Computers and Electronics in Agriculture 123 (2016) 74–79

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Spatial interpolation quality assessment for soil sensor transect datasets

Elia Scudiero [a,*], Dennis L. Corwin [a], Francesco Morari [b], Ray G. Anderson [a], Todd H. Skaggs [a]

[a] USDA-ARS, United States Salinity Laboratory, 450 West Big Springs Rd., Riverside, CA 92507-4617, USA[1]

[b] Department of Agronomy, Food, Natural Resources, Animals and Environment (DAFNAE), University of Padua, Viale dell'Università 16, Legnaro 35020, Italy

ARTICLE INFO

Article history:
Received 24 July 2015
Received in revised form 20 February 2016
Accepted 22 February 2016
Available online 5 March 2016

Keywords:
Apparent electrical conductivity
Spatial interpolation
NDVI
On-the-go sensors
R
h-block resampling

ABSTRACT

Near-ground geophysical soil sensors provide valuable information for precision agriculture applications. Indeed, their readings can be used as proxy for many soil parameters. On-the-go soil sensor surveys are, typically, carried out intensively (e.g., every 2 m) over many parallel transects. Two types of soil sensors measurements are considered in this paper: apparent electrical conductivity (4 fields in California, USA) and reflectance (1 field in Italy). Two types of spatial interpolations are carried out, universal kriging (model-based) and inverse distance weighting (deterministic). Interpolation quality assessment is usually carried out using leave-one-out (*loo*) resampling. We show that *loo* resampling on transect sampling datasets returns overly-optimistic, low interpolation errors, because the left-out data point has values very close to that of its neighbors in the training dataset. This bias in the map quality assessment can be reduced by removing the closest neighbors of the validation observation from the training dataset, in a (spatial) *h-block* (SHB) fashion. The results indicate that, for soil sensor data acquired along parallel transects: (*i*) the SHB resampling is a useful tool to test the performance of interpolation techniques and (*ii*) the optimal (i.e., rendering the same errors of un-sampled locations between transects) SHB threshold distance (*h.dist*) for neighbor-exclusion is proportional to the semi-variogram range and partial sill. This procedure provides research scientists with an improved means of understanding the error of soil maps made by interpolating soil sensor measurements.

Published by Elsevier B.V.

1. Introduction

The benefits of using on-the-go sensors as proxies for soil properties is well recognized (Adamchuk et al., 2004). The increased coverage provided by geospatial sensor measurements enables the spatial structure of the target soil property to be characterized more accurately than when a limited set of soil samples are used (Corwin and Lesch, 2005a). In order to obtain information across the entire field, spatial interpolation techniques (e.g., kriging, inverse distance weighing) are employed. Once the map is made, it is essential to properly quantify its prediction uncertainty. Indeed, interpolation error is often the greatest contribution to the overall prediction error in a soil map (Nelson et al., 2011).

To assess the quality of the spatial interpolations, leave-one-out (*loo*) resampling techniques are usually employed (Robinson and Metternicht, 2006). The *loo* resampling is particularly effective when removing a single observation allows estimating the interpolation error over the farthest-away-as-possible (in terms of distance and/or value) location from the observed data (i.e., where the highest prediction uncertainties are expected). Unfortunately this does not always apply to soil sensor data. On-the-go soil sensors are, generally, used to acquire data intensively (e.g., every 2 m), along many parallel transects. Unless the transect spacing is narrow enough for the sampling scheme to be considered a disperse grid, transect sampling is clustered (i.e., large difference between average nearest neighbor and transect spacing). In clustered sampling, neighboring measurements tend to be very similar. Therefore, removing a single location may not provide comprehen-

* Corresponding author. Tel.: +1 951 369 4864.
E-mail addresses: elia.scudiero@ars.usda.gov, scudiero@dmsa.unipd.it (E. Scudiero).

[1] Note: The U.S. Department of Agriculture (USDA) prohibits discrimination in all its programs and activities on the basis of race, color, national origin, age, disability, and where applicable, sex, marital status, familial status, parental status, religion, sexual orientation, genetic information, political beliefs, reprisal, or because all or part of an individual's income is derived from any public assistance program. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (Braille, large print, audiotape, etc.) should contact USDA's TARGET Center at (202) 720-2600 (voice and TDD). To file a complaint of discrimination, write to USDA, Director, Office of Civil Rights, 1400 Independence Avenue, S.W., Washington, D.C. 20250-9410, or call (800) 795-3272 (voice) or (202) 720-6382 (TDD). USDA is an equal opportunity provider and employer.

http://dx.doi.org/10.1016/j.compag.2016.02.016
0168-1699/Published by Elsevier B.V.

sive information on the error at un-sampled locations. Unrealistically low error estimates may, then, be expected (Ruß and Brenning, 2010; Brenning, 2012).

To overcome this issue in the interpolation quality assessment, nearly identical neighbors of the validation observation can be removed from the training dataset. This particular case of *loo* is called *h-block* (HB) resampling (Burman et al., 1994; Telford and Birks, 2009). Spatial HB (SHB) resampling is generally used to remove the spatial bias in the evaluation of the performance of different spatial regression models. The SHB is generally employed in large-scale (e.g., hundreds of km) applications, to select/validate spatial models, mostly, in ecology studies. Here, we propose a version of the SHB resampling for spatial interpolation quality assessment at the field-scale (e.g., hundreds of m). The proposed application specifically targets the interpolation of intense transect surveys carried out with on-the-go proximal soil sensors.

2. Materials and methods

2.1. Sensor data

On-the-go soil sensing was carried out by electromagnetic induction (EMI) over four fields in California, USA and with an active radiometer over one field in northeastern Italy (Supplemental Fig. A.1).

Transect EMI surveys were used to measure apparent electrical conductivity (EC_a), following the EC_a survey protocols of Corwin and Lesch (2005b), over the 0–1.50 m soil depth in four agricultural fields in California, USA (Supplemental Fig. A.1 and Table A.1): Fields 1, 2, 3 in the western San Joaquin Valley, and Field 4 in the San Jacinto Valley. Data for Fields 1, 2, and 4 were taken from Scudiero et al. (2014). Data for Field 4 was taken from Corwin et al. (2010). Measurements were carried out using an EM38 (Geonics Ltd., Mississauga, Ontario, Canada[2]) sensor, connected to a Trimble (Sunnyvale, CA, USA[2]) GPS system with decimetric precision in horizontal positioning and mounted on a nonmetallic sled (as shown in Fig. 5 of Corwin and Lesch, 2005b). Field 1 (20.7 ha) was surveyed with 111 transects, on average ~6 m apart, totaling 13,440 EC_a readings (Fig. 1). Field 2 (6.4 ha) was surveyed with 8 transects, on average ~9 m apart, totaling 1311 EC_a readings (Supplemental Fig. A.2). Field 3 (40.5 ha) was surveyed with 18 transects, on average ~32 m apart, totaling 1204 EC_a readings (Supplemental Fig. A.3). Field 4 (6.9 ha) was surveyed with 44 transects, on average ~4 m apart, totaling 3502 EC_a readings (Supplemental Fig. A.4).

For Field 5, on-the-go bare-soil reflectance at 590 ± 5.5 nm (VIS) and at 880 ± 5.5 nm (NIR) was measured with an active spectrometer (ACS-210-CropCircle, Holland Scientific, Lincoln, NE, USA) linked with a Trimble (Sunnyvale, CA, USA[2]) GPS system with decimetric precision in horizontal positioning over a 25.8-ha field in Chioggia, Italy (Supplemental Fig. A.1 and Table A.1). The NIR and VIS readings were used to calculate the normalized difference vegetation index (NDVI) (Rouse et al., 1973):

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{VIS}}{\mathrm{VIS} + \mathrm{NIR}} \tag{1}$$

The survey at Field 5 was carried out over 22 transects, on average ~27 m apart, totaling 7403 NDVI readings (Supplemental Fig. A.5). Data for Field 5 was taken from Scudiero et al. (2013).

[2] Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

2.2. Spatial interpolations specifications

In this paper, we discuss the quality assessment of model-based (i.e., universal kriging) and deterministic (i.e., inverse distance weighting) spatial interpolation techniques.

2.2.1. Kriging

At all fields, EC_a and bare-soil NDVI data were characterized by the presence of spatial trend and were interpolated using Universal Kriging (UK). Data for Field 1 and 3 were normalized using square-root transformation and Field 4 was normalized with natural logarithm transformation. To carry out the interpolation, the spatial correlation structures of EC_a and of bare-soil NDVI were modeled by an isotropic penta-spherical semi-variogram, $v(EC_a)$:

$$v(\delta_i) = \begin{cases} (\eta + \sigma) \times \left[\frac{15}{8}\frac{h}{r} - \frac{5}{4}\left(\frac{h}{r}\right)^3 + \frac{3}{8}\left(\frac{h}{r}\right)^5\right] & \text{for } h \leqslant r \\ (\eta + \sigma) & \text{for } h > r \end{cases} \tag{2}$$

where η represents the nugget variance, σ the spatial variance component (partial sill), h the lag distance, and r the range. Semi-variograms were considered accurate when the *loo* resampling average kriging standard error (i.e., the squared-root average of the kriging variance at all locations) was very close to the RMSE (Robinson and Metternicht, 2006). Semi-variogram specifications are reported in Table 1 (for Field 1) and Supplemental Table A.3 (for the other fields). Kriging interpolations were performed using a maximum of 40 neighbors.

2.2.2. Inverse distance weighting

Inverse distance weighting (IDW) estimates values at un-sampled locations as weighted average of the known data points within a selected number of neighbors of the un-sampled location:

$$x_0 = \frac{\sum_{i=1}^{n} x_i \times d_i^{-w}}{\sum_{i=1}^{n} d_i^{-w}} \tag{3}$$

where x_0 is the value to be estimated, x_i is the know value at location i within the neighborhood of n known points (i.e., $n = 40$), d is the distance of x_0 to x_i, and w (>0) is the IDW weighting exponent. The lower w, the more uniformly the n neighbors are incorporated into the calculation of x_0. Contrarily, with high weighting exponent values, the estimation of x_0 is mainly determined by the closest x_i values (Robinson and Metternicht, 2006). The *Model Optimization* feature in Arc Map's (version 10.1; ESRI, Redlands, CA, USA) *Geostatistical Analyst* was used to determine the best w by minimizing the *loo* resampling residual sum of squares.

2.3. Interpolation quality assessment: spatial h-block (SHB) resampling

In the SHB, each observation is removed from the dataset and used for validation. Then, according to an arbitrary threshold neighborhood size, neighboring locations to the validation observation are removed. The threshold neighborhood is, in this manuscript, a circular area of radius of size *h.dist*. The remaining observations (i.e., training dataset) are used to interpolate the selected variable at the validation location. The interpolated prediction is then compared to the observed (left-out) value. Similar to the classical *loo* resampling, the above described procedure is repeated for every observation of the dataset. Finally, the size of the error of the SHB predictions from the actual observed data is used as the metric to evaluate the quality of the spatial interpolation model (i.e., interpolation prediction errors). In this work, we analyze the resampling root mean square error (RMSE) of spatial interpolations.

The SHB procedure for UK and IDW was carried out in the R (version 3.2.0, R Core Team, 2015) environment. For each valida-