International Journal of Applied Earth Observation and Geoinformation 22 (2013) 127–138

Contents lists available at SciVerse ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

Mapping efficiency and information content

T. Hengl[a,*], M. Nikolić[b], R.A. MacMillan[a]

[a] *ISRIC – World Soil Information, P.O. Box 363, 6700 AJ Wageningen, The Netherlands*
[b] *Faculty of Mathematics, University of Belgrade, Studentski trg 16, Belgrade, Serbia*

ARTICLE INFO

Article history:
Received 14 July 2011
Accepted 17 February 2012

Keywords:
Scale
Effective information content
Complexity
Compression
Regression-kriging
Soil mapping

ABSTRACT

This paper proposes two compound measures of mapping quality to support objective comparison of spatial prediction techniques for geostatistical mapping: (1) *mapping efficiency* – defined as the costs per area per amount of variation explained by the model, and (2) *information production efficiency* – defined as the cost per byte of effective information produced. These were inspired by concepts of complexity from mathematics and physics. Complexity i.e. the total effective information is defined as bytes remaining after compression and after rounding up the numbers using half the mapping accuracy (effective precision). It is postulated that the mapping efficiency, for an area of given size and limited budget, is basically a function of inspection intensity and mapping accuracy. Both measures are illustrated using the Meuse and Ebergötzen case studies (gstat, plotKML packages). The results demonstrate that, for mapping organic matter (Meuse data set), there is a gain in the mapping efficiency when using regression-kriging versus ordinary kriging: mapping efficiency is 7% better and the information production efficiency about 25% better (3.99 vs 3.14 EUR B^{-1} for the GZIP compression algorithm). For mapping sand content (Ebergötzen data set), the mapping efficiency for both ordinary kriging and regression-kriging is about the same; the information production efficiency is 29% better for regression-kriging (37.1 vs 27.7 EUR B^{-1} for the GZIP compression algorithm). Information production efficiency is possibly a more robust measure of mapping quality than mapping efficiency because: (1) it is scale-independent, (2) it can be more easily related to the concept of effective information content, and (3) it accounts for the extrapolation effects. The limitation of deriving the information production efficiency is that both reliable estimate of the model uncertainty and the mapping accuracy is required.

1. Introduction

Every time we produce maps, GIS and/or geographical databases, the methods used to generate those products can be evaluated for their efficiency. Unfortunately, most evaluation of maps produced by land resource inventories in the world is still done on a *'look-good'* assessment and the inherent uncertainty of the product is often underreported (Lunetta and Lyon, 2004). Mapping efficiency in, for example, conventional soil survey is often expressed only as costs per area (e.g. 2 US$ per km^2), while the actual quality of the produced maps is often neglected (Eldridge, 1997; Legros, 2004). The main reason for this is that explicit measures of mapping quality are lacking or are not widely accepted by soil survey agencies (Finke, 2006; Brus et al., 2011).

There is now an increasing need to quantify economic aspects of producing maps in relation to (geo)information use (de Bruin and Hunter, 2003; Carrick et al., 2010) and answer questions such as: what are the costs of producing bits of information? what is the value of geoinformation for spatial planning and environmental modeling? how much does any given geostatistical mapping method costs (while producing equally accurate information)? how much does environmental remediation cost? what is the cost–benefit ratio between mapping and exploitation of land resources?

The purpose of this paper is to argue for adoption and use of objective measures of mapping efficiency in the context of geostatistical mapping i.e. spatial prediction of continuous environmental variables. We start by extending the information content concepts and then propose two new measures of mapping quality potentially interesting for both operational planning of surveys and statistical interpretation of results. We first introduce and explain some basic theoretical concepts – mapping scale, effective precision or numeric resolution, complexity and compression (effective information content) – and then illustrate how to calculate mapping efficiency and information production efficiency using real data.

Although we primarily refer to digital soil mapping examples, we assume that these concepts are applicable to other mapping fields such as vegetation and/or land cover mapping, species distribution mapping, climatic mapping and similar. R code used to

* Corresponding author. Tel.: +31 0317 484199.

0303-2434/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jag.2012.02.005

produce all numbers and graphs in this paper are available via the journal website.

2. Theory

2.1. Mapping scale and survey costs

Concepts of effective scale, effective pixel size, detection limit, effective precision, mapping efficiency and information content are often unclear to non-computer scientists and/or non-statisticians. We hence first focus on clarifying each of these in simple terms.

Mapping scale defines many of the aspects of a geostatistical mapping. Although we can provisionally set a nominal scale of a mapping project to any arbitrary number, the actual (effective) mapping scale is, in fact, determined by the actual inspection intensity i.e. density of field observations (Lam and Quattrochi, 1992; Eldridge, 1997; Stein et al., 2001). A cartographic rule used, for example, in traditional soil mapping is that there should be at least one, and ideally four, observations per 1 cm^2 of the map (Avery, 1987, see Table 1), hence for an area of size A that contains N field observations, the effective scale number (SN) is:

$$\mathrm{SN} = \sqrt{4 \cdot \frac{A}{N}} \cdot 10^2 \quad \ldots \quad \mathrm{SN} = \sqrt{\frac{A}{N}} \cdot 10^2 \tag{1}$$

where A is the surface of the study area in m^2 and N is the total number of observations (Fig. 1).

Inspection intensity can also be related to the level of detail in the remotely sensed imagery that typically have a block support. The corresponding pixel size can be related to scale by using the following simple rule (McBratney et al., 2003):

$$p = 0.0005 \cdot \mathrm{SN} \tag{2}$$

i.e. the pixel size should be at least 0.5 mm in the map scale. When working with high resolution remote sensing imagery, Topan et al. (2009) suggest that the ground sampling distance (GSD) should be at least 0.1 mm GSD in the map scale, hence the correct pixel size is probably somewhere between 0.5 and 0.1 mm in the map scale.

Based on Eqs. (1) and (2), and assuming an average of 2.5 observations per 1 cm^2, we can see that an objective estimate of *effective pixel size* can be determined by using (Hengl, 2006):

$$p = 0.0791 \cdot \sqrt{\frac{A}{N}} \tag{3}$$

This principle can also be expressed as: one point sample can be used to generate predictions at 160 pixels on a map, i.e. 100 point samples can be used to produce 16,000 pixels in a gridded map. In practice, most of soil surveys work with lower sampling intensities.

Legros (2004, p. 75) is somewhat less strict considering the number of field observations needed per ha:

$$\hat{\mathrm{SN}} = \exp(8.669545 + 0.652391 \cdot \log(\mathrm{ha}_a)) \tag{4}$$

where ha_a is the hectares per auger bore observations. This means that one could collect 10 auger observations for 100 ha of land to produce ca. 5800 pixels i.e. 580 pixels per point.

Mapping scale also directly determines the survey costs. Burrough et al. (1971), Bie and Ulph (1972), and Bie et al. (1973) postulated in early 70s that the survey costs are a direct function of the mapping scale:

$$\log \left\{ \begin{array}{c} \text{cost per km}^2 \\ \text{or} \\ \text{man} - \text{days per km}^2 \end{array} \right\} = a + b \cdot \log(\text{map scale}) \tag{5}$$

This model typically explains >75% of the survey costs (Burrough et al., 1971). Based on Eq. (5), we have fitted a linear model to the empirical table data from e.g. Legros (2004, p. 75), to produce the following model:

$$\hat{\mathrm{X}} = \exp(19.0825 - 1.6232 \cdot \log(\mathrm{SN})) \tag{6}$$

where X is the minimum cost/ha in Euros (based on estimates in 2002). To map 1 ha of soil at 1:200,000 scale (at the beginning of the 21st century), for example, one needs at least 0.48 Euros (i.e. 48 EUR to map a square-km); to map soil at 1:20k would costs about 25 EUR per ha. Legros (2004) indicates that these are the all-inclusive costs that include salaries and time in the office needed for the work of synthesis and editing.

In practice, estimated standard costs per area for conducting soil survey differ from country to country. The USDA estimates that the total costs of soil mapping at their most detailed scale (1:20k) costs about 1.50 US$ per acre i.e. about 4 US$ per ha (Durana,2008); in Canada, typical costs of producing soil maps at 1:20k are in the range 3–10 CA$ per ha (MacMillan et al., 2010); in the Netherlands 3.4 EUR per ha (Kempen, 2011, pp. 149–154); in New Zealand 4 US$ per ha (Carrick et al., 2010).

Based on these global estimates we can consider the following rule of thumb: to map 1 ha of land at 1:20k scale, one would need (at least) 5 US$. Therefore the model of Legros (2004) would need to be calibrated to [USD]:

$$\hat{\mathrm{X}}_c = .167 \cdot \exp(19.0825 - 1.6232 \cdot \log(\mathrm{SN})) \tag{7}$$

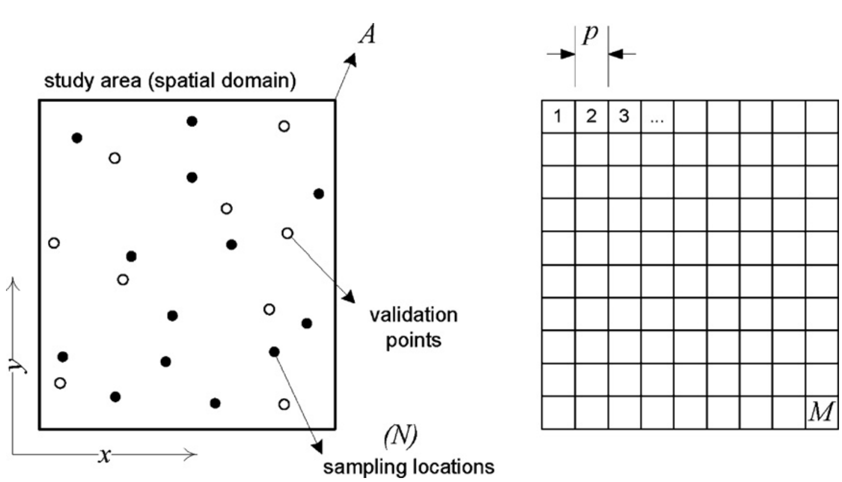

Fig. 1. Some basic concepts of geostatistical mapping: spatial domain (A – area), sampling locations (N – total number of points), validation points, prediction grid (M – total number of cells) and the grid cell size (p – pixel size).