Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Forecasting maize yield at field scale based on high-resolution satellite imagery

Rai A. Schwalbert [a,b,c,*], Telmo J.C. Amado [b,c], Luciana Nieto [c], Sebastian Varela [c], Geomar M. Corassa [a,c], Tiago A.N. Horbe [b], Charles W. Rice [c], Nahuel R. Peralta [d], Ignacio A. Ciampitti [c,**]

[a] Agricultural Engineering Department, Federal University of Santa Maria, Rural Science Centre, Santa Maria, RS, Brazil

[b] Soil Department, Federal University of Santa Maria, Rural Science Centre, Santa Maria, RS, Brazil

[c] Department of Agronomy, Kansas State University, 2004 Throckmorton Plant Science Center, Manhattan, KS 66506, USA

[d] Monsanto Argentina, Department of Technology and Development of Corn and Sorghum, Pergamino, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 25 October 2017
Received in revised form
7 March 2018
Accepted 25 April 2018

Keywords:
Yield forecasting models
Maize
Satellite imagery
Yield maps
Model validation
Sentinel-2

Estimating maize (*Zea mays* L.) yields at the field level is of great interest to farmers, service dealers, and policy-makers. The main objectives of this study were to: i) provide guidelines on data selection for building yield forecasting models using Sentinel-2 imagery; ii) compare different statistical techniques and vegetation indices (VIs) during model building; and iii) perform spatial and temporal validation to see if empirical models could be applied to other regions or when models' coefficients should be updated. Data analysis was divided into four steps: i) data acquisition and preparation; ii) selection of training data; iii) building of forecasting models; and iv) spatial and temporal validation. Analysis was performed using yield data collected from 19 maize fields located in Brazil (2016 and 2017) and in the United States (2016), and normalised vegetation indices (NDVI, green NDVI and red edge NDVI) derived from Sentinel-2. Main outcomes from this study were: i) data selection impacted yield forecast model and fields with narrow yield variability and/or with skewed data distribution should be avoided; ii) models considering spatial correlation of residuals outperformed Ordinary least squares (OLS) regression; iii) red edge NDVI was most frequently retained into the model compared with the other VIs; and iv) model prediction power was more sensitive to yield data frequency distribution than to the geographical distance or years. Thus, this study provided guidelines to build more accurate maize yield forecasting models, but also established limitations for up-scaling, from farm-level to county, district, and state-scales.

* *Corresponding author.* Agricultural Engineering Department, Federal University of Santa Maria, Rural Science Centre, Santa Maria, RS, Brazil.

** *Corresponding author.* Department of Agronomy, Kansas State University, 2004 Throckmorton Plant Science Center, Manhattan, KS 66506, USA

E-mail addresses: rai.schwalbert@gmail.com (R.A. Schwalbert), ciampitti@ksu.edu (I.A. Ciampitti).

https://doi.org/10.1016/j.biosystemseng.2018.04.020

1. Introduction

Precise and reliable yield forecast tools could play a fundamental role in supporting policy formulation, and decision-making process in agriculture (e.g. storage and transport) (Córdoba, Bruno, Costa, Peralta, & Balzarini, 2016; Kantanantha, Serban, & Griffin, 2010; Stone & Meinke, 2005). Historically, most models developed for yield forecasting purposes are focused to large domains (between-field variability) (DiRienzo, Fackler, & Goodwin, 2000; Doraiswamy, Moulin, Cook, & Stern, 2003; Hamar, Ferencz, Lichtenberger, Tarcsai, & Ferencz-Arkos, 1996; Lopresti, Di Bella, & Degioanni, 2015; Reeves, Zhao, & Running, 2005; Sibley, Grassini, Thomas, Cassman, & Lobell, 2014), mostly because, in the past there was limited source of data with a sufficient temporal and spatial resolution for accurate within-field crop yield estimates. Nowadays, satellite data have become more accessible (Azzari, Jain, & Lobell, 2016) with more options of high resolution imagery such as Skysat, RapidEye, and Sentinel-2 satellites, and more studies have portrayed the benefits of using high-resolution satellite imagery for identifying within-field yield variation (Azzari et al., 2016; Jin, Azzari, Burke, Aston, & Lobell, 2017; Peralta, Assefa, Du, Barden, & Ciampitti, 2016). Among the high-resolution satellites, the publically accessible Sentinel-2, a joint initiative of the European Commission (EC) and the European Space Agency (ESA), represents a great opportunity towards fine-resolution yield forecast models, since it was designed to provide systematic global acquisitions of high-resolution (10- to 20-m) multi-spectral imagery with a high revisit frequency (5 days at equator) (Drusch et al., 2012).

The potential to forecast yield using satellite information is already known and a wide set of statistical approaches have been explored. Some approaches rely on the statement that total biomass production is closely related to the fraction of photosynthetically active radiation (fAPAR) absorbed by vegetation over the course of the growing season (Monteith, 1977). Estimations of fAPAR are most often derived from VIs (Lobell, 2013), since the linear relationships between those two variables are well-known (Myneni & Williams, 1994). However, considering that most remote sensing data are not available on a daily basis, some interpolation is needed to estimate daily fAPAR.

Empirical relationships between ground-based yield measures and remote sensing data have been considered as the simplest approach to forecast yield with low computational power demanding (Hatfield, Gitelson, Schepers, & Walthall, 2008; Lobell, 2013), and have been successfully implemented in several studies with maize (Bognár et al., 2011; Bu, Sharma, Denton, & Franzen, 2017; Lobell, Thau, Seifert, Engle, & Little, 2015; Peralta et al., 2016; Shanahan et al., 2001; Sibley et al., 2014). The success of this approach is directly related to the selection of ground-truth data to build models. During the model building process the separation of data into training and validation datasets is a common practice allowing self-test model replicability irrespective of the difference between the two datasets in space or time. The selection of training data is known to have a direct impact on the model quality (Hatfield et al., 2008; Schwalbert et al., 2018) but, despite that, the majority of the published scientific literature randomly selected a subset of the data for comprising training or validation data (Assefa et al., 2016; Gholap, Ingole, Gohil, Gargade, & Attar, 2012; Gonzalez-Sanchez, 2014; Peralta et al., 2016; Sheridan, 2013) without following any guideline or statistical procedure.

Moreover, the choice of the statistical model employed to forecast yield has a large impact on the final result (Anselin, Bongiovanni, & Lowenberg-DeBoer, 2004; Peralta et al., 2016). Mostly empirical yield forecasting models based on VIs utilise classical ordinary least squares (OLS)-based on simple or multiple regression techniques (Noureldin, Aboelghar, Saudy, & Ali, 2013; Rembold, Atzberger, Savin, & Rojas, 2013; Shanahan et al., 2001), without properly accounting for the spatial autocorrelation structure amongst these variables (Imran, Zurita-Milla, & Stein, 2013; Peralta et al., 2016). The latter situation can lead to problems with inflated variance and likely resulting in wrong conclusions (Anselin et al., 2004; Bongiovanni, Robledo, & Lambert, 2007).

Models derived from simple empirical relationships usually tend to be time- and space-limited, valid only under similar conditions as when the correlation was established (Hatfield et al., 2008; Lobell, 2013; Tucker, 1979). Currently, the potential to forecast yield using satellite information through empirical models is already known, but the challenge is to extend these tools beyond the environment where the study was done (Hatfield et al., 2008). Lastly, the selection of adequate VIs is also an important step for model development (Peralta et al., 2016). The normalised difference vegetation index (NDVI) (Rouse, Haas, & Schell, 1973) is one the most widely used VIs to assess crop growth and yield (Peralta et al., 2016; Raun, Solie, & Johnson, 2002; Rembold et al., 2013; Solie, Dean Monroe, Raun, & Stone, 2012), and it becomes as a benchmark for researchers developing new VIs (Hatfield et al., 2008). However, there are some constraints related to saturation in medium to high leaf area index (LAI) values with NDVI (Haboudane, Miller, Pattey, Zarco-Tejada, & Strachan, 2004; Nguy-Robertson et al., 2012; Tucker, 1979). Thus, the incorporation of other VIs that still have sensitivity in high LAI values such as green NDVI (NDVIG) (Gitelson, Kaufman, & Merzlyak, 1996) and red-edge NDVI (NDVIre) (Gitelson & Merzlyak, 1994) have been reported improving empirical models (Hatfield et al., 2008; Peralta et al., 2016).

Following this rationale, guidelines for implementing yield forecasting models derived from empirical relationships and for validating their spatio-temporal relevancy still remain unknown. Thus, the objectives of this study were to: i) identify parameters to guide data selection aiming at building yield forecasting models using Sentinel-2 satellite imagery; ii) compare different approaches (OLS vs. spatial correlation) and different VIs during the model building process; iii) perform spatial and temporal model validation using independent datasets to identify potential limitations in up-scaling yield forecasting models. The main hypothesis is that model predictability power increases as the yield frequency distribution of the training data becomes more similar to the validation data even when considering diverse spatio-temporal scales (geography, time, or years).