Available online at www.sciencedirect.com
ScienceDirect
Geoderma 141 (2007) 1 – 14
GEODERMA
www.elsevier.com/locate/geoderma

Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping

F. Carré [a,*], Alex B. McBratney [b], B. Minasny [b]

[a] European Commission, DG Joint Research Centre, Institute of Environment and Sustainability, Land Management Unit, TP 280, 21020 Ispra (Va), Italy
[b] Australian Centre for Precision Agriculture, Faculty of Agriculture, Food & Natural Resources, The University of Sydney, NSW 2006, Australia

Received 19 January 2006; received in revised form 24 August 2006; accepted 24 January 2007
Available online 29 May 2007

Abstract

Legacy soil data form an important resource for digital soil mapping and are essential for calibration of models for predicting soil properties from environmental variables. Such data arise from traditional soil survey. Methods of soil survey are generally empirical and based on the mental development of the surveyor, correlating soil with underlying geology, landforms, vegetation and air-photo interpretation. There are no statistical criteria for traditional soil sampling, and this may lead to biases in the areas being sampled. The challenge is to test the use of legacy data for large-area mapping (e.g. national or continental extents) in order to limit the funds of field survey for large-area mapping. The problem is then to assess the reliability and quality of the legacy soil databases that have been mainly populated by traditional soil survey, and if there is a possibility of additional funding for sampling, to determine where new sampling units should be located. This additional sampling can be used to improve and validate the prediction model.

Latin hypercube sampling (LHS) has been proposed as a sampling design for digital soil mapping when there is no prior sample. We use the principle of hypercube sampling to assess the quality of existing soil data and guide us to locations that need to be sampled.

First an area is defined and the empirical environmental data layers or covariates are identified on a regular grid. The existing soil data are matched with the environmental variables. The HELS algorithm is used to check the occupancy of the legacy sampling units in the hypercube of the quantiles of the covarying environmental data. This is to determine whether legacy soil survey data occupy the hypercube uniformly or if there is over- or under-observation in the partitions of the hypercube. It also allows posterior estimation of the apparent probability of sample units being surveyed. From this information we can design further sampling. The methods are illustrated using legacy soil samples from Edgeroi, New South Wales, Australia, and from a large part of the Danube Basin. One third of the total number of sampling units are added to the original dataset. These new sampling units improve the representation of the feature space of the covariate. The standard deviation of the overall density is consequently smaller.

Keywords: Legacy soil data; Soil sampling; Hypercube sampling; Pedometrics; Soil survey; Digital soil mapping

1. Introduction

Legacy soil data arise from traditional soil survey (Bui and Moran, 2001). Methods of soil survey are generally empirical and based on the mental development of the surveyor, correlating soil with underlying geology, landforms, vegetation and air-photo interpretation. There are no statistical criteria for traditional soil sampling, this may lead to bias in the areas being sampled.

de Gruijter et al. (2006) offer some very thoughtful definitions in relation to sampling which we paraphrase here and use subsequently. *Sampling sensu lato* comprises selecting parts from a universe with the purpose of taking observations on them. The selected parts may be observed *in situ*, or material may be taken out from them for subsequent measurement in a laboratory. It is the collection of selected parts that is referred to as the *sample*. A single part that is, or could be, selected, is referred to as a *sampling unit*. The total number of sampling

* Corresponding author. Tel.: +39 0332 78 65 46; fax: +39 0332 78 63 94.
E-mail address: Florence.Carre@jrc.it (F. Carré).

0016-7061/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.geoderma.2007.01.018

units in a sample is referred to as the *sample size*. The material possibly taken from a sampling unit is referred to as an *aliquot*.

Using these definitions a legacy soil sample is a collection of sampling units that have been selected (probably) with unequal probability.

Usually, for national scale survey, the possibility of using legacy soil observation data is really expected since it can avoid the expense of new soil survey. But these data can also have limitations depending on:

- their description: the data have to answer the target mapping issue;
- their location/number: if the data must answer specific issues like “where to find soils with high agronomic potential”, their location might not be representative of the pedogenesis of the whole area. The main issue is that usually, these existing data have been sampled at different times in order to answer various and manifold questions. In this case, soil samples can have different soil variable descriptions and some areas of interest can be relatively over- or undersampled.

In digital soil mapping, soil samples may be used to elaborate quantitative relationships or models, between soil attributes and soil covariates, the *scorpan* variables (McBratney et al., 2003). Because the relationships are based on the soil observations, the quality of the resulting soil map depends also on the soil observation quality. Usually, a digital soil mapper tries to optimise the accuracy of the models and minimize the errors (Heuvelink et al., 2006), without taking into account the quality of the legacy data. (Legacy) sample quality evaluation has been done in forest ecology (Vanclay et al., 1995) and in computer science (Bisbal et al., 2005) but not in soil science. In this paper, we focus only on the location quality of the soil samples in the feature and geographic spaces.

An appropriate sampling design for digital soil mapping depends on how much data are available and where the data are located. Since the ‘90s, some statistical methods have been developed for optimizing data sampling for soil surveys. Some deal with the use of ancillary information. Heuvelink et al. (in press) designed the sampling minimizing the spatially averaged universal kriging variance. Simbahan and Dobermann (2006) compared three different optimization criteria: the minimization of the mean of the shortest distances, a uniform distribution of point pairs for variogram estimation (Warrick and Myers, 1987) and a combination of both. All of these methods involved simulated annealing algorithms (Van Groeningen and Stein, 1998; Fereyra et al., 2002). Hengl et al. (2003) proposed sampling along the principal components of the ancillary variables. The number of samples taken from each component is proportional to the total variance described by each of the component. Other methods do not involve ancillary information. Lark (2000) introduced fuzzy sets of grid spacings when there are uncertainties

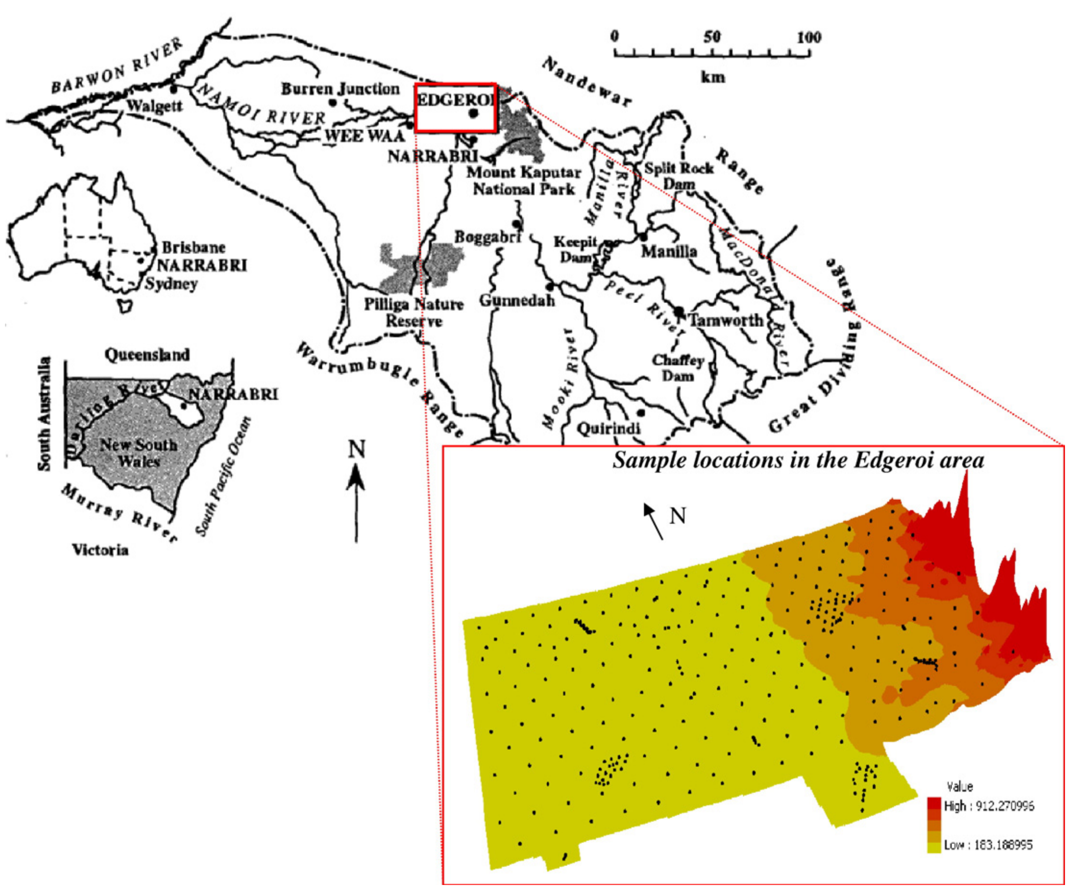

Fig. 1. Study location and the legacy sample location on the altitude of the Edgeroi area.