Europ. J. Agronomy 77 (2016) 111–121

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Geostatistical interpolation and aggregation of crop growth model outputs

Luc Steinbuch [a,b,*], Dick J. Brus [c], Lenny G.J. van Bussel [d], Gerard B.M. Heuvelink [a,b]

[a] Soil Geography and Landscape Group, Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands
[b] ISRIC—World Soil Information, P.O. Box 353, 6700 AJ Wageningen, The Netherlands
[c] Team Soil, Water and Land Use, Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands
[d] Plant Production Systems, Wageningen University, P.O. Box 430, 6700 AK Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 14 January 2016
Received in revised form 12 March 2016
Accepted 20 March 2016
Available online 28 April 2016

Keywords:
Geostatistics
Spatial aggregation
Spatial prediction
Uncertainty
Yield potential
Yield gap

ABSTRACT

Many crop growth models require daily meteorological data. Consequently, model simulations can be obtained only at a limited number of locations, i.e. at weather stations with long-term records of daily data. To estimate the potential crop production at country level, we present in this study a geostatistical approach for spatial interpolation and aggregation of crop growth model outputs. As case study, we interpolated, simulated and aggregated crop growth model outputs of sorghum and millet in West-Africa. We used crop growth model outputs to calibrate a linear regression model using environmental covariates as predictors. The spatial regression residuals were investigated for spatial correlation. The linear regression model and the spatial correlation of residuals together were used to predict theoretical crop yield at all locations using kriging with external drift. A spatial standard deviation comes along with this prediction, indicating the uncertainty of the prediction. In combination with land use data and country borders, we summed the crop yield predictions to determine an area total. With spatial stochastic simulation, we estimated the uncertainty of that total production potential as well as the spatial cumulative distribution function. We compared our results with the prevailing agro-ecological Climate Zones approach used for spatial aggregation. Linear regression could explain up to 70% of the spatial variation of the yield. In three out of four cases the regression residuals showed spatial correlation. The potential crop production per country according to the Climate Zones approach was in all countries and cases except one within the 95% prediction interval as obtained after yield aggregation. We concluded that the geostatistical approach can estimate a country's crop production, including a quantification of uncertainty. In addition, we stress the importance of the use of geostatistics to create tools for crop modelling scientists to explore relationships between yields and spatial environmental variables and to assist policy makers with tangible results on yield gaps at multiple levels of spatial aggregation.

1. Introduction

To support research and decision making related to global food security, mechanistic crop growth models are frequently used to calculate the potential yield of a food crop in a certain area and context. These models describe the build-up of harvestable biomass as a result of the interaction between plant physiology and environment (Roudier et al., 2011; van Ittersum et al., 2013). Many of these models require accurate daily meteorological data, preferably observations, instead of interpolated grid based data, due to the non-linearity of many weather–crop relationships (van Bussel et al., 2011; van Wart et al., 2013a). In addition, detailed and locally relevant information about crop management and soil information are required for accurate crop growth simulations (van Ittersum et al., 2013). Consequently, model simulations can only be obtained on a limited number of locations, i.e. close to weather stations with long-term records.

In several studies average crop estimates for large areas have been obtained by spatially aggregating location-specific crop model simulations, see e.g. Rosenzweig and Parry (1994), Wolf and Diepen

Abbreviations: CZ, agro-ecological climate zones; GYGA, global yield gap atlas; KED, kriging with external drift; LOOCV, leave one out cross validation; REML, restricted maximum likelihood estimation; RVH, regressor variable hull; RWS, reference weather station, reference weather station location; SCDF, spatial cumulative distribution function; sd, standard deviation; se, standard error; SPAM, spatial plant allocation model; WOFOST, world food studies; Yp, yield potential; Yw, water-limited yield potential.

* Corresponding author.
E-mail addresses: luc.steinbuch@wur.nl, lst@dds.nl (L. Steinbuch).

http://dx.doi.org/10.1016/j.eja.2016.03.007
1161-0301/

(1995) and Alexandrov et al. (2002). A recently implemented approach is based on so-called agro-ecological Climate Zones (CZ) (van Wart et al., 2013b), applied e.g. in the Global Yield Gap Atlas (GYGA; www.yieldgap.org) (van Bussel et al., 2015). In this approach it is assumed that CZ are regions that are homogeneous with respect to climate conditions. The CZ approach is a straightforward and clear example of the calculate > interpolate > aggregate class of spatial aggregation approaches. An important drawback of this approach is that it ignores spatial variation of crop growth simulations within the climate zones, i.e. within these zones the simulated crop growth is assumed constant. Incorporating spatial variation could improve the spatial resolution and accuracy of the final results and thus help supporting national and local policy decisions, prioritizing investment strategies of fertilizer and seed companies and NGO's. The CZ approach also fails to quantify the uncertainties associated with the interpolation and aggregation steps, which is essential information to guide accuracy improvement strategies (van Bussel et al., 2016). In this study we therefore explore whether the drawbacks of the CZ approach can be overcome with the help of a geostatistical approach. Geostatistics provides tools for a coherent quantification of site-specific as well as aggregated modelled crop yield prediction uncertainties. It produces continuous spatial maps that provide valuable location-specific information for crop modellers as well as decision makers and yields graphs that indicate areal proportions below or above a potential yield level threshold for regions or countries. It also offers means to explore the relationships between calculated yields and explanatory environmental variables.

The aims of this study are to present a state-of-the art model-based geostatistical method for spatial interpolation and aggregation of simulated yields, to illustrate it with a case study, and to compare the results with those of the common CZ approach. More specifically, we use kriging with external drift (KED), supported by restricted maximum likelihood parameter estimation (REML; Lark 2000; Diggle and Ribeiro, 2007). Additionally, we use spatial stochastic simulation to predict aggregated crop production at country level and its associated uncertainty. As a case study, we interpolate and aggregate modelled yields of sorghum (*Sorghum bicolor*) and millet (e.g. *Pennisetum glaucum*, *Eleusine coracana*) in West Africa, as provided by the crop growth model WOFOST (Wolf et al., 2011; Supit et al., 2012).

2. Materials and methods

2.1. Study area

This study has been carried out in West Africa, focussing on Burkina Faso, Mali, Ghana, Niger and Nigeria. Most of this area consists of a low plateau of maximal 500 m above sea level, with some mountainous areas up to 2040 m. The daily mean temperature is almost always and everywhere (except at high altitudes) above 18 °C and relatively stable during the year. The most dynamic weather pattern is precipitation, dictated by dry winds from the Sahara in the north, dominant from November until February, and by the moist southwest marine wind, dominant in July (von Kaufmann et al., 1983).

2.2. Modelled crop yield data

Modelled crop yields for sorghum at 38 (Fig. 1) and millet at 37 Reference Weather Station locations (RWS) were obtained from the Global Yield Gap Atlas (www.yieldgap.org). Two yield levels, yield potential (Yp) and water-limited yield potential (Yw), were simulated using the crop growth model WOFOST version 7.1.3 (release March 2011) (Wolf et al., 2011; Supit et al., 2012). The yield potential is determined by solar radiation, temperature and carbon dioxide concentration; there are no limitations due to water stress, low soil fertility, weeds, pests, etc. The yield potential is further influenced by management practices like sowing date and cultivar choice. The water-limited yield potential, i.e. rainfed yield, is defined similar as Yp, except that possible water stress is taken into account (Evans 1996; van Ittersum and Rabbinge, 1997).

Table 1
Summary statistics of crop growth model yields Yp and Yw, for sorghum and millet. 'n' is the number of observations, i.e. the number of modelled crop yields per case. 'Skewness' refers to the asymmetry of the dataset values.

	Units	Research cases			
		Sorghum Yp	Sorghum Yw	Millet Yp	Millet Yw
n	[–]	38	38	37	37
Mean	[t/ha]	7.50	6.21	4.23	3.02
Standard deviation	[t/ha]	1.17	2.00	1.02	1.40
Skewness	[–]	−0.15	−0.45	−0.66	0.09
Min	[t/ha]	5.04	2.04	1.21	0.56
Max	[t/ha]	9.96	9.71	5.90	5.80

The 38 and 37 locations used in the crop yield modelling were selected on: (1) the basis of proximity of weather stations with high-quality weather data, which are located in areas with high crop densities as indicated by You et al. (2006) and You et al. (2009); see also http://mapspam.info) and (2) the dominant representation of the crop growing conditions in terms of weather, soils and cropping system for the countries of interest. Sorghum and millet share the same location 32 times. The final numbers of Yw on each location are area-weighted means of several simulations for dominant soil types; both Yp and Yw are averaged over multiple years of simulation (Grassini et al., 2015; van Bussel et al., 2015). Summary statistics of simulated Yp and Yw for sorghum and millet are provided in Table 1.

2.3. Trend model covariates

In the kriging procedure described hereafter, we used grid maps of environmental and meteorological variables that are expected to be related to the simulated crop yield. To stay as close as possible to the agro-ecological Climate Zones method, we decided to use for all four cases a trend model with the three covariates used in the CZ approach (Table 2).

2.4. Geostatistical modelling

2.4.1. General framework: the geostatistical model

To build a geostatistical model, we first need to introduce the idea of a random field. A random field is a set of random variables indexed by location (Plant, 2012). Additional to the statistical model of a random variable, a random field has parameters describing its spatial correlation.

In this paper, we build a statistical model of a random field for each of the four cases defined in Section 2.2. Thus, the crop growth model outputs for each case are considered realisations of four separate random fields. Our general statistical model of each random field is denoted by $Z = \left\{Z(s), s \in D\right\}$ (unit: t/ha; s is a two-dimensional vector, representing geographic location, D is the geographic domain of interest). At each location $s \in D, Z(s)$ is modelled as the sum of a spatial trend (a linear regression part) and a stochastic residual (a random variable):

$$Z(s) = \beta_0 + \sum_{i=1}^{p} \beta_i \times x_i(s) + \varepsilon(s) = X(s)^T \times \beta + \varepsilon(s) \qquad (1)$$

where β_0 is the regression intercept, $\beta_i\,(i = 1 \cdots p)$ are regression coefficients associated with the covariates, $x_i(s)$ is the i^{th} environ-