Geoderma 251–252 (2015) 24–32

Contents lists available at ScienceDirect

Geoderma

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

The implicit loss function for errors in soil information

R.M. Lark [a,*], K.V. Knights [b,c]

[a] British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, UK
[b] Geological Survey of Ireland, Beggars Bush, Haddington Road, Dublin 4, Ireland
[c] Centre for Environmental Geochemistry, British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, UK

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 4 March 2015
Accepted 8 March 2015
Available online 27 March 2015

Keywords:
Soil sampling
Soil monitoring
Uncertainty
Value of information
Loss function

ABSTRACT

The loss function expresses the costs to an organization that result from decisions made using erroneous information. In closely constrained circumstances, such as remediation of soil on contaminated land prior to development, it has proved possible to compute loss functions and to use these to guide rational decision making on the amount of resource to spend on sampling to collect soil information. In many circumstances it may not be possible to define loss functions prior to decision making on soil sampling. This may be the case when multiple decisions may be based on the soil information and the costs of errors are hard to predict. We propose the implicit loss function as a tool to aid decision making in these circumstances. Conditional on a logistical model which expresses costs of soil sampling as a function of effort, and statistical information from which the error of estimates can be modelled as a function of effort, the implicit loss function is the loss function which makes a particular decision on effort rational. After defining the implicit loss function we compute it for a number of arbitrary decisions on sampling effort for a hypothetical soil monitoring problem. This is based on a logistical model of sampling cost parameterized from a recent survey of soil in County Donegal, Ireland and on statistical parameters estimated with the aid of a process model for change in soil organic carbon. We show how the implicit loss function might provide a basis for reflection on a particular choice of sampling regime, specifically the simple random sample size, by comparing it with the values attributed to soil properties and functions. In a recent study rules were agreed to deal with uncertainty in soil carbon stocks for purposes of carbon trading by treating a percentile of the estimation distribution as the estimated value. We show that this is equivalent to setting a parameter of the implicit loss function, its asymmetry. We then discuss scope for further research to develop and apply the implicit loss function to help decision making by policy makers and regulators.

1. Introduction

The collection of soil information, both inventory and monitoring over time, is sponsored by various end-users including land-managers, regulators and policy-makers. In all cases the end-user must accept that there is uncertainty in the information which they obtain. This uncertainty could result in a cost due, for example, to over- or under-application of a fertilizer, a decision to implement unnecessary land remediation or failure to identify decline in soil quality and respond with appropriate policy. The uncertainty of soil information, given some fixed methodology, depends on the effort that can be deployed in field sampling, and so the cost to the sponsor. The sponsor is therefore faced with the problem of deciding how much effort it is appropriate to invest in soil sampling.

A rational approach to this problem is to choose a level of investment in soil sampling such that the benefit to the sponsor from the information over the cost of obtaining it is maximized. Yates (1949) was, perhaps, the first to point this out formally. To do this requires the specification of a *loss function*. A loss function expresses the costs incurred by a data-user (which may be an individual, a business or society at large) which result from using some estimate, $\tilde{x}$, of a quantity (for example, an estimate of the mean concentration of available phosphorus in the soil of a field) to make a decision (e.g., a fertilizer rate) when the true value of the quantity is x_t. The loss is, in general, non-zero when $\tilde{x} \neq x_t$, i.e., the information is erroneous. In our example the loss is incurred because of under-application of fertilizer and consequent loss of potential profitable yield ($\tilde{x} > x_t$) or wasteful over-fertilization ($\tilde{x} < x_t$) such that the marginal gain in yield does not cover the marginal cost of the input, and other costs may be incurred because of the environmental impact of the surplus nutrient. Because overestimation and underestimation incur losses for different reasons the loss function may be asymmetrical. Given a loss function and an error distribution for the information, one may make a decision which minimizes expected loss (e.g., Journel, 1984; Goovaerts, 1997). Some form of loss function, not necessarily a continuous function of the target variable, may be used to plan optimal sampling for decision-making (e.g., Yates, 1949; Ramsey et al., 2002; Boon et al., 2011) or to make decisions as to whether and how to

* Corresponding author.
E-mail address: mlark@nerc.ac.uk (R.M. Lark).

http://dx.doi.org/10.1016/j.geoderma.2015.03.014

supplement existing soil data by further sampling (e.g., Marchant et al., 2013).

Such rational planning of soil sampling requires that loss functions can be determined. This is plausible in some cases, where the analysis of decisions based on the soil information is relatively simple (e.g., remediate or do not remediate) and where reasonable values can be obtained for costs under different combinations of decision and future scenarios (chose to remediate — land was not contaminated; chose not to remediate — land was contaminated etc.). Some of the most sophisticated analyses of decision-making from uncertain soil information have been undertaken in the context of contaminated land where relatively simple decision trees based on single variables can be defined (e.g., Ramsey et al., 2002). Similar analyses have been undertaken for nutrient sampling at field scale by arable growers (Marchant et al., 2012). There is a wider literature on the use of loss functions for planning and control, particularly in manufacture (e.g., Freisleben, 2008; Pan and Chen, 2013), and these methodologies may be useful in environmental management and regulation. We call loss functions that can be developed in this way *explicit loss functions*.

In many cases, however, this is not a feasible approach. For example, when considering the design of a national-scale soil monitoring system for the UK, Black et al. (2008) asked sponsors (a range of regulators, government departments and public bodies responsible for environmental management) to give acceptable tolerances on estimates of regional and global mean values of soil properties, and changes in these properties. They then computed the costs of achieving these targets under different sampling regimes. Note that the process of defining acceptable tolerances was not straightforward, and was identified as an area for continued attention. Note also that the process was essentially 'open-loop'. There is no consistent method to evaluate whether the final costs are commensurate with the benefits of achieving the original target precision. Effectively it is assumed that the target precision must be achieved regardless of cost. However, if the sponsor decided that the total cost of the resulting scheme was unaffordable then it is not clear how to proceed, other than by assuming that the cost is fixed and reporting the corresponding precision.

It is, perhaps, not surprising that sophisticated decision analysis is possible for soil sampling on possibly-contaminated land, whereas planning of regional or national-scale soil monitoring and inventory remains 'open-loop'. In the former case there is generally a fairly simple binary decision to be supported (remediate or do not), and the costs under different decisions and scenarios (e.g., of remediating a site prior to development, of undertaking remediation after development on discovery that contaminants do exceed regulatory thresholds, etc.) can be reasonably approximated. For example, Ramsey et al. (2002) use approximate remediation costs, legal costs and liabilities in their case studies. In contrast, a soil monitoring scheme at regional or national scale will serve a range of purposes, not all of them foreseeable, and support a range of decisions and actions the consequences of which it is difficult to predict or quantify, let alone cost. One may therefore think it unlikely that policy makers or their advisors would be any more able to specify explicit loss functions for errors in soil information than they can specify acceptable confidence limits for estimates.

This could be regarded as an argument against any attempt to use a cost–benefit analysis when considering the design of soil inventory and monitoring, consistent with the criticisms of the ecosystem services valuation approach (Robinson et al., 2013) as voiced, for example, by Matulis (2014). However, Hansjürgens (2004) suggests, without conceding the broader agenda of monetizing the value of ecosystem components, that approaches based on cost–benefit analysis can provide a useful framework for the collection and evaluation of environmental information. That is the basis of our approach. Specifically we develop the concept of the *implicit loss function*. Consider a case of the 'open-loop' approach to planning of inventory and monitoring where a sponsor states that 'N samples are affordable'. The implicit loss function is the loss function implicit in that decision. That is to say it is the particular loss function which would lead to a selection of sample size N to maximize the benefit of sampling over its costs. In short, the implicit loss function, given some decision on how to undertake sampling, is the loss function under which that decision is rational. Our contention is that, by computing and examining implicit loss functions, one may, without entirely closing the planning loop, provide a basis for more rational reflection on sample effort by examining whether the form of the implicit loss function is congruent with the sponsor's expectations and any valuations of the target soil variable.

In this paper we develop the concept of the implicit loss function. While implicit loss functions have been used in financial analysis, we believe that they are a novel technology in the valuation of environmental information. There are three novel developments in this paper. First, we show that, for a specified sampling strategy which determines the precision of the resulting estimate as a function of sample size (e.g., a simple random sample from a variable of standard deviation σ), a given relationship between sample size and the cost of sampling and a specified asymmetry of the loss function, a unique implicit loss function exists for some specified sample size. Second, we point out that the asymmetry of the general linear loss function is implicit in certain criteria agreed in Australia for valuing soil carbon stocks from uncertain estimates. This suggests that the asymmetry of loss functions could be elicited from data users. Third, we use soil sampling records from a part of Ireland with rugged terrain and relatively sparse communications to develop a simple logistical model for sampling which allows us to estimate costs for particular sampling intensities. On the basis of these we present a hypothetical example of the implicit loss function for a case of monitoring change in soil carbon.

2. Theory

In this section we review the loss function and its use to determine optimal sample size, and develop the explicit expected loss under normal errors with a linear loss function. We then introduce the implicit loss function.

2.1. The loss function and optimal sample size

The most general form of the loss function is

$$L(\tilde{x}|x_{\mathrm{t}}) \tag{1}$$

which is the loss incurred as a result of a decision made on the assumption that some variable X takes the value $\tilde{x}$ when the true value is x_{t}. We define the loss as the difference between all costs incurred as a result of the decision between the present and some future time horizon over and above any costs that would be incurred as a result of making the decision on the assumption that $X = x_{\mathrm{t}}$. It follows that

$$L(\tilde{x}|x_t) = 0, \quad \forall\ \tilde{x} = x_t, \tag{2}$$

so one may think of $L(\tilde{x}|x_{\mathrm{t}})$ as the difference between the value of imperfect information $\tilde{x}$ and perfect information x_{t}. However,

$$L(\tilde{x}|x_{\mathrm{t}}) \geq 0, \quad \forall\ \tilde{x} \neq x_{\mathrm{t}}, \tag{3}$$

the perfect information is never worth less than the imperfect information, but is not necessarily worth more. If, for example, X is the concentration of a soil contaminant and remediation is required if and only if the concentration exceeds a regulatory threshold, $x > x_{\mathrm{R}}$, then the loss function in respect of decisions on remediation is zero for all cases where

$$\{\tilde{x} \leq x_{\mathrm{R}}, x_{\mathrm{t}} \leq x_{\mathrm{R}}\},$$