Environmental Modelling & Software 80 (2016) 100–112

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Evaluating the precision of eight spatial sampling schemes in estimating regional means of simulated yield for two crops

Gang Zhao [a, *], Holger Hoffmann [a], Jagadeesh Yeluripati [b, e], Specka Xenia [c], Claas Nendel [d], Elsa Coucheney [e], Matthias Kuhnert [f], Fulu Tao [g], Julie Constantin [h], Helene Raynal [h], Edmar Teixeira [i], Balázs Grosz [j], Luca Doro [k], Ralf Kiese [l], Henrik Eckersten [m], Edwin Haas [l], Davide Cammarano [n], Belay Kassie [n], Marco Moriondo [o], Giacomo Trombi [p], Marco Bindi [p], Christian Biernath [q], Florian Heinlein [q], Christian Klein [q], Eckart Priesack [q], Elisabet Lewan [e], Kurt-Christian Kersebaum [d], Reimund Rötter [g], Pier Paolo Roggero [k], Daniel Wallach [h], Senthold Asseng [n], Stefan Siebert [a], Thomas Gaiser [a], Frank Ewert [a]

[a] *Crop Science Group, Institute of Crop Science and Resource Conservation (INRES), University of Bonn, Katzenburgweg 5, 53115, Bonn, DE, Germany*
[b] *The James Hutton Institute, Craigiebuckler, Aberdeen, Scotland, AB15 8QH, UK*
[c] *Institute of Landscape Biogeochemistry, Leibniz Centre for Agricultural Landscape Research, 15374, Müncheberg, DE, Germany*
[d] *Institute of Landscape Systems Analysis, Leibniz Centre for Agricultural Landscape Research, 15374, Müncheberg, DE, Germany*
[e] *Department of Soil and Environment, Swedish University of Agricultural Sciences, Lennart Hjelms väg 9, 750 07, Uppsala, Sweden*
[f] *Institute of Biological and Environmental Sciences, School of Biological Sciences, University of Aberdeen, 23 St Machar Drive, Aberdeen, AB24 3 UU, Scotland, UK*
[g] *Environmental Impacts Group, Natural Resources Institute Finland (Luke), Jokiniemenkuja1, FI-01370, Vantaa, Finland*
[h] *INRA, UMR 1248 AGIR & UR0875 MIA-T, F-31326, Auzeville, France*
[i] *Systems Modelling Team (Sustainable Production Group), The New Zealand Institute for Plant and Food Research Limited, Canterbury Agriculture & Science Centre, Gerald St, Lincoln, 7608, New Zealand*
[j] *Thünen-Institute of Climate-Smart-Agriculture, Bundesallee 50, 38116, Braunschweig, DE, Germany*
[k] *Nucleo di Ricerca sulla Desertificazione (NRD) and Dipartimento di Agraria, Università degli Studi di Sassari, Viale Italia 39, 07100, Sassari, Italy*
[l] *Institute of Meteorology and Climate Research, Atmospheric Environmental Research, Karlsruhe Institute of Technology, Kreuzeckbahnstraße 19, 82467, Garmisch-Partenkirchen, DE, Germany*
[m] *Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Ulls väg 16, 750 07, Uppsala, Sweden*
[n] *Agricultural & Biological Engineering Department, University of Florida, Frazier Rogers Hall, Gainesville, FL, 32611, USA*
[o] *CNR-Ibimet, Via Madonna del Piano 10, 50019, Sesto Fiorentino, Italy*
[p] *Department of Agri-food Production and Environmental Sciences, University of Florence, Piazzale delle Cascine 18, 50144, Firenze, Italy*
[q] *Institute of Biochemical Plant Pathology, German Research Center for Environmental Health, Ingolstädter Landstraße 1, D 85764, Neuherberg, DE, Germany*

ARTICLE INFO

Article history:
Received 7 August 2015
Received in revised form
3 January 2016
Accepted 15 February 2016
Available online xxx

Keywords:
Crop model
Stratified random sampling
Simple random sampling
Clustering
Up-scaling
Model comparison
Precision gain

ABSTRACT

We compared the precision of simple random sampling (SimRS) and seven types of stratified random sampling (StrRS) schemes in estimating regional mean of water-limited yields for two crops (winter wheat and silage maize) that were simulated by fourteen crop models. We found that the precision gains of StrRS varied considerably across stratification methods and crop models. Precision gains for compact geographical stratification were positive, stable and consistent across crop models. Stratification with soil water holding capacity had very high precision gains for twelve models, but resulted in negative gains for two models. Increasing the sample size monotonously decreased the sampling errors for all the sampling schemes. We conclude that compact geographical stratification can modestly but consistently improve the precision in estimating regional mean yields. Using the most influential environmental variable for stratification can notably improve the sampling precision, especially when the sensitivity behavior of a crop model is known.

* Corresponding author.
E-mail address: gzhao@uni-bonn.de (G. Zhao).

http://dx.doi.org/10.1016/j.envsoft.2016.02.022
1364-8152/

1. Introduction

Dynamic crop models are developed for simulating crop growth and yield in response to various environmental conditions and management practices at a field scale (Keating et al., 2003; van Diepen et al., 1989; Williams et al., 1989). To provide summarized information (e.g. mean/total crop production inside a political boundary) for agricultural impact and risk assessment to support policy making, crop models need to be applied over large areas. Due to data paucity and computing cost, simulations are typically conducted at a limited number of sample locations across a region, through which results are up-scaled to regional or larger scales (Ewert et al., 2011). For example, Rötter et al. (1995) chose 18 sites to represent a large watershed, the Rhine basin. Trnka et al. (2014) chose 14 sites to represent Europe to simulate the adverse weather events for wheat. Asseng et al. (2015) chose 30 sites across the world to simulate temperature effects on global wheat production. The methods used to select simulation locations, called sampling design, can be used to improve the representativeness of the simulation results (Roleček et al., 2007).

Many environmental characteristics show a spatial continuity, i.e. data at two nearby locations are on average more similar than data at two widely spaced locations. For this reason, when using environmental data as input to a crop model, the simulation results are spatially dependent (Caeiro et al., 2003). Despite this, classical sampling theory is perfectly valid for such spatially structured populations (Brus and De Gruijter, 1997; Brus and DeGruijter, 1993; De Gruijter and Ter Braak, 1990). Model-based and design-based are two widely used schemes of sampling (Cassel et al., 1977; Wang et al., 2013). For estimating global and regional means, design-based strategies can be advantageous (Brus and De Gruijter, 1993), while simple random sampling (SimRS) and stratified random sampling (StrRS) are two of the most important design-based strategies (Hirzel and Guisan, 2002; Ripley, 2005). In SimRS, a given number of sampling units are selected independently from each other and with equal inclusion probability (Cochran, 1977). In StrRS, the entire study area is separated into sub-regions, called strata (or zoning), frequently according to prior information on the population and then random sampling is applied to each stratum. These two design-based schemes have been widely evaluated in monitoring of natural resources (Brus, 1994; De Gruijter et al., 2006), species distribution modeling (Stockwell and Peterson, 2002; Wisz et al., 2008) and demographic health surveys (Kumar, 2007, 2009). In a vegetation survey, Austin and Heyligers (1989) found that stratifying the population by combined information on climate, topographic and lithological characteristics could better represent the environmental variability in the area, especially when the stratification is coupled with well-tuned sampling rules based on aspect and topographic position. Wang et al. (2002) found that zoning of the population based on prior knowledge of the influential variables could reduce the sample size to achieve the same efficiency in monitoring the area of cultivated land. Brus (1994) found that the estimation accuracy can be improved by stratifying the population based on soil and land use maps when estimating the spatial means of phosphate sorption characteristics. Wang et al. (2010) found that stratification of population in the study area could reduce the variance of estimators in surveys of non-cultivated land in China.

In these survey and monitoring applications, the prior information that is used to stratify the population is normally obtained from other correlated variables or historical survey data. In crop modeling, the output population is simulated with the input of environmental variables and management practices, which can be used as prior information to create the strata. Many types of strata or zones including climate zones (Rötter et al., 2012), agro-climatic zones (Rötter et al., 1995), environmental zones (Metzger et al., 2005; Olesen et al., 2011), agro-ecological zones (Aggarwal, 1993), and climate-soil zones (Bryan et al., 2014; Zhao et al., 2015a), have been used for regional or global crop modelling studies. However, only very few studies explicitly investigated the precision of these stratification methods and spatial sampling strategies. Nendel et al. (2013) showed that one soil profile and weather station were not sufficient to represent the observed mean grain yields of winter wheat in Thuringia, a region in Germany covering more than 16 000 km^2. By using one soil profile and gridded weather data at 1 km spatial resolution, van Bussel et al. (2016) evaluated the effects of sample size of StrRS on simulations of winter wheat yields under two production conditions, i.e. potential and water-limited in North Rhine-Westphalia. They recommended that detailed soil properties should be included in the simulations to further consolidate the conclusions from their study. To our best knowledge, no study has compared the efficiency of different stratum types (i.e. variables used to create the strata) and stratum number for estimating regional mean of simulated crop yields.

This study aims to compare the precision of SimRS and seven types of StrRS in estimating regional mean yield for two crops (winter wheat and silage maize). We investigated how the precision, indicated by mean squared error (*MSE*), depends on the sample sizes, the variables used for stratification, the number of strata, the crop types and the crop models.

2. Methods

2.1. Sampling precision

Crop yields of a region (A) constitute a continuous surface that can be infinitely divided. However, due to computing cost and input data availability, it was not possible even to do the simulations for each individual field of the entire study area. Instead, we divided the A into 1 × 1 km grid cells and simulated yield for each cell. The results were treated as the full population (N = 34,168) and the average over all cells was treated as the true regional yield $\overline{Y}(A)$. We sampled the population with a range of sample sizes and sampling schemes to giving various estimates $\widehat{\overline{Y}}(A)$ of $\overline{Y}(A)$.

Eight design-based sampling schemes were evaluated, including simple random sampling (SimRS) and seven stratified random sampling (StrRS) with strata based on different environmental variables. A stratification method with L strata divides the population of grid cells into L non-overlapping groups. SimRS can be treated as a one stratum StrRS ($L = 1$). For any particular stratification method, let N_h denote the number of cells within stratum h. This is determined by the stratification scheme, and is known. Suppose from each stratum a simple random sample without replacement is selected. The sample size within stratum h is noted n_h. The symbols used in this study are shown in Table 1.

The estimated mean using stratified random sampling $\widehat{\overline{Y}}(A)$ was calculated as

$$\widehat{\overline{Y}}(A) = \frac{\sum_{h=1}^{L} N_h \widehat{\overline{Y}}_h}{N} = \sum_{h=1}^{L} w_h \widehat{\overline{Y}}_h \tag{1}$$

where $\widehat{\overline{Y}}_h$ is average yield in stratum h, estimated using the samples from that stratum.

The estimator $\widehat{\overline{Y}}_h$ is unbiased, since the mean of all possible samples equals to the *true* population mean of stratum h. Therefore, $\widehat{\overline{Y}}(A)$ is also an unbiased estimator of the population mean $\overline{Y}(A)$ of the entire region according to Theorem 5.1 in Cochran (1977). To