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

$$L(\widetilde{x}|x_t) \tag{1}$$

which is the loss incurred as a result of a decision made on the assumption that some variable X takes the value $\widetilde{x}$ 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 $X = x_t$. It follows that

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

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

$$L(\widetilde{x}|x_t) \geq 0, \quad \forall\ \widetilde{x} \neq x_t, \tag{3}$$

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, $x > x_R$, then the loss function in respect of decisions on remediation is zero for all cases where

$$\{\widetilde{x} \leq x_R, x_t \leq x_R\},$$

or

$$\{\widetilde{x}>x_R, x_t>x_R\}.$$

In some conditions we may treat the loss function as a function only of the error of $\widetilde{x}$ as an estimate of x_t:

$$L(\widetilde{x}-x_t). \tag{4}$$

This loss function assumes that the loss is independent of the absolute value of x_t, and is commonly used in discussion of estimation error and its implications. See, for example, Journel (1984) and Goovaerts (1997). We use it in this paper, although the ideas developed here could be extended to the more general case of Eq. (1).

Consider a case where we obtain an estimate of x_t by sampling. Sampling and application of an appropriate estimator, invoking some assumptions, gives us a conditional distribution for x_t (conditional on the sample). We assume that the sampling procedure is unbiased. The probability density function (pdf) for the conditional distribution is denoted by $f(x)$ and the cumulative distribution function (cdf) by $F(x)$ where

$$F(x_1) = \int_{-\infty}^{x_1} f(x)\mathrm{d}x. \tag{5}$$

If some value $\widetilde{x}$ is used as the estimate of x_t for decision making then the expected loss from the decision is

$$\mathcal{L} = \int_{-\infty}^{\infty} f(x)L(\widetilde{x}-x)\mathrm{d}x, \tag{6}$$

that is to say, the statistical expectation of the loss given the estimate and the conditional distribution of the sample mean.

If the loss function is a quadratic function of the error (as assumed by Yates, 1949) then the expected loss is minimized by using the expectation of the conditional distribution of x_t as the estimate, i.e., the sample mean. In this paper we follow Journel (1984) by using a general linear loss function:

$$\begin{aligned} L_l(\widetilde{x}-x_t) &= \alpha_1|\widetilde{x}-x_t| \quad x_t<\widetilde{x} \\ &= \alpha_2|\widetilde{x}-x_t| \quad x_t\geq\widetilde{x}. \end{aligned} \tag{7}$$

The parameters α_1 and α_2 have positive real values. In this case it can be shown (Journel, 1984) that the expected loss is minimized by setting $\widetilde{x}$ to:

$$\widetilde{x}_{\min} = F^{-1}\left(p_{\mathrm{opt}}\right), \tag{8}$$

where

$$p_{\mathrm{opt}} = \frac{\alpha_2}{\alpha_1+\alpha_2} \tag{9}$$

and $F^{-1}(p)$ denotes the inverse of the cdf, i.e., the pth quantile of the conditional distribution of x_t. For a symmetrical loss function $\widetilde{x}$ is therefore the median of the conditional distribution. This is equal to the mean if the conditional distribution is assumed to be Gaussian, which is justified in the simple random sampling case or under other probability sampling designs where independence of the samples allows the central limit theorem to be invoked for the distribution of the sample mean. However, in a case where $\alpha_2 > \alpha_1$ (i.e., a larger loss is incurred when $\widetilde{x}$ underestimates x_t than when it overestimates it by the same amount) the value of $\widetilde{x}$ is larger than the median. With the linear loss function the expected loss at $\widetilde{x}$ is given, following Journel (1984), by

$$\mathcal{L}_l = \alpha_2\left(x_t-\widetilde{x}+\frac{F(\widetilde{x})}{p_{opt}}\left[\widetilde{x}-\widetilde{\mu}_{\widetilde{x}}\right]\right), \tag{10}$$

where

$$\widetilde{\mu}_{\widetilde{x}} = \frac{1}{F(\widetilde{x})}\int_{-\infty}^{\widetilde{x}} xf(x)\mathrm{d}x. \tag{11}$$

If $\widetilde{x}$ is set to $\widetilde{x}_{\min}$ (Eq. (8)) then the minimum expected loss is

$$\breve{\mathcal{L}}_l = \alpha_2 x_t - (\alpha_1+\alpha_2)\int_{-\infty}^{\widetilde{x}_{\min}} xf(x)\mathrm{d}x. \tag{12}$$

If we obtain an estimate of x_t by simple random sampling across a domain of interest with a sample size of n, and the variable, X, has variance σ^2, then the conditional distribution of x_t is a Gaussian distribution (from the central limit theorem) with mean x_t (from the design-unbiasedness of simple random sampling) and variance σ^2/n (from the independence of the observations in simple random sampling). We write the pdf of this distribution as $f_G(x|x_t, \sigma^2/n)$.

Because we are considering a loss function which depends only on the estimation error and not on the absolute value of x_t we can, without loss of generality, set x_t to zero and write the minimum expected loss as a function of sample size n:

$$\breve{\mathcal{L}}_{l,G}\left(n|\sigma^2,\alpha_1,\alpha_2\right) = -(\alpha_1+\alpha_2)\int_{-\infty}^{\widetilde{x}_{G,\min}} xf_G\left(x|0,\sigma^2/n\right)\mathrm{d}x, \tag{13}$$

where $\widetilde{x}_{G,\min}$ is obtained with the inverse cdf for $f_G(x|0, \sigma^2/n)$:

$$\widetilde{x}_{G,\min} = F^{-1}\left(\frac{\alpha_2}{\alpha_1+\alpha_2}\middle|0,\sigma^2/n\right). \tag{14}$$

Fig. 1 shows a simple example of minimum expected loss as a function of sample size for a hypothetical case. The target variable is soil pH estimated to select the liming rate on a farm assuming a known buffering capacity. We assume that the standard deviation of pH across the

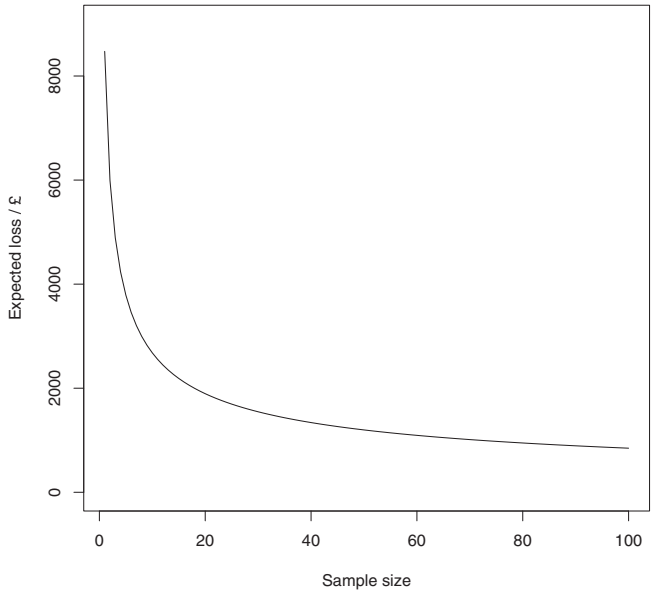

Fig. 1. Expected loss as a function of sample size for an asymmetrical loss function for error in determination of soil pH with α_1 = £10000 per pH unit error, $\alpha_2 = \alpha_1/3$ for a decision based on a simple random sample and the standard deviation of soil pH of 2 units.

farm is 2 pH units and that the slope of the linear loss function for a particular farm has $\alpha_1 = £10000$ per unit error in pH and $\alpha_2 = \alpha_1/3$. That is to say we assume that the loss due to a unit overestimation of pH and consequent under-liming and loss of potential profitable yield is three times the loss due to an equivalent underestimation of pH leading to overliming. Note that increasing sample size reduces the minimum expected loss, but with diminishing returns as the variance of the conditional distribution is proportional to n^{-1}. Assuming that the loss function gives loss in the same units as we may measure the costs of obtaining data for n samples, a sample size can be chosen at which the marginal cost of an additional sample is equal to the reduction in expected loss that the sample achieves.

In this paper we limit our discussion to simple random sampling, but one could extend the approach to more complex cases. For example, if an exhaustively-measured covariate such as a remote sensor image were available for the region, and this were correlated with the target variable, then one could consider the variance of the regression estimator for x_t (Brus, 2008). At a given sample size this variance would be smaller than for simple random sampling so the expected loss would be less.

2.2. The implicit loss function

As stated in Section 1, an implicit loss function is a loss function which makes some specific sample size a rational choice, given the marginal costs of sampling and the conditional distribution of x_t given the sample size. If the specified sample size is denoted by $\overline{n}$, and the variance of the variable X is σ^2 then the implicit loss function has parameters $\overline{\alpha}_1$ and $\overline{\alpha}_2$ such that

$$\breve{\mathcal{L}}_{l,G}\left(\overline{n}-1|\sigma^2,\overline{\alpha}_1,\overline{\alpha}_2\right)-\breve{\mathcal{L}}_{l,G}\left(\overline{n}|\sigma^2,\overline{\alpha}_1,\overline{\alpha}_2\right)=C(\overline{n})-C(\overline{n}-1), \quad (15)$$

where $C(n)$ is a function which returns the costs of a sample of n observations, assumed to be a real positive value for any positive n. Because the loss function is defined by two parameters there is not a unique solution $\{\overline{\alpha}_1,\overline{\alpha}_2\}$ to Eq. (15). However, if the asymmetry ratio is fixed, $\alpha_1/\alpha_2 = a$, then

$$\frac{\alpha_2}{\alpha_2+\alpha_1}=\frac{1}{a+1},$$

and so the integral on the right-hand-side of Eq. (13) depends only on a, n and σ^2. For notational simplicity we denote this value by $I(n, a, \sigma^2)$, then

$$\begin{aligned}&\breve{\mathcal{L}}_{l,G}\left(\overline{n}-1|\sigma^2,\overline{\alpha}_1,\overline{\alpha}_2\right)-\breve{\mathcal{L}}_{l,G}\left(\overline{n}|\sigma^2,\overline{\alpha}_1,\overline{\alpha}_2\right) \quad (16)\\ &\quad=-\alpha_2(1+a)\left\{I\left(\overline{n}-1,a,\sigma^2\right)-I\left(\overline{n},a,\sigma^2\right)\right\},\end{aligned}$$

which, with a, $\overline{n}$ and σ^2 all fixed, is linearly proportional to α_2. Because $C(\overline{n})-C(\overline{n}-1)$ is a positive constant for fixed $\overline{n}$ on the assumption that an additional sample point inevitably incurs some additional cost, it follows from Eq. (16) that, with specified a, $\overline{n}$ and σ^2 there exists a unique value of α_2 which provides a solution of Eq. (15). A numerical solution is necessary because the equation includes integrals of the normal density function.

In order to find a unique solution the asymmetry ratio a must be specified. In general one would expect loss functions to be asymmetric because the consequences of over- and under-estimation are generally different in kind and magnitude. Underestimation of soil carbon content may result in certain social costs from loss of production and unnecessary payment of incentives to land managers, whereas overestimation may result in insufficient investment in soil protection and incentives to improve soil management with long-term consequences for a range of soil functions.

An example of the selection of the loss asymmetry (albeit implicit) is provided by the Australian Government in their adoption of a practice for trading soil carbon stocks of uncertain magnitude (Department of the Environment, 2014). The practice is to use the 40th percentile of the sample distribution of soil carbon stock as the estimated value for trading purposes. This was selected both as an incentive for efficient sampling of stocks, and in explicit recognition that errors of over- and under-estimation have different consequences. With a linear loss function (Eq. (7)), the use of the 40th percentile as the effective estimate of the carbon stock implies (Eq. (9)) that

$$\frac{\alpha_2}{\alpha_1+\alpha_2}=0.4,$$

from which it follows that

$$a=\frac{\alpha_1}{\alpha_2}=1.5.$$

The asymmetry ratio is larger than 1.0 because the loss from trading carbon stocks overestimated by some amount is regarded as greater than that from trading similarly underestimated stocks. The selection of the percentile was not based explicitly on loss functions, but implies a slight preference for the interests of subsequent owners of the carbon stocks and for the environment, given the benefits of carbon sequestration, over those of the landowner.

The fact that data users in Australia were able to agree on a percentile of the sampling distribution to use as an estimator of soil carbon stocks is encouraging because it suggests that an elicitation procedure for the asymmetry ratio might be based on a consideration of percentiles of the sample distribution to treat as effective estimates. This is beyond the scope of the present paper. In general we propose that the implicit loss function is estimated for a range of asymmetry ratios which can be presented to the sponsor for consideration.

The idea of an implicit loss function is not entirely novel. It has been used in finance, for example to model how auditors make decisions about the collection of evidence (Scott, 1975). Estimation of the implicit loss function has been proposed by Elliott et al. (2005) as a method to elucidate the basis on which experts make financial forecasts. If one thinks of the expert's forecasting procedure as a tacit model estimation, then a loss function is effectively minimized, much as the quadratic loss function of a standard statistical estimation algorithm such as ordinary least squares. The recovery of an implicit loss function may explain apparent biases of forecasts in terms of asymmetry of the function. This procedure has been used to examine how members of the Federal Open Market Committee (FOMC) of the US Federal Reserve weight under- and over-prediction of economic variables such as inflation, growth rate and unemployment in terms of possible impacts through effects on the FOMC's decisions (Pierdzioch et al., 2013). We are not aware of a previous extension of this concept to our sampling problem.

3. Case study

3.1. The case study

We now illustrate the implicit loss function with an example. We consider a hypothetical region of 10000 km^2. We are interested in determining change in the regional mean stock of soil organic carbon over a period of time. To obtain the implicit loss function requires that we can approximate the variance of the sample mean of the target variable as a function of sample size, and the marginal cost of the nth soil sample. We discuss how this was done below. In brief, we follow Lark (2009) in using the soil carbon model of Nye and Greenland (1960) to compute distributions of soil carbon stocks and their changes under a change in land use based on a sample from a distribution of model parameters for the lowland tropics. We use detailed information on

sampling rate from a recent soil geochemical survey in Ireland (Knights and Scanlon, 2013) as a basis for the logistical component of the cost model. To make the logistical model as consistent as possible with a soil carbon model for the lowland tropics we extracted information on sampling rate in County Donegal in northwest Ireland, where relatively sparse communications and rugged terrain made field work most challenging.

3.2. Information on variability

To compute the implicit loss function we require information on the variance of the target variable. For purposes of this study we used a simple single-pool model of soil carbon, and sampled from a distribution of model parameters extracted from the literature for the lowland tropics, to obtain means and variance for soil carbon stock (t ha^{-1}) at two time points in a particular scenario. Lark (2009) describes the procedure in detail. The scenario we considered was forest land cleared for agriculture, with the initial or baseline survey undertaken 25 years after conversion and the resampling after a further 10 years. We added to the variance of the simulated data an analytical variance on the assumption that the coefficient of variation of analytical error is 5% (Landon, 1984). On this basis the mean carbon stocks 25 and 35 years post-clearance were 104 and 82 t ha^{-1} with standard deviations of 53 and 46 t ha^{-1} respectively. These are comparable with reported results for similar conditions in tropical and subtropical South America (Assad et al., 2013). If n samples are collected independently and at random on each date, and the standard deviations of carbon stock on the two dates are σ_1 and σ_2, then the standard error of the estimated mean change in carbon stock is

$$\sigma_c = \sqrt{\frac{\sigma_1}{n} + \frac{\sigma_2}{n}}. \tag{17}$$

3.3. The costs model

We developed a cost model on the basis of an analysis of the rate of soil sampling during the recently-completed Tellus Border survey (Knights and Scanlon, 2013) in six counties of Ireland (Donegal, Sligo, Leitrim, Cavan, Monaghan and Louth). This sampling was undertaken at an average rate of 0.25 samples km^{-2} by teams each of two workers. Analysis of the daily records of GPS locations allowed us to estimate the mean rate at which the teams sampled sites per county. For purposes of this paper we use the sampling rate for part of County Donegal, which was seven sites per team day, excluding local duplicate sampling. The rate of progress of sample teams across terrain in this part of Ireland was relatively slow. This can be attributed to the marked relief and complexity of the terrain which, over most of the land area of the county, is used for extensive grazing. The pronounced regional strike from northeast to south-west (Whittow, 1974) is reflected in the topography and restricted road access across the region. Rather than a uniform and isotropic road network allowing good access across the region, major roads follow the orientation of the regional strike, with relatively short branching access roads.

In this paper we consider a simple random sampling strategy. The empirical sampling rate from Donegal is from systematic sampling, because sample teams aimed to visit sample sites at the centre of 2-km square grid cells (Knights and Scanlon, 2013; Knights, 2013). One may expect such systematic sampling to be somewhat slower than an equivalent random sample because of the absence of short trips between points closer than the mean grid spacing. To adjust the Donegal sample rate to a simple random sampling equivalent we considered a notional 6 × 6-km region encompassing nine sample points set out according to the Tellus Border survey design. The shortest route around all these points is 18.83 km, assuming that the landscape can be traversed in a straight line. We then considered 1000 realizations of a simple random sample of 9 points in a 6 × 6-km region, computing the shortest route around each sample using the solve_TSP procedure from the TSP package for the R platform (Hahsler and Hornik, 2014; R core team, 2013). The mean distance travelled between points in a simple random sample was 16.66 km. The time per sample point in the systematic and simple random sampling regimes are therefore in the ratio 18.83/16.66 = 1.13. On the assumption that traveling speed is the same for random and for systematic sampling and that total time to undertake sampling is dominated by travelling time between points, the rescaled number of sample points per day under a simple random sampling scheme in Donegal is approximated as 7 × 1.13 = 7.9.

Following Beardwood et al. (1958), we assume that the distance travelled, d_n to visit n independently and randomly selected locations in a fixed area scales with n according to

$$\frac{d_n}{d_{n_1}} = \sqrt{\frac{n}{n_1}}. \tag{18}$$

On the assumption that the speed of travel is constant, and that total sampling time is dominated by travel between sites, we assume that sampling time is linearly proportional to distance travelled. On that basis the time to sample a unit area at density r, t_r (days km^{-2}), scales with sample density (r samples per km^2) as

$$\frac{t_r}{t_{r_1}} = \sqrt{\frac{r}{r_1}}. \tag{19}$$

The time per unit area in the Tellus Border survey in County Donegal at density 0.25 points km^{-2} was 0.25/7.9 = 31.6 × 10^{-3} days km^{-2}. The corresponding time per km^2 to sample at density r samples per km^2, where r is of similar order to the density of the Tellus Border survey, is therefore assumed to be

$$t_r = 31.6 \times 10^{-3}\sqrt{\frac{r}{0.25}}. \tag{20}$$

We assume that this scaling relationship holds over sample densities such that the sample rate per day is between about 2 and 11 (the range of sample rates in Donegal was 1 to 15).

On this basis one may compute the costs of sampling an area A km^2 with $n = rA$ points as

$$C(n) = \Omega + \nu n + \beta A \times 31.6 \times 10^{-3}\sqrt{\frac{n}{0.25A}}, \tag{21}$$

where Ω is the fixed costs, ν is the unit analytical cost and β is the field work cost of a team-day. For purposes of this study we assumed that ν = £20, based on preparation and analytical costs quoted in early 2014 by a UK-based company. We assumed that a team-day cost is β = £270, based on salary costs of technical staff and a two-person team. These figures are for illustrative purposes to develop the concept of the implicit loss function. Further refinement would be possible, for example to allow for economies of scale on analytical costs. In the case study we consider a monitoring programme with two time points and so we double the costs to visit sample points twice and compute two analyses at additional expense. Once again, further refinement would be possible to compute the costs at the start and end of the survey on a common net present value.

Fig. 2 shows variable costs ($c_n - \Omega$) and marginal costs for a single sampling campaign in a region of 10 000 km^2 with different sample sizes and using Eq. (21) with the constants in the previous paragraph. and assuming a baseline and resampling campaign. Note that the marginal cost of an extra sample decreases with sample size.

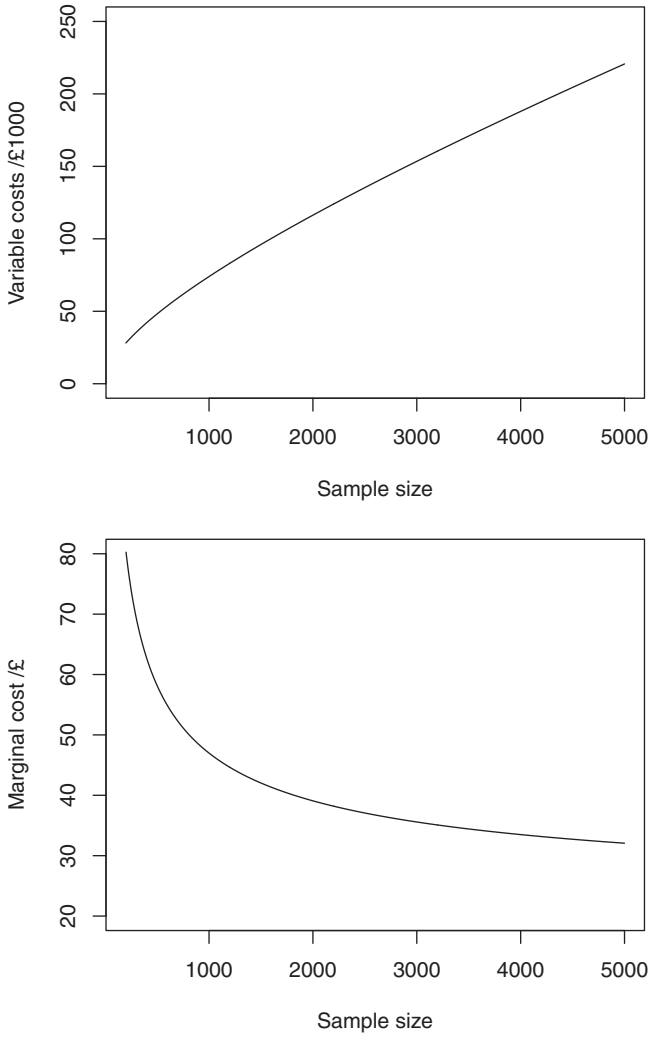

Fig. 2. Variable costs (top) and marginal cost (bottom) for sampling a 10000 km^2 region with different sample sizes (one campaign), $v = £20, \beta = £270$.

3.4. Implicit loss functions

Given a sample size n the standard error of the estimate of the change in soil carbon stock from two independent random samples was computed from Eq. (17). The cost of sampling at this intensity (over fixed costs) was computed from Eq. (21). For some sample size and specified value of the asymmetry ratio, a, we found the value of α_1 that satisfies Eq. (16) using the optim procedure in the R platform (R core team, 2013).

In this case study we consider an environment where we expect ongoing reductions in soil carbon stocks because, even with no change in the mean inputs of carbon to the soil, it is likely still to be approaching a new steady-state soil carbon stock under new land use. The policy maker wishes to know the mean rate of this change across the 10000-km^2 region to formulate policy in respect of the role of soil in the carbon budget and likely implications for soil functions including agricultural production, the modulation of surface water flows and stability of soil against erosion by water or wind.

The variable that we consider is the loss of soil carbon stock, and so positive errors mean that this loss is underestimated. We considered an asymmetry ratio of 1, and alternatives smaller than one. By excluding asymmetry ratios larger than 1 we make an assumption that underestimation of the loss of soil carbon never incurs smaller costs than overestimation. This seems reasonable, since underestimation may result in complacency about soil quality, the amount of carbon that remains sequestered in soil and the success of existing policy on land use and soil protection with implications for future food security, water resource management etc. However, overestimation of the loss may result in undue regulatory burdens on producers, excessive expenditure if it is decided to offset loss of soil carbon from agricultural land and possible distortions in land use which may have implications for food prices.

Some comparisons may be drawn between the asymmetry of the loss function in this case and that for soil carbon trading, referred to in Section 2.2 above, where there is a small preference for underestimation of stock. The carbon trading case is simpler in that we are considering only the value of the soil carbon in a particular market. However, at least in principle, this integrates at least some of the factors of interest here: specifically the value of soil carbon as an offset for carbon emissions, and also the asymmetry of interests between landowners selling the carbon, and the subsequent owners and environmental beneficiaries of the offset. The asymmetry ratio for estimation of tradable carbon stocks was 1.5, in the case of estimates of loss of soil carbon the equivalent ratio is the reciprocal of this, $0.\dot{6}$, because the equivalent preference is for an overestimate of loss of stock. We therefore considered this asymmetry ratio for our case study, along with two rather smaller ratios, 0.4 and 0.2 which imply stronger preferences for overestimation. For completeness we also present the symmetrical implicit loss function.

Fig. 3 shows the implicit loss functions with the four asymmetry ratios for the case where 500 sample points are proposed for each of the baseline and resampling surveys. The values of the α_2 parameter for $a = 1, 0.\dot{6}$, 0.4 and 0.2 are approximately £50000, £60000, £80000 and £124000 t^{-1} ha soil carbon. In a 10000 km^2 region these units are equivalent to £per Mt error in the estimated loss of total soil carbon stock.

Fig. 4 shows the slope of the implicit loss function (α_1 and α_2) for the same asymmetries and different proposed sample sizes.

3.5. Interpretation of the implicit loss functions

How might a policy-maker interpret the implicit loss function? First, recall that we propose the implicit loss function for situations where a loss function is not straightforward to specify. This is because of the complexity of the policy decisions informed by soil information, the relevance of this information to different sectors, uncertainty about future costs of interventions and uncertainty about the efficacy of policy options

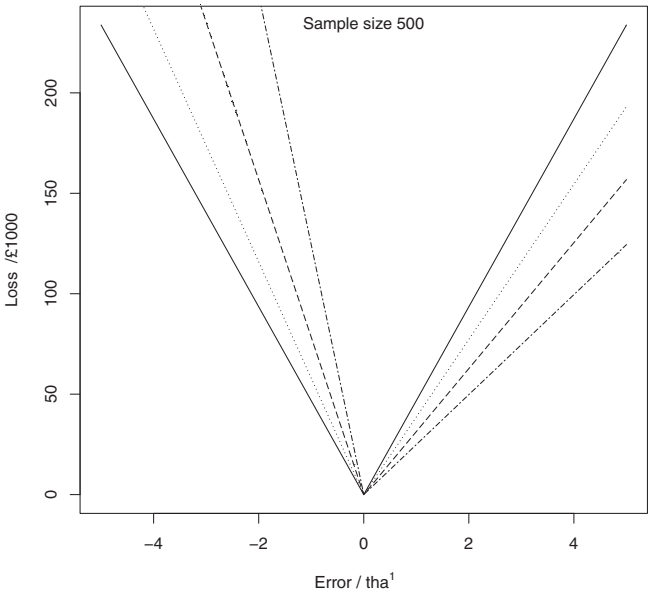

Fig. 3. Implicit loss function for error in estimated mean reduction in soil carbon stock assuming a total sample size of 500. Symmetrical function (solid line) or asymmetrical with $a = \alpha_1/\alpha_2 = 0.\dot{6}$ (dotted line), $a = 0.5$ (dashed line) or $a = 0.2$ (dashed and dotted line).

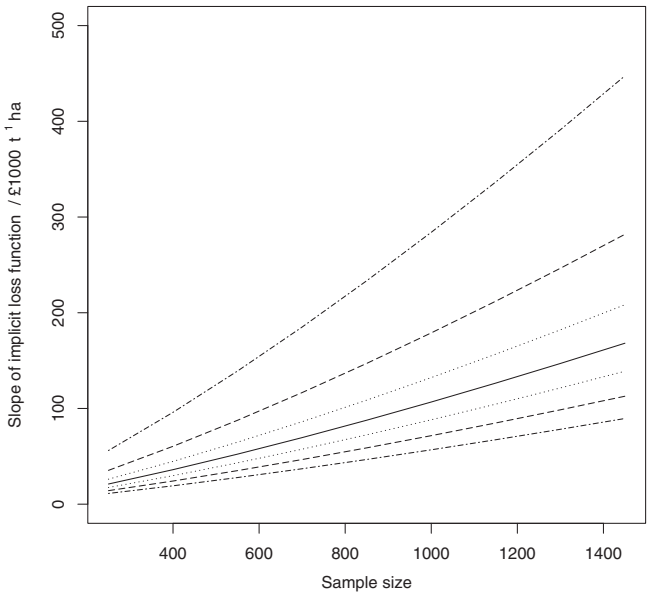

Fig. 4. Slope of the implicit loss function for error in estimated mean reduction in soil carbon stock for a range of sample sizes. Symmetrical function (solid line) or asymmetrical with $a = \alpha_1/\alpha_2 = 0.\dot{6}$ (dotted line), $a = 0.5$ (dashed line) or $a = 0.2$ (dashed and dotted line).

or specific interventions which might be based on the information. It is this complexity that makes it impossible to make the sampling decision a closed-loop process in the sense of Section 1. The point of the implicit loss function is to exhibit the assumptions that are implicit in a particular decision on sampling so that they can be open to general scrutiny.

Consider a hypothetical example. In our 10 000-km^2 region soil scientists have proposed a sample size of 2000 for the two-phase soil monitoring procedure (somewhat sparser than the Tellus Border sample density). However, based on initial budgetary considerations, the officials who make the decisions on resources propose reducing this to 500. If the policy-maker takes an asymmetry $a = 0.\dot{6}$ on the grounds that this loss function implies a mild preference for environmental considerations, then the value of α_2 for the implicit loss function for a sample size of 500, expressed as loss per unit error in the reduction of total soil carbon stock for the region is £60 000 Mt^{-1} and for a sample size of 2000 it is £300 000 Mt^{-1}.

To support a decision on sampling one must decide which of these two loss functions is most plausible. For reasons already enunciated this process cannot be formal, but it can be systematic. First, one may identify the possible consequences of an error. For example, the possible consequences of underestimation of the loss of soil carbon stock include:

1. Failure to prioritize soil protection with respect to competing policy areas.
2. Failure to implement appropriate soil protection measures and in consequence of this
   - Failure to improve food production — c.f. examples presented by Lal (2004) who quotes yield gains for cereal or legume crops between 1 and 40 kg ha^{-1} from increases of soil organic carbon of 1 t ha^{-1}.
   - Failure to sustain the soil's capacity to modulate water flows by accepting infiltration and allowing groundwater recharge.
3. Failure to account correctly for the role of soil in the regional greenhouse gas budget.

A similar set of consequences for overestimation of loss of soil carbon can be identified, for example:

1. Imposition of excessive regulation on producers, with consequences for sustainability, employment and food security.
2. Distortions in policy priorities with respect to other areas.
3. Overestimation of soil contribution to the regional greenhouse gas budget, with financial costs if this is offset by carbon trading or other mechanisms.

Reflection on these lists may allow refinement of any previous choices of asymmetry ratio and the implicit balance of preferences between these considerations. One may then consider any numerical information pertinent to these factors. For example, one might assume that the value of carbon in trading schemes reflects social costs of emissions. The cost of 1 Mt of carbon may then be approximated as £18 m based on a cost of €6 t^{-1} under the European Emissions Allowance scheme (EUA) (costs and exchange rates in August 2014). Some severe qualifications are required here. First, it is certainly not clear that EUA prices at present reflect social costs but rather particular policy objectives and various institutional factors affect their price (Lutz et al., 2013). Second, the EUA scheme can only be indicative, at present carbon stocks associated with land use and land use change are not tradable in the scheme.

Accepting these qualifications, one may take as a starting point the observation that an underestimation by 1 Mt of the soil carbon lost from a region is an underestimation by £18 m of the costs imposed by soil carbon loss. However, we may not assume that this underestimation, through its effects on policy decisions, results directly in a social cost of the same size. This does not translate simply into costs due to the effect of this error on future policy. First, one must note that future carbon losses, other factors remaining equal, will be smaller as soil carbon stocks approach a steady state. Second, one must ask how successful any policy or mitigation measures would be in reducing these losses even if they were based on error-free information. Nonetheless, it may be argued that very severe discounting for future uncertainty and uncertainty about the consequences of policy decisions is required to reduce the market-based value of 1 Mt of soil carbon to a value smaller than α_2 for the implicit loss function for a sample size of 2000.

4. Discussion

We have defined the implicit loss function for errors in soil information and shown that a unique implicit loss function exists for any specified sample size given a logistical model of sampling costs, and information on the variability of the target property. With an example we have shown how the implicit loss function allows us to exhibit the assumptions implicit in any decision on sample size in an 'open loop' context where the costs and benefits of environmental information cannot be simply and directly compared. In the context of our example we have made some tentative suggestions about how the implicit loss function could be used to reflect on such sampling decisions, without claiming that this 'closes the loop'. The implicit loss function is a novel concept in the valuation of environmental information, and we suggest that it merits further investigation.

The amount of field effort to be deployed to address a question in environmental science, management or policy is often a fraught matter between field scientists and their sponsors. Yates (1952), discussing the investment of resources in research for agricultural development wrote: 'With the present drive for economy there is serious danger that even such facilities as are available for experimental work of this kind will be curtailed or not used to full advantage. It is therefore important to stress that such curtailment will result in much more substantial and immediate losses through failure to determine the best practices.' We recognize that we write from one side of this fence, but offer the implicit loss function method as a tool to improve communication across that fence. Environmental scientists increasingly recognize the importance of effective communication of environmental information to decision makers in the presence of uncertainty, and we suggest that the implicit loss function is potentially a contribution to this task.

There is scope for further development of this work. It would be informative to undertake experiments with groups with policy or

regulatory responsibilities to examine implicit loss functions for notional tasks in the commissioning of soil inventory or monitoring. The objective would be to assess whether and how the implicit loss function helps the decision-making process. One approach would be to present the experimental subjects with a series of implicit loss functions corresponding to different sample sizes, without initially disclosing the sample sizes or total sampling cost, and to elicit a view as to which function best represents the socio-economic and environmental costs of error in environmental information. One could then ask the group to make a decision on sample effort given the sample effort and cost that corresponds to the selected loss function, and others close to it. One could then compare these results with decisions made purely from the costs of sampling. One useful extension of this work would be to show how the approach could be used to choose a partition of a fixed total resource between two or more competing projects.

Another way to develop this approach would be to work with a policy or regulatory group in a post-hoc analysis of past projects, regarded as more or less successful. One might ask managers, for example, to assign such projects to groups characterized in terms such as:

1. 'A "Rolls-Royce" study: it was useful but we suspect that the effort was excessive when we look at the true costs'
2. 'This provided useful information, we would pay for it again in comparable circumstances'
3. 'This was a waste of time and resources. There was too much uncertainty in the final results, which were therefore hard to interpret'

One would then undertake a comparable elicitation of plausible implicit loss functions for each project, and test the hypothesis that these results would be congruent with the classification (i.e., the selected loss function for projects in class 1 would have smaller slopes than the implicit loss function for the actual project sample size, the selected loss function would more or less match the implicit loss function for the actual project sample size in class 2, and the selected loss function would be steeper than the implicit loss function for cases in class 3).

There is scope for further work on using elicitation to obtain the asymmetry of loss functions. It was interesting that the asymmetry of the general linear loss function is implicit in the percentile-based approach used in the Australian Carbon trading scheme, and suggests that percentiles may provide a basis for such an elicitation. This could be useful for development of the implicit loss function, but could also facilitate the development of explicit loss functions for rational sample planning.

Further work is also needed to include economies of scale and Net Present Values in the implicit loss function, and to improve the logistical model to make it more flexible. Beckett (1981) presents a review and evaluation of logistical models in the context of agricultural extension and soil survey. There may be scope to develop this model and to calibrate it with records from the Tellus Border survey and similar sampling exercises. While attempts have been made in the past to compute costs for notional soil sampling schemes of different intensity (Black et al., 2008) we are not aware of previous systematic attempts to calibrate models from GPS records of the movement of sampling teams. Given the widespread use of GPS in field work, and the scope to download daily records, the collation of such information from sampling schemes with different designs in different conditions could be informative and useful for planning.

5. Conclusions

We defined the implicit loss function and exemplified it, using a process model to compute statistical parameters for a soil monitoring problem and records from a survey in Ireland to provide a logistical model. The implicit loss function is offered as a method to aid decision making on soil sampling problems where the costs of errors in soil information are not sufficiently clear cut to support a classical value of information analysis. This will often be the case in soil sampling and monitoring at regional and national scale. In such circumstances the selection of a level of investment in sampling may not be based on the information required but rather on arbitrary constraints. The implicit loss function allows one to exhibit the implicit assumptions in making a decision to invest a certain amount of resource in soil sampling, and we propose that this could help in reflection on this decision and on comparisons between levels of investment in different projects.

Acknowledgments

This paper is published with the permission of the Director of the British Geological Survey (NERC) and the Director of the Geological Survey of Ireland. Tellus Border is part-financed by the European Union's INTERREG IVA cross-border Programme managed by the Special EU Programmes Body. We acknowledge the contribution of Bob Cooper who organized and edited the data on sample times from which we extracted sampling rates for the Tellus Border survey. We are grateful to Dr Ichsani Wheeler for helpful exchanges on soil carbon trading schemes.

References

Assad, E.D., Pinto, H.S., Martins, S.C., Groppo, J.D., Salgado, P.R., Evangelista, B., Vasconcellos, E., Sano, E.E., Pavão, E., Luna, R., Camargo, P.B., Martinelli, L.A., 2013. Changes in soil carbon stocks in Brazil due to land use: paired site comparisons and a regional pasture soil survey. Biogeosciences 10, 6141–6160.

Beardwood, J., Halton, J.H., Hammersley, J.M., 1958. The shortest path through many points. Proc. Camb. Philos. Soc. 55, 299–327.

Beckett, P.H.T., 1981. Logistics of agricultural extension — foreword and part 1: the component parts of a logistic model. Agric. Adm. 8, 177–208.

Black, H., Bellamy, P., Creamer, R., Elston, D., Emmett, B., Frogbrook, Z., Hudson, G., Jordan, C., Lark, M., Lilly, A., Marchant, B., Plum, S., Potts, J., Reynolds, B., Thompson, P., Booth, P., 2008. Design and operation of a UK soil monitoring network. Science Report — SC060073. Environment Agency, Bristol.

Boon, K.A., Rostron, P., Ramsey, M.H., 2011. An exploration of the interplay between the measurement uncertainty and the number of samples in contaminated land investigations. Geostand. Geoanal. Res. 35, 353–367.

Brus, D., 2008. Using regression models in design-based estimation of spatial means of soil properties. Eur. J. Soil Sci. 51, 159–172.

Department of the Environment, 2014. Measurement-based methodology for sequestering carbon in soils in grazing systems. Government of Australia, Department of the Environment, Canberra (http://www.climatechange.gov.au/sites/climatechange/files/files/reducing-carbon/cfi/methodologies/cfi-sequestering-carbon.pdf).

Elliott, G., Komunjer, I., Timmermann, A., 2005. Estimation and testing of forecast rationality under flexible loss. Rev. Econ. Stud. 72, 1107–1125.

Freisleben, J., 2008. A proposal for an economic quality loss function. Int. J. Prod. Econ. 113, 1012–1024.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.

Hahsler, M., Hornik, K., 2014. TSP: traveling salesperson problem (TSP). R package version 1.0–9. (http://CRAN.R-project.org/package=TSP).

Hansjürgens, B., 2004. Economic valuation through cost–benefit analysis: possibilities and limitations. Toxicology 205, 241–252.

Journel, A.G., 1984. mAD and conditional quantile estimators. In: Verly, G., David, M., Journel, A.G., Marechal, A. (Eds.), Geostatistics for Natural Resources Characterization. D. Reidel, Dordrecht, pp. 261–270 (Part 1).

Knights, K.V., 2013. Quality control and statistical summaries of Tellus Border topsoil regional geochemical data. Report Version 1.0. Geological Survey of Ireland and Geological Survey of Northern Ireland joint report.

Knights, K.V., Scanlon, R.P., 2013. Tellus: regional-scale baseline geochemical mapping of soil, stream sediment and stream water for the island of Ireland. Mineral. Mag. 77, 1482.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627.

Landon, J.R., 1984. Booker Tropical Soil Manual. Longman, New York.

Lark, R.M., 2009. Estimating the regional mean status and change of soil properties: two distinct objectives for soil survey. Eur. J. Soil Sci. 60, 748–756.

Lutz, B.J., Pigorsch, U., Rotfuß, W., 2013. Nonlinearity in cap-and-trade systems: the EUA price and its fundamentals. Energy Econ. 40, 222–232.

Marchant, B.P., Dailey, A.G., Lark, R.M., 2012. Cost-effective sampling strategies for soil management. Home-Grown Cereals Authority Research and Development Project Report No, 485. HGCA, London (Available at http://www.hgca.com/media/252469/pr485.pdf).

Marchant, B.P., McBratney, A.B., Lark, R.M., Minasny, B., 2013. Optimized multi-phase sampling for soil remediation surveys. Spat. Stat. 4, 1–13.

Matulis, B.S., 2014. The economic valuation of nature: a question of justice? Ecol. Econ. 104, 155–157.

Nye, P.H., Greenland, D.J., 1960. The Soil under Shifting Cultivation. Commonwealth Bureau of Soils Technical Communication No 51. Commonwealth Agricultural Bureaux, Farnham Royal.

Pan, J.-N., Chen, S.-C., 2013. A loss-function based approach for evaluating reliability improvement of an engineering design. Expert Syst. Appl. 20, 5703–5708.
Pierdzioch, C., Rülke, J.-C., Tillmann, P., 2013. Using forecasts to uncover the loss function of FOMC members. Joint Discussion Paper Series in Economics by the Universities of Aachen, Gießen, Göttingen, Marburg and Siegen. No. 02 (https://www.uni-marburg.de/fb02/makro/forschung/magkspapers/02-2013_tillmann.pdf).
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 3-900051-07-0 (URL http://www.R-project.org/).
Ramsey, M.H., Taylor, P.D., Lee, J.C., 2002. Optimized contaminated land investigation at minimum overall cost to achieve fitness-for-purpose. J. Environ. Monit. 4, 809–814.
Robinson, D.A., Jackson, B.M., Clothier, B.E., Dominati, E.J., Marchant, S.C., Cooper, D.M., Bristow, K.L., 2013. Advances in soil ecosystem services: concepts, models and applications for earth system life support. Vadose Zone J. 12.
Scott, W.R., 1975. Auditor's loss functions implicit in consumption-investment models. J. Account. Res. 13, 98–117.
Whittow, J.B., 1974. Geology and Scenery in Ireland. Penguin, Harmondsworth.
Yates, F., 1949. Sampling Methods for Censuses and Surveys. Griffin, London.
Yates, F., 1952. Principles governing the amount of experimentation in development work. Nature 170, 138–140.