Los Ríos, Manabí, Guayas in the coastal zone, and Loja in the highlands. Three major production zones with their corresponding sub-regions were identified (Figure 1). Zone 1: Province of Los Ríos, counties Ventanas and Mocache; Zone 2: Provinces of Manabí and Guayas, counties Tosagua and Balzar; and Zone 3: Province of Loja, county Pindal. Within those counties some farms were monitored since sowing to harvesting date. These farms had been monitored on the above mentioned projects and a field survey was carried out in this study. Each farm had an area between 1.5 to 3.5 ha. All farmers and direct participants in the maize production chain, contributed to generate land cover and use maps (scale 1:25,000).

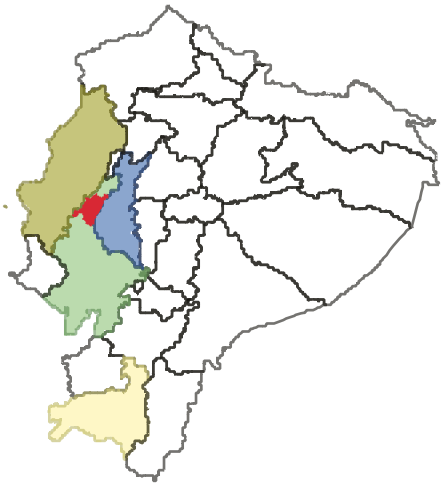

Figure 1. Maize producing Provinces in Ecuador and their respective sampled counties

Agro-climatic conditions in these provinces have traditionally shown better suitability for maize production. The coastal zone of Ecuador is influenced by the Humbolt ocean current and the warm phase of El Niño Southern Oscillation, which produce a climate combination of the Tropical savannah and Tropical monsoon, with high temperatures almost the whole year. [21] Ventanas and Mocache are located at an altitude of 6–11 m above sea level (a.s.l.) and has a clearly marked dry season between June and November with around 128 mm, while the precipitation in the rainy season reaches 1800 mm; temperatures vary between 22 and 33° C. [3] Tosagua is between 6 to 350 m a.s.l. of altitude; the dry season occurs from June through December with 85 mm, while the rainy season receives nearly 604 mm of precipitation, and the temperature oscillates from 22 to 32° C. Balzar is between 4 and 6 m a.s.l. of altitude; the dry season goes from June through December with nearly 90 mm, while the rainy season receives 1181 mm of precipitation and the average temperature is 25.6° C. Although in the highlands, Pindal has a tropical climate, its altitude is around 774 m a.s.l. The dry season lasts from June through November with nearly 302 mm, while the rainy season receives 837 mm of precipitation; the average temperature is 23° C.

2.2 Field assessments

From each farm in the selected counties, yield was measured within squares of 5 × 5 m, repeated at least two times within each farm if conditions were more homogeneous and the farm was smaller than 1 ha. The number of samples increased in bigger farms and more heterogeneous conditions. However, in some farms only one sample was allowed to be taken, in order to affect insignificantly the farmers's profit. Yield parameters measured within each square were: harvested plant density, number of cobs per plant, fresh weight of cobs and grain weight at 13% humidity. Grain yield was measured as dry weight in kg per 25 m^{-2} and the values were extrapolated to the whole field size in t ha^{-1}.

Across the three zones for maize production in Ecuador during the growing season 2013–2014, sampling areas were depicted as a grid in a map. GIS tools were applied using the software ArcGIS® version 10.1 (ESRI©, Environmental Systems Research Institute, Inc., 1995–2014) to draw the grid. The borders of the counties were marked at each maize producing zone and the surrounding area was calculated. Within each county, a grid of 250 × 250 m cell size was created, delivering a raster file of 6.25 ha pixel resolution. A total of 119 farms were sampled, which covered nearly 323 ha. The sample size was calculated using Equation (1), according to Ryan [16]:

$$n = \left(\frac{Z . \frac{a}{2} . \sigma}{E} \right)^2 \quad (1)$$

where n is the sample size, $Z\frac{a}{2}$ is the critical statistical value, the positive Z value that is at the vertical boundary of the area of $\frac{a}{2}$ in the right tail of the standard normal distribution. σ is the population standard deviation, and E is the maximum difference observed between the sample mean and the value of the population mean μ.

Spectral information was acquired across the study zones in each of the 129 farms; maize yield could be measured in 119 farms. A Fieldspec 4 Hi-Res Spectroradiometer (Analytical Spectral Devices, Inc., ASD, Boulder, Colorado) was used to measure the reflectance. This device measured the spectral range between 350 and 2500 nm, covering the visible and near-infrared spectral region (Vis / NIR). This instrument can capture spectral signatures of objects, due to its sensitivity to the radiation reflected at different wavelengths. Prior to measurements, the Spectroradiometer was calibrated using a standard white reference: Spectralon©Labsphere Inc., North Sutton, New London, USA). The optical fiber of the Spectroradiometer was pointed perpendicular to the Spectralon during 10s, and the measured value corresponded to pure or zero reflectance. A field of view (FOV) of 25° was achieved for the optical fiber, at a distance of 80 cm from the measured surface. In order to capture the spectral reflectance on maize plants, the optical fiber was perpendicularly pointed to canopy. Measurements took place at two development stages of maize, full leaf development (BBCH 17–19) and beginning of tassel emergence (BBCH 51), according to Zadoks et al. [20] Normalized Difference Vegetation Index (NDVI) was calculated from the obtained spectral signatures. NDVI calculated from spectral data taken at BBCH 17–19 is referred to hereafter as NDVI_1; NDVI calculated from spectral data assessed at BBCH 51 is referred to hereafter as NDVI_2.

2.3 Model development

Machine learning techniques were applied to develop algorithms, program and train the models. Machine learning is a prediction-making discipline in computer science that allows to create a model that “learns” from example inputs to make predictions, such that the prediction results improves with every model run. Statistical tools such as simple linear regression, logistic regression, polynomial regression and multinomial logistic regression with polynomial features were explored to propose an efficient yield estimation model. However, simple linear regression models do not consider other influential variables, such as topography and climatic conditions.

Six models were constructed using the whole country dataset of NDVI_1 and NDVI_2 calculated from spectral signatures. Simple linear regression was the basis for models 1 to 3. NDVI_1 was separately used to identify a relationship with observed yield in model 1, while NDVI_2 was used in model 2. Model 3 used combined features of NDVI_1 and NDVI_2. However, these three models did not show a significant predictive capability. Therefore, polynomial regression (model 4), multinomial logistic regression (model 5) and multinomial logistic regression using polynomial features (model 6) fitted the data better and allowed more accurate yield estimation. Cross validation was applied to all models, in order to determine the accuracy level; R^2 was also calculated when possible.

III. RESULTS AND DISCUSSION

Linear regression did not permit an accurate yield prediction, as they fitted poorly NDVI data, which was visible due to the low calculated R2 (Table 1). Equations (2), (3) and (4) show the mathematical structure of models 1, 2 and 3, respectively:

$$E_Y = 65.05 + 117.68 \cdot NDVI\_1 \quad (2)$$

$$E_Y = -20.37 + 234.86 \cdot NDVI\_2 \quad (3)$$

$$E_Y = -47.92 + 102.32 \cdot NDVI\_1 + 188.31 \cdot NDVI\_2 \quad (4)$$

where E_Y is the estimated maize yield, and NDVI_1 and NDVI_2 are the calculated NDVI at BBCH 17–19 and BBCH 51, respectively.

Table I. Comparison of accuracy of the tested maize yield estimation models.

Model	Mathematical approach	R^2	Accuracy (%)
1	Linear regression	0.43	NA *
2	Linear regression	0.35	NA
3	Linear regression	0.53	NA
4	Six degree polynomial regression	0.86	NA*
5	Multinomial logistic regression	NA	52
6	Multinomial logistic regression using polynomial features	NA	61

* Non estimable

Developed models using polynomial regression, multinomial logistic regression and multinomial logistic regression with polynomial features, delivered a better yield estimation capability than simple linear regression. In linear and polynomial regression (models 1–4), R^2 determines how well the model fits the data. However, in logistic regression R does not explain the goodness of fit to the data, because the output is binomial (either 0 or 1). Thus, the percentage of accuracy of prediction was calculated as a measure of how well the model fitted the observed data. Training accuracy was calculated with the formula *training_accuracy* = *mean_predicted_yield* x *(yield_predicted – yield_observed*) x 100.

Table I displays the comparison of accuracy for models 5 and 6. Prediction models 4–6 used combined features NDVI_1 and NDVI_2. Model 4 exhibited better prediction accuracy compared with the others and to linear regression models. Due to its best performance, six degree polynomial regression can be recommended to forecast maize yield under Ecuadorian conditions, according to the observed data. In Figure 2, sixth degree

polynomial regression (model 4) is shown. Normalized features and feature scaling were used to avoid over-fitting problems. Prediction models of higher degree of polynomial regression over-fitted the data and did not perform an acceptable prediction, while models using lower order polynomial under-fitted the data. For Models 5 and 6, "one vs. all" approach was used for multinomial logistic regression. Models 5 and 6 showed an acceptable accuracy in prediction, especially compared with linear regression models. Monitored yield data was classified in eight classes (Table II) and predicted yield was estimated within those ranges.

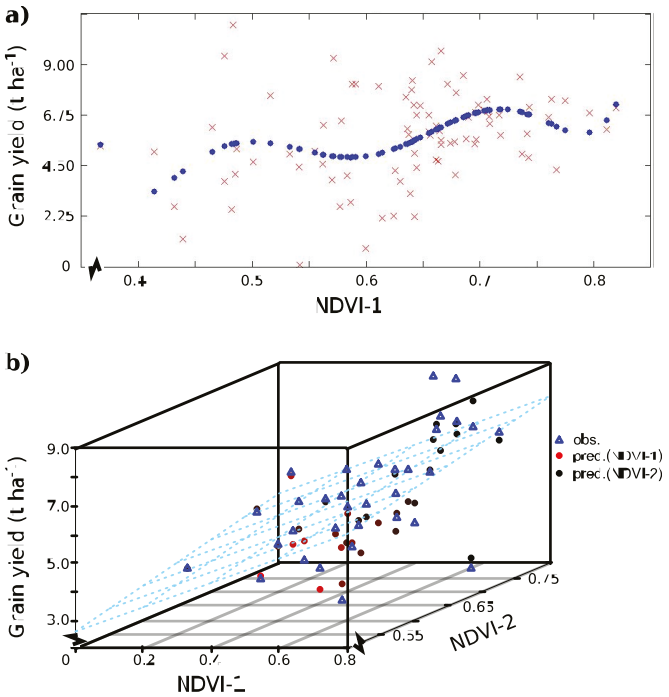

Figure 2. Yield prediction using a six degree polynomial regression algorithm (model 4) and calculated NDVI from spectral data acquired at maize stages BBCH 17-19 (NDVI-1) and BBCH 51 (NDVI-2). a) Depiction of predicted (blue symbols) and observed yield (red symbols) and NDVI-1; b) 3D comparison of observed with predicted yield and NDVI-1 and NDVI-2 values.

Table II. Classification yield range obtained from the field assessments and yield prediction using multinomial logistic regression, within those class codes.

Yield range (t ha^{-1})	Yield class code*
9.2–10	8
8.1–9.1	8
7.2–8.0	7
6.3–7.1	6
5.4–6.2	5
4.5–5.3	4
3.6–4.4	3
2.7–3.5	2
1.8–2.6	1

* Yield predicted using model 5 was within these classes

Table III. Maize yield classification and prediction using features NDVI-1 and NDVI-2 from predictive models using six degree polynomial regression compared with multinomial logistic regression using polynomial features.

NDVI-1	NDVI-2	Province	Actual yield* (t ha^{-1})	Estimated yield† (t ha^{-1})	Yield class‡
0.37	0.69	Guayas	5.4	5.3	4
0.49	0.69		5.2	4.8	4
0.52	0.56		7.6	7.5	5
0.53	0.62		5.0	4.9	4
0.54	0.5		4.4	4.5	4
0.56	0.56		4.6	5.2	4
0.63	0.71		7.5	7.7	7
0.64	0.76		2.2	2.6	1
0.64	0.78		6.1	5.9	4
0.71	0.73		7.4	8.4	4
0.02	0.65	Loja	3.3	3.3	2
0.26	0.64		5.4	5.5	5
0.34	0.63		4.4	4.3	4
0.39	0.64		5.8	4.8	5
0.51	0.66		4.7	4.9	4
0.54	0.7		6.3	6.1	5
0.55	0.55		5.7	5.2	5
0.62	0.6		4.6	4.8	4
0.67	0.64		5.2	5.4	4
0.58	0.6	Manabí	2.7	3.2	1
0.64	0.54		4.4	3.7	3
0.66	0.67		4.7	5.4	5
0.66	0.64		6.0	4.7	4
0.66	0.69		6.3	6.4	5
0.68	0.72		7.8	7.3	4
0.74	0.75		7.1	6.8	5

* Observed yield from 4 farms with similar NDVI values, within each county
†Using model 4, six degree polynomial regression, $R^2 = 0.86$ (Table I)
‡Estimated using model 6 (61% accuracy); yield class according to Table II

Table III shows the measured yield averaged from four farms having similar NDVI values, correspondingly within each county and Province. Yield data from Los Ríos showed many inconsistencies such as extreme low or high values, thus they were not used for prediction purposes. The predicted yield according to model 4 and predicted yield class by model 6 is presented. Both models generated an acceptable yield estimation. However, further model improvement can be done with other machine learning techniques, in order to recommend a definitive model for the whole country.

Climatic conditions, soil type and crop management practices need to be included in the model for model improvement and better yield estimation. Among the climatic data, temperature data, humidity and precipitation would provide the algorithm with more specificity, in order to perform better to local (county) assessments. Likewise, soil type, soil texture, water capacity retention and soil fertility can be variables which would improve the predictive capability of the algorithm. Moreover, agricultural practices such as fertilization, irrigation, pest management can be

included in the model, in order to determine which variable in fact affects the yield. The majority of higher yield ranges and classes were obtained in the provinces of Guayas, followed by Manabí. Therefore, strategies to increase the productivity should be prioritized in the province of Loja. Model 4, using six degree polynomial regression, delivered the best yield predicting capability. This model should be evaluated in future years and locations in order to be fine-tuned with out-of-sample testing. After validation, this model could be recommended for decision making on imports strategies, to prevent the overlapping with the national production. Moreover, since NDVI indices were good features in the modeling process, NDVI extracted from satellite images could be obtained before the maize harvesting season, thus an early warning strategy can be designed for the sites requiring technical assistance and practices to improve yield. Special attention could be given to earlier development stages for calculating the NDVI, since NDVI at BBCH 17–19 showed to be better inputs for yield forecasting.

IV. CONCLUSION

Yield estimation models are used in precision Agriculture to increase yield production to meet demand and to recommend to the government in regard to decision making on imports of maize to avoid overlapping. Data were collected from four provinces with a sampled area of nearly 300 ha. In this paper, six models were tested in their yield prediction capabilities. Spectroradiometer readings were used for model inputs. Machine learning algorithms offered acceptable estimation accuracy, although higher predictive power may be obtained when other variables, such as climate, agricultural practices and soil characteristics are including in the model development. The model using six degree polynomial regression could be recommended for Ecuadorian conditions. In Ecuador, yield predictive models are not existent for any crop. Using results from this study, the Ministry of Agriculture could have a tool for decision taking about the accurate amounts of maize to be imported, and avoid the overlapping with the national production. This model can be reformulated using other crop assessments in the future, to develop strategies for increasing yield and land territorial management in other crops of importance, such as rice and potato.

Acknowledgements

The authors acknowledge the projects “Producción de semillas de alto rendimiento”, MAGAP (Ministry of Agriculture, Livestock and Fisheries of Ecuador) and “Generación de Geoinformación para Gestión Territorial” (Instituto Espacial Ecuatoriano -IEE), which defined the study area. We are also thankful to Gabriela Carrera, Grace Benavídez-Gutiérrez, Andrea Córdova-Cruzatty, Alejandra Cabrera and Christian Fernández for data collection and technical support. Special thanks to the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) of the Republic of Ecuador for financial support and to the PROMETEO project for support during writing of this manuscript.

References

[1] J. S. Ahlrichs, M. E. Bauer (1983) “Relation of Agronomic and Multispectral Reflectance Characteristics of Spring Wheat Canopies”, *Agron*. J. 75(6), 987–993.

[2] T. N. Carlson, D. A. Ripley (1997) “On the relation between NDVI, fractional vegetation cover, and leaf area index”, *Remote Sens. Environ*. 62(3), 241 – 252.

[3] Climate-Data.org (2015) “Clima: Ecuador”, Available online, Last accessed 25 July 2015.

[4] Ecuavisa (2012) “Noticias: Desperdicio de maíz destapa corrupción en Ministerio de Agricultura (News: Wasting of maize production uncovers corruption inside the Ministry of Agriculture)”, Available online, Last accessed 16 March 2015.

[5] J. C. D. M. Esquerdo, J. Z. Júnior, J. F. G. Antunes (2011) “Use of NDVI/AVHRR time-series profiles for soybean crop monitoring in Brazil”, *Int. J. Remote Sens*. 32(13), 3711–3727.

[6] C. Ferencz, P. Bognár, J. Lichtenberger, D. Hamar, G. Tarcsai, G. Timár, G. Molnár, S. Pásztor, P. Steinbach, B. Székely, O. E. Ferencz, I. Ferencz-Árkos (2004) “Crop yield estimation by satellite remote sensing”, I*nt. J. Remote Sens*. 25(20), 4113–4149.

[7] R. R. V. Gonçalves, J. Z. Jr, L. A. S. Romani, C. R. Nascimento, A. J. M. Traina (2012) “Analysis of NDVI time series using cross-correlation and forecasting methods for monitoring sugarcane fields in Brazil”, *Int. J. Remote Sens*. 33(15), 4653–4672.

[8] H. J. Heege, E. Thiessen (2013) “Sensing of Crop Properties”, In H. J. Heege, (Ed.) Precision in Crop Farming, 1 edn., pp. 103–141, Springer Netherlands.

[9] INIAP (2014) “Instituto Nacional de Investigaciones Agropecuarias (National Institute of Agricultural Research, Ecuador)”, Available online, Last accessed 14 November 2014.

[10] M. Kalubarme, M. Hooda,R.S.and Yadav, G. Saroha (2006) “Spectral vegetation indices and its response to in-situ measured leaf area index of cotton”, In ISPRS, (Ed.) Symposium of ISPRS Commission IV, 25-30 September, Goa, India, vol. Volume XXXVI Part 4, pp. 1–6, International Society for Photogrammetry and Remote Sensing, ISPRS, Commission IV, Goa, India: ISPRS.

[11] H. R. Matinfar (2013) “Modeling wheat yield estimation base upon spectral data and field measurement, case study: Razan plain, Iran”, *Technical Journal of Engineering and Applied Sciences* 3(17), 2123–2130.
[12] J. Moore, N. Holden (2003) “Examining the development of a potato crop nutrient management trial using reflectance sensing”, In ASAE, (Ed.) 2003 ASAE Annual Meeting, Paper number 031133, pp. 1–9, American Society of Agricultural and Biological Engineers (ASAE), St. Joseph, Michigan, USA: American Society of Agricultural and Biological Engineers.
[13] R. R. Nemani, S. W. Running (1989) “Testing a theoretical climate-soil-leaf area hydrologic equilibrium of forests using satellite data and ecosystem simulation”, *Agric. For. Meteorol.* 44(3–4), 245–260.
[14] N. Noureldin, M. Aboelghar, H. Saudy, A. Ali (2013) “Rice yield forecasting models using satellite imagery in Egypt”, *Egyptian Journal of Remote Sensing and Space Science* 16(1), 125–131.
[15] J. J. Rouse, R. Haas, J. Schell, D. Deering (1974) “Monitoring Vegetation Systems in the Great Plains with Erts”, In NASA Goddard Space Flight Center 3d ERTS-1 Symposium (1974), vol. 351, pp. 309–317.
[16] T. P. Ryan (2013) Sample size determination and power, Wiley Series in Probability and Statistics, Hoboken, New Jersey, USA.: John Wiley & Sons, Inc., 1–404 pp.
[17] C. Wiegand, S. Maas, J. Aase, J. Hatfield, P. P. Jr., R. Jackson, E. Kanemasu, R. Lapitan (1992) “Multisite analyses of spectral-biophysical data for wheat”, *Remote Sens. Environ.* 42(1), 1–21.
[18] Q. Xin, P. Gong, C. Yu, L. Yu, M. Broich, A. E. Suyker, R. B. Myneni (2013) “A Production Efficiency Model-Based Method for Satellite Estimates of Corn and Soybean Yields in the Midwestern US”, *Rem. Sens.* 5(11), 5926–5943.
[19] C. Yang, J. Everitt, Q. Du, B. Luo, J. Chanussot (2013) “Using High-Resolution Airborne and Satellite Imagery to Assess Crop Growth and Yield Variability for Precision Agriculture”, *Proceedings of the IEEE* 101(3), 582–592.
[20] C. Zadoks, T. Chang, F. Konzak (1974) “A decimal code for the growth stages of cereals”, *Weed Res*. 14(6), 415–421.
[21] E. Zambrano, F. Hernández (2007) “Inicio, duración y término de la estación lluviosa en cinco localidades de la costa ecuatoriana”, *Acta Oceanográfica del Pacífico* 14(1), 7–11.
[22] P. Zhang, B. Anderson, B. Tan, D. Huang, R. Myneni (2005) “Potential monitoring of crop production using a satellite-based Climate-Variability Impact Index”, *Agric. For. Meteorol.* 132(3–4), 344–358.