European Journal of Soil Science

European Journal of Soil Science, March 2017, **68**, 221–234 doi: 10.1111/ejss.12413

Controlling the marginal false discovery rate in inferences from a soil dataset with α-investment

R. M. Lark
British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, UK

Summary

Large datasets on soil provide a temptation to search for relations between variables and then to model and make inferences about them with statistical methods more properly used to test preplanned hypotheses on data from designed experiments or sample surveys. The control of family-wise error rate (FWER) is one way to improve the robustness of inferences from tests of multiple hypotheses. In its simplest form, hypothesis testing with FWER control lacks statistical power. The α-investment approach to controlling the marginal false discovery rate is one method proposed to improve statistical power. In this paper I outline the α-investment approach and then demonstrate it in the analysis of a dataset on the rate of CO_2 emission from incubated intact cores of soil from a transect over Cretaceous rocks in eastern England. Hypotheses are advanced after considering the literature and examining relations among the available soil variables that might be proposed as explanatory factors for the variation of CO_2 emissions. They are then tested in sequence with α-investment, such that the rejection of null hypotheses increases the power to reject later ones, while controlling the overall marginal false discovery rate at a specified value. This paper illustrates the use of α-investment to test a multiple set of hypotheses on a soil dataset; statistical power is improved by ordering the sequence of hypotheses on the basis of process knowledge. The approach could be useful in other areas of soil science where covariates must be selected for predictive statistical models, notably in the development of pedotransfer functions and in digital soil mapping.

Highlights

- α-investment controls marginal false discovery rate in statistical inference.
- Hypotheses were advanced about soil factors that affect CO_2 emission from soil.
- These hypotheses were tested in sequence with control of marginal false discovery rate.
- Soil properties, land use and parent material were significant predictors.

Introduction

Increasingly, soil science is undertaken with neither specific experiments nor data from bespoke surveys, but rather with pre-existing datasets, sometimes public, collected for a different primary purpose. Such data may be very useful, but the problem of how to evaluate statistically the weight of evidence that they provide for a hypothesis is subtly different from the classical case of the designed experiment. This is not recognized sufficiently in scientific practice.

In a bespoke experiment a scientist proposes a set of hypotheses. These hypotheses correspond to the main effects of factors that are varied in the experiment, or their interactions. The hypotheses can be examined by evaluating a particular contrast between treatment means, which would be zero under a 'null hypothesis'. In conventional frequentist inference the evaluation of a hypothesis is supported by computing a *P*-value, the probability of obtaining evidence as strong as or stronger than the observed contrast if the null hypothesis were true. The *P*-value therefore helps the scientist to decide whether the experimental data are sufficient to support a rejection of the null hypothesis and a scientifically interesting interpretation of the observed contrast. Soil scientists may also work with observational data rather than experiments, but in principle it is still possible to use a set of preplanned contrasts to test prior hypotheses. For example, Lark & Scheib (2013) tested preplanned hypotheses about the effects of land use on lead content of topsoil with data from the British Geological Survey's London Earth sample survey.

Correspondence: R. M. Lark. E-mail: mlark@bgs.ac.uk
Received 22 August 2016; revised version accepted 11 December 2016

© 2017 British Geological Survey, NERC *European Journal of Soil Science* © 2017 John Wiley & Sons Ltd.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

This paper is concerned with a different problem. Here datasets are unstructured, in the sense that they do not arise from an experimental or sampling design selected to test a set of predetermined hypotheses. They may contain a single target variable of interest (e.g. some function of the soil system) and many other variables that describe the state of the soil system, its outputs or its environment. The temptation is to trawl the data with plots and correlation coefficients to search for variables that have some apparent statistical relation to the variable of interest, and then to evaluate the evidence for such a relation by fitting linear models and applying some kind of significance test. This approach is sometimes called '*P*-hacking' (Wasserstein & Lazar, 2016).

The problem with this approach is that the *P*-value for a test of the significance of a term in a linear model indicates the probability that evidence as strong as or stronger than that provided by the data would emerge against a null hypothesis of no effect when that particular null hypothesis has been specified in advance. This interpretation is not valid if the null hypothesis has been selected for testing because the experimental evidence against it looks strong in a table of means or on a plot. Consider the following thought experiment. Generate a notional dataset with one soil system response variable of interest and 100 potential continuous explanatory variables of interest, but generate these 'data' as independent draws from a random number generator. Next, undertake a regression of the response variable on each explanatory variable, testing the null hypothesis of no linear relation and rejecting a potential explanatory variable for which the *P*-value of this test exceeds 0.05. By definition of the *P*-value, the expected number of null hypotheses rejected with $P < 0.05$ (potential explanatory factors regarded as significant) would be 5. This means that in a test of all 100 null hypotheses, all of which are true, the probability of rejecting at least one with $P < 0.05$ is 0.994. This is clearly not satisfactory.

Reflection on this thought experiment shows why a crude *P*-hacking exercise can be expected to lead to 'discovery' of effects in a dataset that do not have the statistical evidence that is claimed for them. The *P*-hacking strategy must be eschewed. However, the thought experiment points to difficulties in approaches to data analysis that might be less evidently problematic. Consider a case where a soil scientist wishes to use a large dataset to identify factors that control some response of the soil system. It is proposed to test several statistical models in which the response is the dependent variable and subsets of soil properties constitute the independent variables. The aim would be to select the model that is best-fitting on some criterion and for which the data provide evidence of a significant relation. The key to identifying the problem with this approach is to recognize that, in the analysis of an unstructured dataset with k proposed models, we do not really undertake k separate tests, each of which represents a scientifically framed hypothesis. Rather, we test a family of k hypotheses to discover whether one or more are of interest. This is known as multiple hypothesis testing (Tukey, 1991). If we adopt a rule of rejecting the null hypothesis for a model (proposed in advance) if $P < \alpha_T$ then we know that the probability of falsely rejecting a true null hypothesis is α_T. But in the multiple-testing case, the probability that we will reject at least one true null hypothesis among our set of k is somewhat larger. Approaches that use multiple testing can be controlled statistically provided that the models to be tested are proposed in advance. In this paper I consider an approach to multiple hypothesis testing that can be used by the soil scientist who rejects *P*-hacking strategies, but who wishes to explore a set of possible explanatory models through a set of multiple tests.

One general approach to statistically controlled multiple testing uses Bonferroni's inequality. If we denote the (unobserved) number of falsely-rejected null hypotheses out of a set of k (not necessarily independent) by the random variable $V(k)$, then the family-wise error rate (FWER) may be defined as:

$$\mathrm{FWER} = P\left[V\left(k\right) \geq 1\right], \tag{1}$$

where $P[\cdot]$ denotes the probability of the term in brackets and Bonferroni's equality provides an upper bound for this:

$$P\left[V\left(k\right) \geq 1\right] \leq \sum_{j=1}^{k} P\left[V_j = 1\right], \tag{2}$$

where V_j is an indicator that is 1 if the jth null hypothesis is falsely rejected and 0 otherwise. A common strategy to apply Bonferroni's inequality is to test each hypothesis against a threshold *P*-value of

$$\frac{\alpha}{k} \tag{3}$$

to impose an upper bound of α on the FWER. This is potentially a very conservative procedure because for a large k the threshold can be very small. A less conservative approach is to control not the FWER but the false discovery rate (FDR) defined as:

$$\mathrm{FDR} = \mathrm{E}\left[\left.\frac{V\left(k\right)}{R\left(k\right)}\right| R\left(k\right) > 0\right] P\left[R\left(k\right) > 0\right], \tag{4}$$

where $R(k)$ is an (observed) random variable, the number of rejected null hypotheses out of k. Methods to control the FDR have been proposed for independent sets of tests (Benjamini & Hochberg, 1995) and for tests that cannot be regarded as independent (Benjamini & Yekutieli, 2001).

These approaches have been used in soil science. For example, Lark *et al.* (2007) used FDR control to select variables for predictive models of soil properties. Turner *et al.* (2008) used FDR control to examine differences between subpopulations of the plant *Arabidopsis lyrata* from sites with soil formed over contrasting parent materials with respect to the occurrence of a large number of genes. The FDR control therefore provides a solution for the soil scientist who wants to test multiple hypotheses on an observational dataset without making spurious claims of statistical significance for selected models. However, whether controlling FWER with a Bonferroni-based procedure or controlling FDR with one of the methods mentioned above, there is a penalty of loss of power to detect effects. Even in the case of FDR control, for the independent

case the smallest P-value must be less than the threshold for FWER control if any null hypotheses are to be rejected.

The common FWER-control criterion in Equation (3) is not the only criterion that will satisfy control at FWER $< \alpha$ given Equation (2) because the latter requires only that the thresholds sum to α over all k tests, and setting all thresholds to k/α is not a unique solution. In principle one might apply different thresholds to different hypotheses, provided that the latter are not selected on the basis of their P-values. This approach is known as α-spending because it entails a prior decision about how to allocate a total 'resource' over a set of k tests. Tukey (1991) points out that this approach allows one to test some preselected hypotheses at a larger threshold for P than is given in Equation (3) provided that the thresholds over all k tests sum to α. This approach has been used to design strategies for hypothesis testing in medical trials.

Foster & Stine (2008) propose an extension of the α-spending strategy to what they call α-investment. This is discussed further in the theory section, but in short they show that control of what is called the marginal false discovery rate can be achieved by a strategy in which hypotheses are tested sequentially at thresholds calculated by a rule under which the rejection of a null hypothesis allows larger thresholds to be used to test subsequent ones. If one thinks of the specified acceptable FWER under α-spending as a total 'wealth' to be distributed over tests, then the strategy of Foster & Stine (2008) offers the opportunity to increase that wealth by the judicious selection of hypotheses early in the sequence to increase the power to detect significant effects in subsequent tests. It should go without saying that this selection of hypotheses is not based on the data, but on scientific considerations. This means that it is inherent in α-investment that the search for significant effects in a dataset is not delegated to an algorithm, but requires an element of scientific judgement.

The approach to statistical inference based on α-investment is new and has had limited application at the present time. Koenig *et al.* (2015) show how the method can be used to ensure robust inference from clinical trials, and Karp *et al.* (in press) discuss its use for large genetic screening studies where multiple traits are examined. The objective of this paper is to demonstrate application of the α-investment approach to inference to a soil dataset. The next section outlines the theory of α-investment and is followed by an application of this to a case study on the rate of CO_2 emission from soil cores. I then discuss the more general implications for inference from soil datasets.

Theory

Marginal false discovery rate

The α-investment strategy of Foster & Stine controls the marginal false discovery rate, mFDR_η, which is defined as

$$\text{mFDR}_\eta = \frac{\text{E}\left[V(k)\right]}{\text{E}\left[R(k)\right] + \eta}, \tag{5}$$

where $V(k)$ and $R(k)$ have the same meaning as in Equation (4). The constant $\eta > 0$ is added to ensure that mFDR_η can be controlled in the case of the 'complete null hypothesis' where all k hypotheses are true.

The term V_j in Equation (2) is an indicator, equal to 1 when the jth null hypothesis is incorrectly rejected and 0 otherwise; therefore, Equation (2) can be written as:

$$\begin{aligned} P\left[V(k) \geq 1\right] &\leq \sum_{j=1}^{k} \text{E}\left[V_j\right], \\ &\leq \text{E}\left[V(k)\right]. \end{aligned} \tag{6}$$

If $\text{mFDR}_\eta \leq \alpha$ then this implies that, for the complete null hypothesis (where all rejected null hypotheses are falsely rejected and $V(k) = R(k)$),

$$\text{E}\left[V(k)\right] \leq \frac{\alpha\eta}{1-\alpha}, \tag{7}$$

and so, if $\eta = 1 - \alpha$, we can combine Equations (6) and (7) to obtain:

$$P\left[V(k) \geq 1\right] \leq \text{E}\left[V(k)\right] \leq \alpha. \tag{8}$$

That is to say, controlling mFDR_η at α with $\eta = 1 - \alpha$ controls FWER at α in the case of the complete null hypothesis, which is called control of FWER in the weak sense. Foster & Stine (2008) report a series of simulation studies that show that control of FDR, as defined in Equation (4), and of mFDR_η, $\eta = 0.95$, gives very similar control of false discovery over a range of conditions.

α-investment

A set of up to k hypotheses is tested in a sequence. Under an α-investing rule, the jth null hypothesis is rejected if $p_j < \alpha_j$. The threshold value for the jth test, α_j, depends on the alpha-wealth after the previous test, $W(j-1)$. If the jth test results in rejection of the null hypothesis then there is an increase in the alpha-wealth and $W(j) > W(j-1)$, otherwise $W(j) < W(j-1)$. The sequence of tests ends either at the kth test or at the jth test if $W(j)$ goes to zero.

Foster & Stine (2008) show that an α-investing rule (governing costs and payouts) of the form:

$$\begin{aligned} W(j) - W(j-1) &= \omega \qquad \text{if } p_j \leq \alpha_j, \\ &= -\frac{\alpha_j}{1-\alpha_j} \qquad \text{if } p_j > \alpha_j, \end{aligned} \tag{9}$$

with initial α-wealth $W(0) \leq \alpha\eta$ and payout $\omega \leq \alpha$ controls mFDR_η at level α if there is a finite stopping time (i.e. the procedure will stop after some not necessarily predetermined number of rejections). They go on to show that a procedure that always spends some (*hopeful*), but not all (*thrifty*), of its α-wealth testing the next hypothesis has a finite stopping time.

The proof of this property by Foster & Stine (2008) requires the condition that:

$$\text{E}\left[V_j | R_{j-1}, R_{j-2}, \ldots, R_1\right] \leq \alpha_j, \tag{10}$$

where $R_i \in \{0, 1\}$ is an indicator, denoting the acceptance or rejection of hypothesis i (i.e. the properties of a particular test hold regardless of the history of rejections) (Aharoni & Rosset, 2014). This is a weaker assumption than that of independence of the tests, although it is easiest to show that it holds if the tests are independent. D.P. Foster (personal communication, 2015) suggests one practical procedure in a linear modelling setting is to orthogonalize a new predictor on any predictors for which the hypothesis was rejected, on the basis that little information is provided by any test i for which $R_i = 0$. This approach is applied by Lin *et al.* (2011). In this study I took a comparable approach. I developed a single model for the response variable sequentially. I tested a hypothesis about each potential explanatory variable in turn, and in the sequence preplanned for the α-investment procedure. Each test was on the likelihood ratio statistic to compare a null model (that includes all variables for which the null hypothesis had previously been rejected) with a full model that includes the variables in the null model and the new one of interest.

One strategy to determine the threshold for rejection of the jth null hypothesis, given the wealth after the previous test, is given by Foster & Stine (2008). They propose that:

$$\alpha_j = W(j-1)\left(\frac{1}{1+j-h^*} \vee \frac{1}{1+k-j}\right), \qquad (11)$$

where h^* is the index of the last rejected hypothesis. The two terms in brackets are alternatives; the second is applied to the kth test. This strategy is hopeful and thrifty, in the senses defined above. Given the rationale for α-investment, the ordering of hypotheses for tests should put those with the strongest scientific rationale early in the sequence. The ordering can be dynamic, in that, for example, one might choose to add a quadratic term in some predictor only if an earlier test of a linear effect led to rejection of the null hypothesis.

In this study I followed a conservative hopeful and thrifty strategy by setting the maximum level for any test to α so that:

$$\alpha_j = \min\left\{\alpha, W(j-1)\left(\frac{1}{1+j-h^*} \vee \frac{1}{1+k-j}\right)\right\}. \qquad (12)$$

I set $\alpha = 0.05$ and followed Aharoni & Rosset (2014) in setting $\omega = W(0) = \alpha\eta$.

Case study

The question and the data

In this case study I consider a dataset from a transect across a part of Bedfordshire in eastern England. Details of this transect have been published previously (Haskard *et al.,* 2010). The objective of this original study was to examine the spatial variation of nitrous oxide emissions, but data on the rate of CO_2 emissions were also measured, although hitherto they have not been analysed.

The first site on the 7.5-km transect was at 508 329, 237 450 on the British National Grid, and subsequent points were at approximately 30-m intervals on a line of bearing 173.5° from due north. The soil types on this transect were described in the survey by King (1969), who mapped them according to a legend of associations of soil series from the classification of the Soil Survey of England and Wales. The soil was under woodland, on arable land (with a germinated autumn-sown crop, recently cultivated or under stubble) or under rough grass (paddocks, field margins and one sports field). Details of the sampling procedure and the laboratory analyses are given elsewhere (Haskard *et al.*, 2010; Lark & Milne, 2016). In addition to the rate of CO_2 emission, the following soil variables were measured: bulk density, volumetric water content, soil pH, soil organic carbon (per cent by mass), total nitrogen (per cent by mass), soil nitrate and ammonium (μg g^{-1} soil). The volumetric water content was converted to the water-filled pore space fraction (WFPS) after Linn & Doran (1984) using the measured bulk density (Minasny *et al.*, 1999; Lark & Milne, 2016). The land use was also recorded at each site.

This study was restricted to the first two soil associations mapped by King (1969) on the transect (locations 1–171), the Cottenham and Wicken associations. The soils of these two associations formed in superficial material over the Lower Greensand and the Gault Clay, respectively. They correspond to associations of Arenosols and Cambisols (Cottenham association) and Cambisols, Luvisols and Acrisols (Wicken association) according to the World Reference Base classification (IUSS Working Group WRB, 2006). The remaining soil on the transect formed over chalk units (see Lark & Milne, 2016) and contains large amounts of calcium carbonate. This was measured on each soil sample by the water-filled calcimeter method of Williams (1949) for correction of the determination of carbon content by combustion. It was decided to exclude the chalk soil because of the possibility of a contribution to the measured emission rate from abiotic sources (Aciego Pietri & Brookes, 2008). In addition, exploratory analysis showed considerable variation of the organic C to total N ratio, suggesting that the determination of organic carbon was prone to error because of the correction for inorganic carbon. Table 1 presents summary statistics on the soil properties for the 171 locations. Histograms are shown in Figure 1 and the variables are plotted against location on the transect in Figure 2. Figure 3 shows the land use at locations on the transect and Figure 4 shows box and whisker plots for the soil properties within the three land uses.

The objective of this case study was to identify variables in this set, or derived from them, that account for spatial variation in the rate of CO_2 emission by soil. To achieve this I first considered reasons why available variables might be hypothesized to be explanatory factors for this process. I then examined exploratory statistics of the predictor variables only, to identify possible redundancies between correlated variables. On the basis of these considerations I then proposed a sequence of hypotheses to test with α-investment. The identification of these hypotheses and the associated exploratory analyses that supported the decision are not reported in the main paper for reasons of space, but are described in Supporting Information.

Table 1 Summary statistics on available soil variables

Variable	Units	Mean	Median	SD	Min.	Max.	Skewness
pH		6.82	7.19	1.16	3.65	8.03	−1.34
SOC	Gravimetric % dry soil	2.99	2.54	1.82	1.23	19.87	5.13
log SOC	log (gravimetric % dry soil)	0.99	0.93	0.41	0.21	2.99	1.15
CN		10.98	9.93	3.22	8.32	24.47	2.42
log CN		2.36	2.30	0.23	2.12	3.20	2.08
WFPS		0.77	0.81	0.17	0.35	1.00	−0.59
BD	g cm^{-3}	1.21	1.21	0.22	0.63	1.86	−0.09
Ammonium	μg g^{-1}	1.69	0.29	5.41	0.02	40.41	5.06
log Ammonium	log (μg g^{-1})	−0.99	−1.24	1.38	−3.91	3.70	1.24

log, natural logarithm; SOC, soil organic carbon; C:N, ratio of total carbon to total nitrogen; WFPS, water-filled pore space fraction; BD, bulk density.

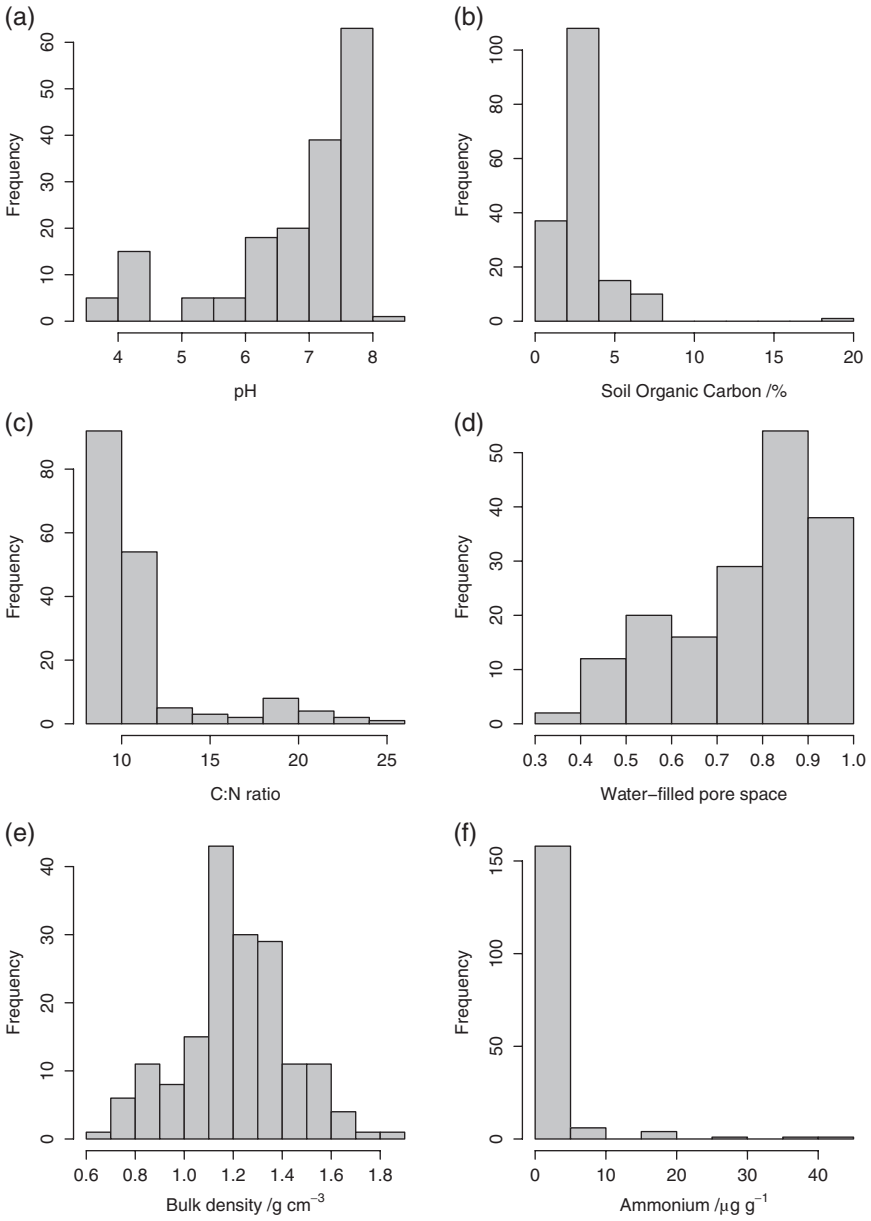

Figure 1 Histograms of predictor variables.

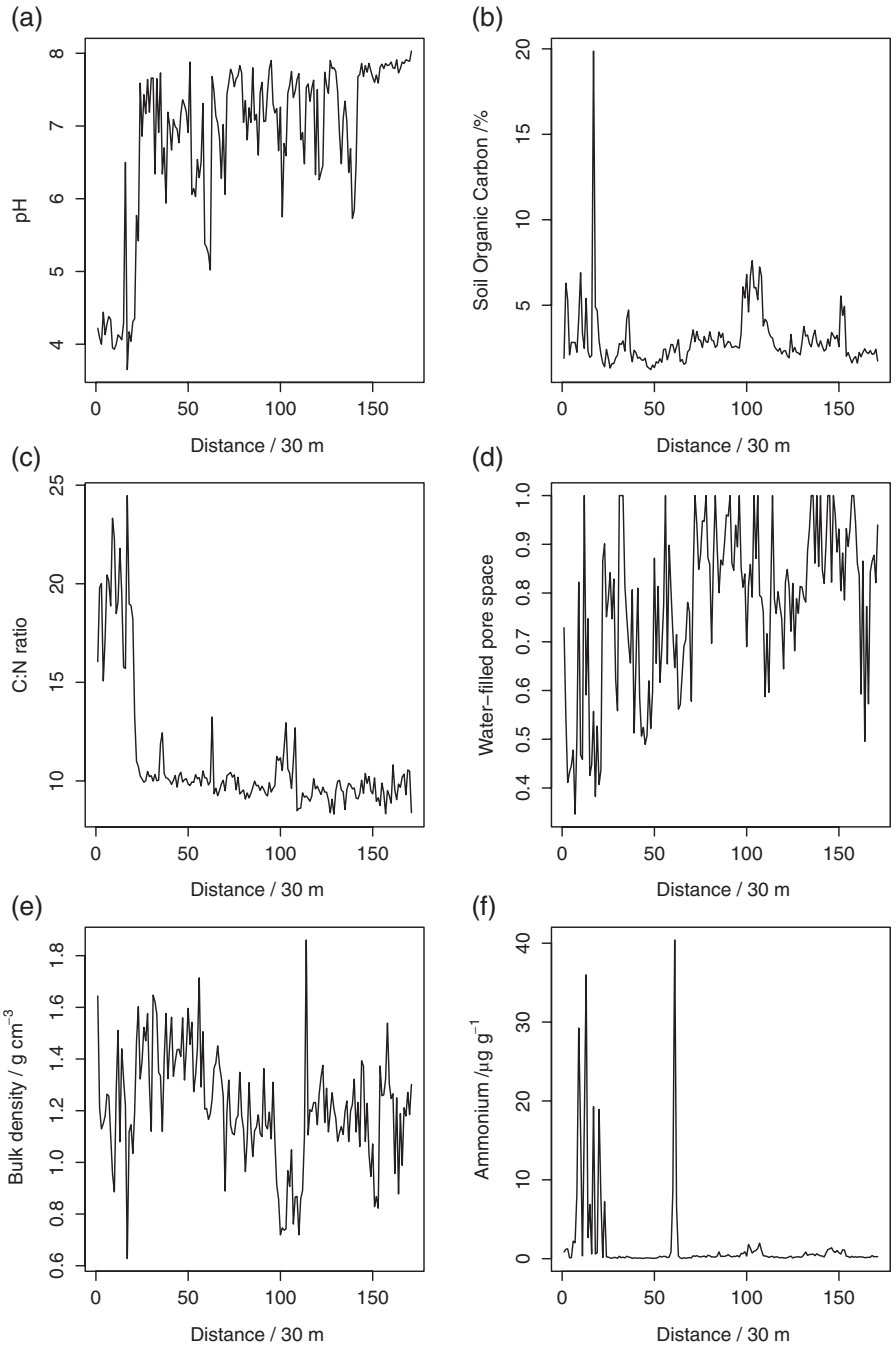

Figure 2 Values of predictor variable plotted against location on the transect. Location is distance rescaled by the sampling interval (30 m).

Hypotheses to be tested

1 *A linear effect of soil organic carbon concentration*. This is included because, of the available variables, it is the best proxy for the substrate available to the soil microflora. A hypothesis that the rate of CO_2 emission from soil depends, in part, on the soil organic carbon (SOC) concentration is plausible biologically. Because this variable is strongly skewed (Table 1) it was decided to transform it to natural logarithms before using it as a predictor in the model. As a scientifically plausible hypothesis, it is sound practice to test this early in the sequence in line with the 'best-foot forward' principle of Foster & Stine (2008). If a strong hypothesis is placed early in the sequence on the basis of scientific plausibility it will boost the initial alpha wealth.

2 *A linear effect of ammonium*. This is included to test the hypothesis that ammonium concentration and CO_2 emission are both signs of mineralization, a process that links the C and N cycles in soil. It is possible that such an effect will not be observed because much of the carbon released by mineralization may be

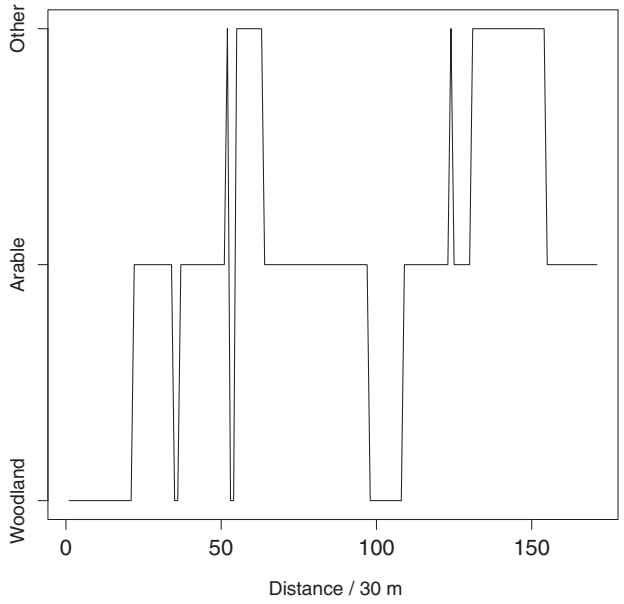

Figure 3 Land uses at locations on the transect. Location is distance rescaled by the sampling interval (30 m).

in simpler organic forms that are not respired directly, and the ammonium released may be assimilated rapidly by plant roots or microorganisms in some circumstances. Because this variable is strongly skewed (Table 1), I transformed it to natural logarithms before using it as a predictor in the model.

3 *A linear contrast between land uses*. As noted above, land use appears to reflect many soil properties, including SOC and ammonium, which are already proposed as predictors. For this reason I proposed an initial test of the linear contrast between arable and non-arable land uses. Arable soil is the most intensively managed, so other factors that might affect CO_2 emission might be represented by this contrast, such as the effect of autumn cultivations or residual fertilizer effects. One more orthogonal contrast between land uses can be proposed (woodland against grass), but this is introduced later in the sequence.

4 *A linear effect of bulk density*. Although bulk density is correlated with SOC, evidence from two previous studies (Moyano *et al.*, 2012; Saiz *et al.*, 2006) suggests that there may be additional effects even when SOC is already represented in a linear model.

5 *A second-order polynomial function of water-filled pore space*. As discussed above, previous studies have shown an effect of WFPS fraction on CO_2 emission from soil. Moyano *et al.* (2013) identified a quadratic effect, so here I propose a polynomial function of WFPS with linear and quadratic terms.

6 *A linear effect of soil pH*. There are some types of soil on the transect with pH values that, according to Aciego Pietri & Brookes (2008), would be sufficiently small to have a limiting effect on soil respiration. For most of these soils, however, the pH is outside the range where these authors expect to see an effect. For this reason this factor was included relatively late in the sequence as a speculative hypothesis.

7 *A linear effect of the ratio of soil organic carbon to total N*. This ratio is fairly stable at around 10 for most soil types on the transect, but there are some larger values on the more acid soil, predominantly under woodland in the north of the transect. Because this variable is somewhat skewed (Table 1), I transformed it to natural logarithms before using it as a predictor in the model. Again, this factor is included late in the sequence as a speculative hypothesis that 'overflow respiration' in the presence of surplus organic carbon might be a factor causing variation in CO_2 emissions.

8 *The contrast between the two soil associations*. This is the contrast between the Cottenham association and the Wicken association. Various soil properties were considered earlier in the sequence so this is included to account for any differences in the soil that these individual variables do not represent. These could include structural and textural differences between soils that contrast markedly with respect to parent material.

9 *A linear contrast between land uses*. This is the remaining contrast orthogonal to the one introduced at (3) above, Woodland against Grassland.

Analysis. After exploration of the predictor variables and formulation of the list of hypotheses presented in the previous section, exploratory analysis was carried out on the CO_2 emission data. Summary statistics are listed in Table 2. Histograms of the raw data and data transformed to natural logarithms are shown in Figure 5, and these variables are plotted against position on the transect in Figure 6. The raw data have a skewness coefficient of 1.26. The assumption of normality applies to the random components in the linear mixed model used to analyse these data, and not the data on the dependent variable itself. I used the transformed data for this modelling and checked the distribution of the residuals when the modelling was complete.

Recall that, to make the condition in Equation (10) plausible, I proposed that these successive hypotheses are tested in the context of a single linear model formed by adding and testing terms successively and retaining those for which the null hypothesis is rejected. This means that a new term is tested by adding it to a model that contains predictors for which the null hypothesis has been rejected. This was done in the context of a linear mixed model (Verbeke & Molenberghs, 2000) and the fixed effects comprised all previouslyselected soil properties and the random effects comprised a spatially correlated random variable and an uncorrelated nugget term (Lark & Cullis, 2004). This model was necessary because the sampled sites were not selected according to an independent random sampling design, and so the residuals from the fixed effects model cannot be treated as independent random variables. I estimated the model parameters by maximum likelihood with the likfit procedure for the geoR package on the R platform (Diggle & Ribeiro, 2007; R Core Team, 2014). Maximum likelihood was chosen because this allows models with different

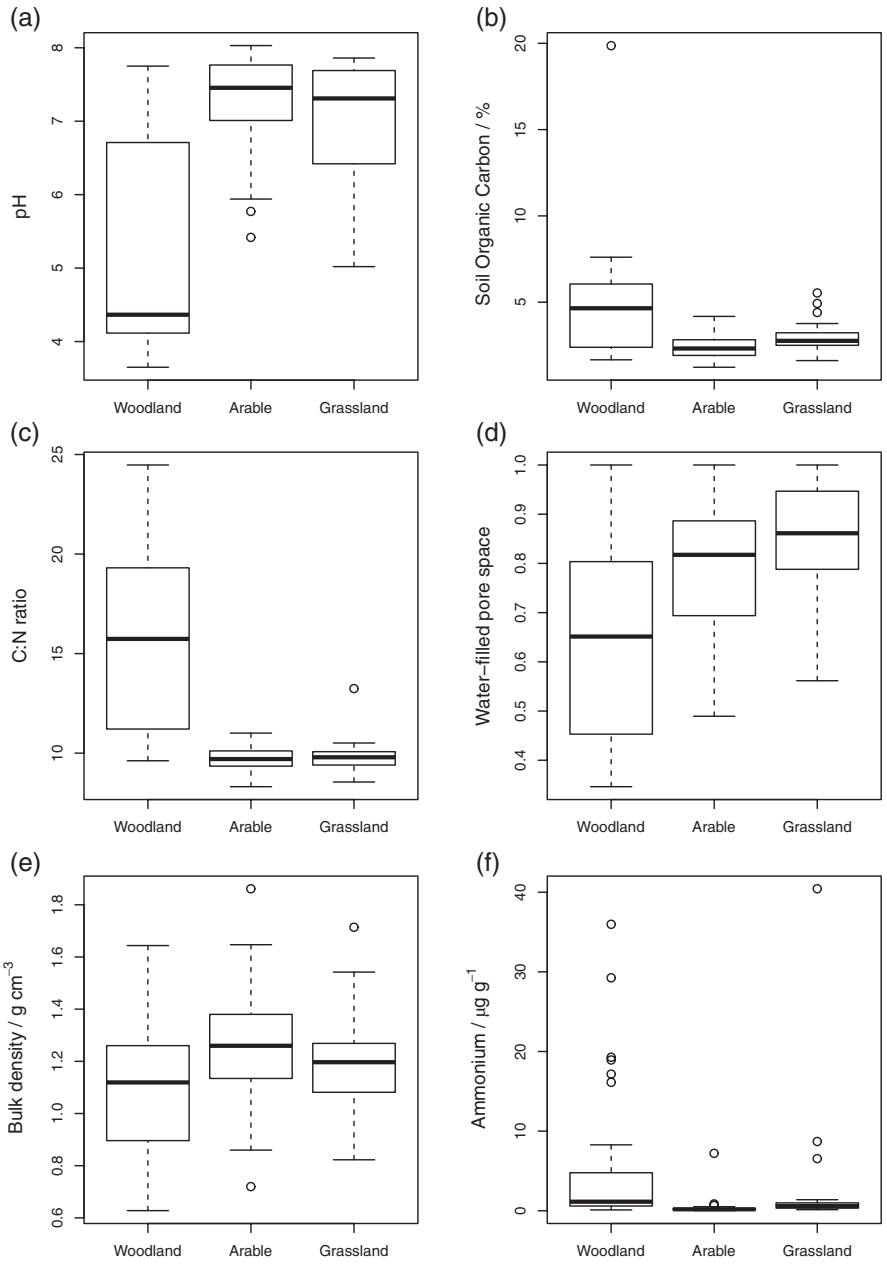

Figure 4 Box and whisker plots for variables within land-use categories. Circles on the plots are outlying values, defined as values more than 1.5 times the interquartile range above the third quartile or below the first quartile of the data.

sets of fixed effects to be compared by a log-likelihood ratio test. If ℓ_N is the maximized log-likelihood for a model with a set of fixed effects selected previously and ℓ_F is the maximized log-likelihood for a model with the same set of fixed effects and p additional terms then the log-likelihood ratio statistic L,

$$L = 2\left(\ell_F - \ell_N\right), \tag{13}$$

has a χ^2 asymptotic distribution with p degrees of freedom under the null hypothesis that there is no effect of the additional p terms. This was the basis for testing terms, with $p = 1$ in each case except for bulk density where a linear and a quadratic term in the proposed variable were added together.

Table 2 Summary statistics on rate of CO_2 emission on original and logarithmic scales (natural logarithms)

Units for rate of CO_2 emission	Mean	Median	SD	Min.	Max.	Skewness
µg kg^{-1} day^{-1}	10 753.6	9246.2	6698.0	1871.8	43 763.3	1.26
log µg kg^{-1} day^{-1}	9.10	9.10	0.66	7.50	10.70	−0.36
Residual log µg kg^{-1} day^{-1}	0	−0.01	0.40	0.43	−0.77	0.62

Also included are statistics for the residuals from the complete model computed at the end of the analysis.

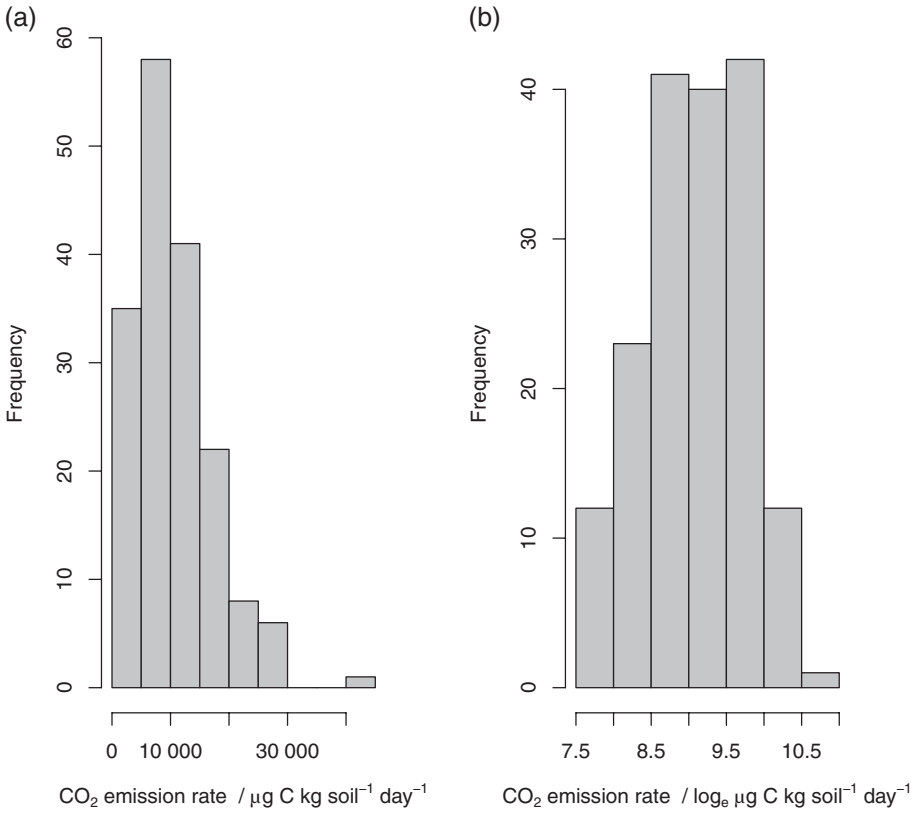

Figure 5 Histograms of rates of CO_2 emission on (a) original and (b) logarithmic scales.

The initial model was fitted with a constant mean as the only fixed effect. Models were fitted with spherical and with exponential covariance functions (Webster & Oliver, 2007), and the model with the largest likelihood was selected, and the particular covariance function was retained for all subsequent models. The first hypothesis about a soil variable, SOC, was tested by comparing a model with fixed effects SOC and an intercept with the initial model in which the constant mean was the only fixed effect. In this case $p = 1$.

The α-investment was undertaken to specify the P-value at which a null hypothesis would be rejected. The marginal false discovery rate ($mFDR_\eta$) was controlled with $\alpha = 0.05$ and η was therefore 0.95. The 'payout' on rejection of the null hypothesis was $\omega = \alpha\eta = 0.0475$. The α-investing rule given in Equation (9) was applied and the level of each test was determined by Equation (12). Successive hypotheses were tested in the order introduced above until either $W(j)$ went to zero or all nine hypotheses were tested.

Results

Figure 7 shows (a) the value of $W(j)$, the alpha-wealth, after each test and (b) P-values for the successive tests (open circles) and the levels, α_j, against which each hypothesis was tested with $mFDR_{0.95}$ controlled at 0.05 and with α-investment (solid discs). The horizontal broken line shows the equivalent threshold for testing nine hypotheses with FWER control at 0.05 by Bonferroni's inequality.

The first three null hypotheses are rejected, which is why the alpha-wealth grows after each of these tests. These are the effects of SOC, ammonium and the contrast between arable and non-arable land use. The next four null hypotheses are accepted (bulk density, WFPS fraction, soil pH and ratio of soil organic carbon to total nitrogen). The null hypothesis is rejected for the comparison between soil associations, although the depletion of alpha-wealth means that the test was against a threshold near 0.01. The null hypothesis is then accepted for the contrast between woodland and grassland. It is notable that the four hypotheses rejected in this case would also be rejected against the Bonferroni thresholds, although it is clear from Figure 7(b) that the power to reject hypotheses under α-investment is much larger.

After the hypothesis testing I examined the fitted models more closely. Table 3 gives the model effects. For each listed variable Table 3 gives the fixed effect coefficient from the model where that variable was first added. For the continuous variables the effect is the regression coefficient, for the contrasts between levels of a categorical variable (land use or soil association) the effect is the (additive) difference between the two levels. In each case the standard error of the effect is also given. Figure 8 shows the variogram models, which describe the random effects of (i) the models with a constant mean as the only fixed effect and (ii) each model where a soil variable was added as an independent variable and the null hypothesis was rejected. It is not possible to compute an R^2 for a linear model fitted by maximum likelihood with correlated random effects. Instead I used these models to compute

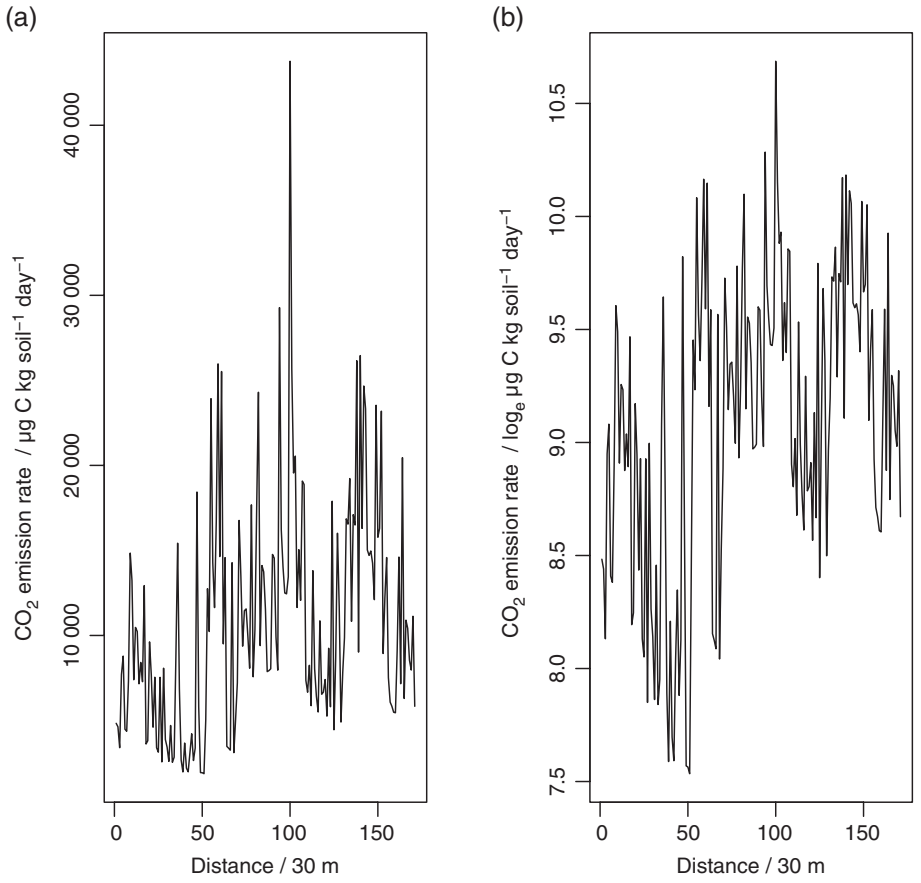

Figure 6 Plots of rates of CO_2 emission against location on the transect, on (a) original and (b) logarithmic scales. Location is distance rescaled by the sampling interval (30 m).

the dispersion variance of the random effects along a 5-km transect computed as the double integral of the variogram over the transect (Webster & Oliver, 2007). If the dispersion variance of the random effects for the model with a constant mean as the only fixed effect is denoted $\breve{\sigma}^2_{N0}$ and the dispersion variance for the jth model in the sequence is denoted $\breve{\sigma}^2_N j$ then I define the approximate adjusted R^2 for the jth model as

$$\breve{R}^2_{adj} = 1 - \frac{\breve{\sigma}^2_{Nj}}{\breve{\sigma}^2_{N0}}. \quad (14)$$

Values of $\breve{R}^2_{adj}$ are provided in Table 3.

The first null hypothesis tested in this sequence is the effect of SOC. The null hypothesis is rejected; note that $\breve{R}^2_{adj}$ for the model with SOC as the only predictor is small (0.14). The SOC effect is positive and about 3.3 times its standard error. Comparison of the variograms for this model with the variogram for the model with a constant mean as the only fixed effect (Figure 8) shows some reduction in the variance of the correlated random term, but little change in the spatial scale of this variation. The null hypothesis of no effect of soil ammonium concentration is also rejected. The fixed effect coefficient is positive and four times its standard error, and including this term increases $\breve{R}^2_{adj}$ to 0.24.

Figure 9 was prepared only after the modelling was complete. It shows plots of the rate of CO_2 emission against (a) SOC and (b) ammonium concentration (all variables are on log scales). These graphs show the positive relations between both variables and rates of CO_2 emission, albeit with relatively shallow slopes.

The contrast between arable and non-arable land use is the next statistically significant effect. The fixed effect coefficient is negative (Table 3) and about five times its standard error. This implies that, other factors being equal, the rate of emission from arable soil is smaller than that from soil under the other land uses on the transect. Figure 10(a) shows this and also the box and whisker plot for rates of emission in the arable and other land use categories. Adding this term to the model increases $\breve{R}^2_{adj}$ to 0.33. The variance of the correlated random effect in the model is larger, but the range of spatial dependence also increases substantially, such that the variogram takes smaller values than for previous models for lags smaller than 3000 m (Figure 8).

The final significant effect is the soil association. The effect is negative, which implies larger emission rates over the Wicken association than the Cottenham association (see Figure 10b). The effect is larger than for the land use contrast, but of similar size relative to its standard error. Including the term raises $\breve{R}^2_{adj}$ to 0.51, and Figure 8 shows a substantial reduction in the variance of the spatially correlated random effect; the random variation in the model is dominated by the uncorrelated nugget term.

Figure 11 shows the histogram of residuals from the final model, and their summary statistics are in the bottom line of Table 2. Note

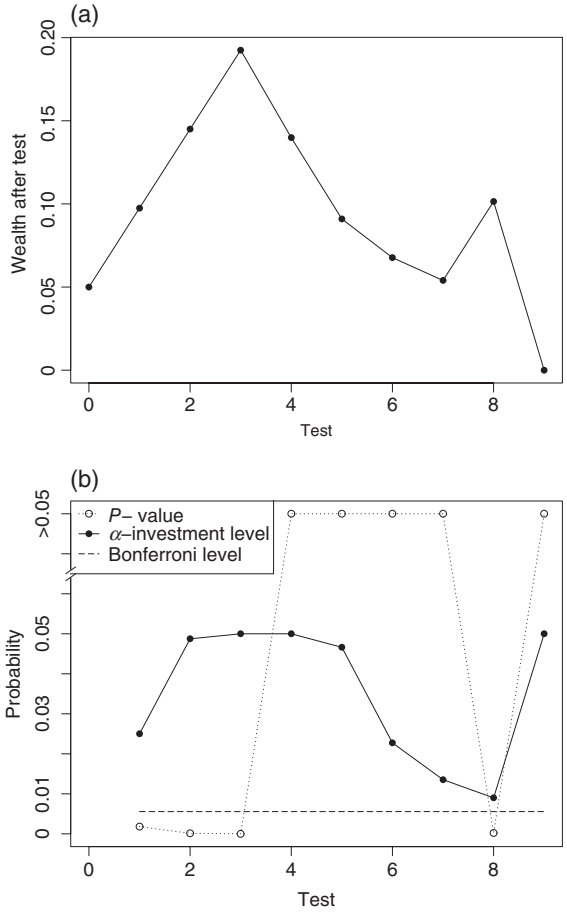

Figure 7 (a) Alpha-wealth after each successive test and (b) *P*-values and thresholds by different testing protocols for successive tests.

Table 3 Fixed effects coefficients and their standard errors and approximate adjusted R^2 for each term for which the null hypothesis was rejected

Added variable	Effect	Standard error	95% confidence interval	$\breve{R}^2_{adj}$
SOC	0.40	0.12	[0.16, 0.64]	0.14
Ammonium	0.16	0.04	[0.08, 0.24]	0.24
Arable against non-arable	−0.59	0.12	[−0.35, −0.83]	0.33
Cottenham against Wicken	−0.81	0.16	[−0.49, −1.13]	0.51

These statistics are for the model in which the listed term was added (so the model includes only that term and those listed above it in the table).
SOC, soil organic carbon.

that the residuals have a positive skew, but this is within the range [−1, 1] so can be regarded as moderate (Bulmer, 1979).

Discussion

The analysis of the data provides evidence for marked differences between soil associations with respect to the rate of CO_2 emission.

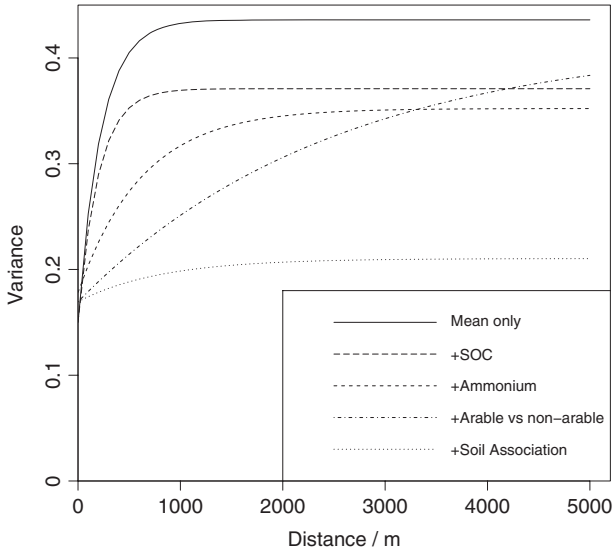

Figure 8 Maximum likelihood-estimated variograms for random components in model with a constant mean as the only fixed effect and terms added in order of the hypotheses.

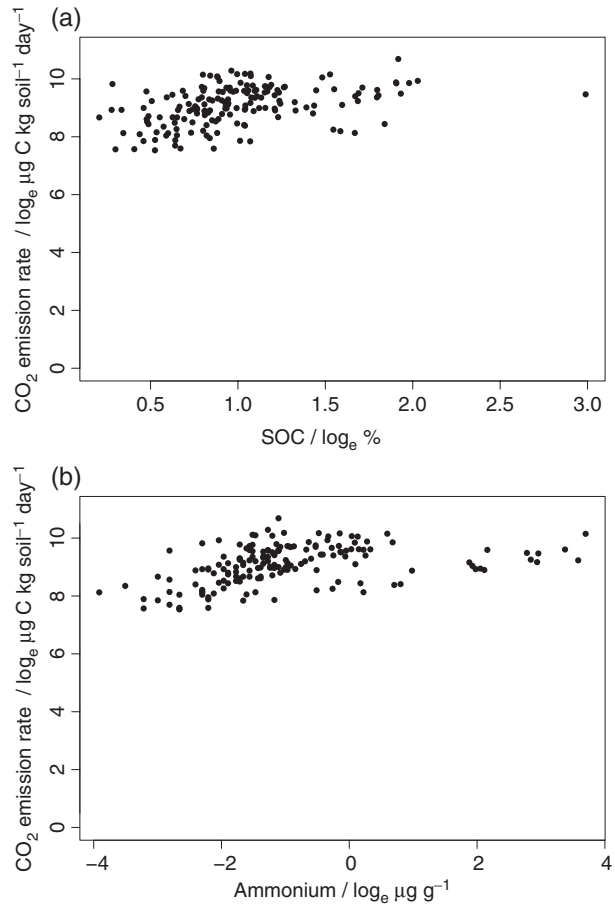

Figure 9 Plots of rates of CO_2 emission against (a) SOC and (b) soil ammonium concentration.

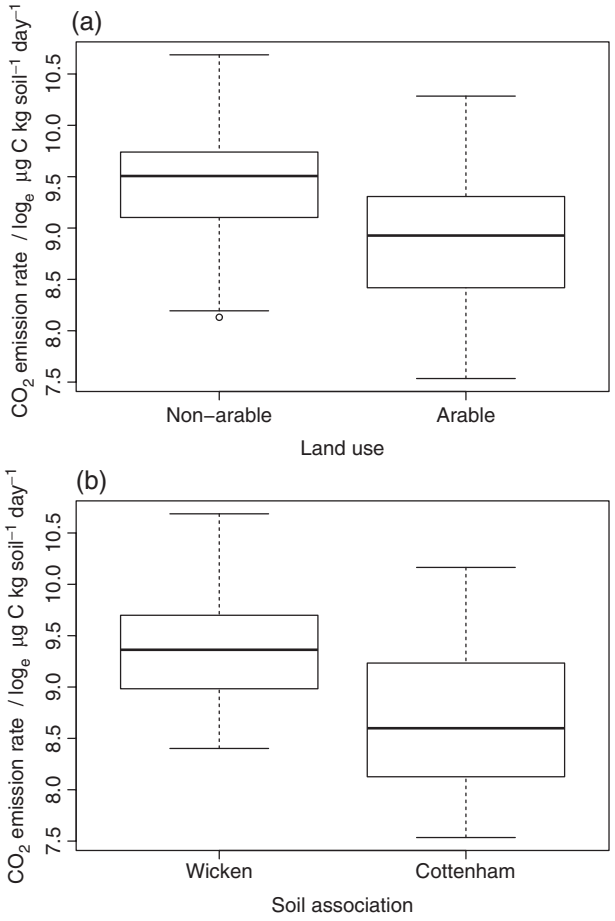

Figure 10 Box and whisker plots of rates of CO_2 emission for (a) arable and non-arable land use and (b) Cottenham and Wicken soil associations.

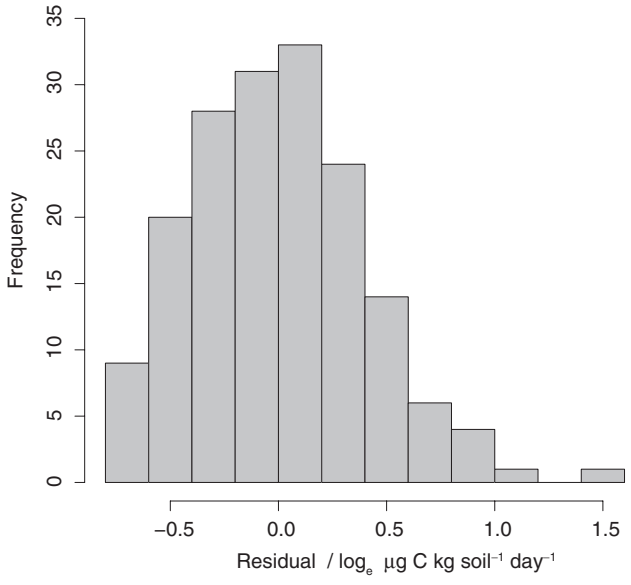

Figure 11 Histogram of residuals from model with all selected fixed effects.

Note that this effect appears in a model where SOC and ammonium content are already included, and a set of soil properties comprising bulk density, water-filled pore space, pH and the SOC to total N ratio have been considered but have not provided evidence to reject the null hypothesis. This suggests that there are differences between these soils with respect to important properties that affect microbial activity, gas transport or both. The two associations are over contrasting geological units (the Gault Clay and Lower Greensand), but both are overlaid with substantial Quaternary superficial deposits, although these do inherit some properties from the bedrock (King, 1969). It is notable that including this factor in the model removes almost all the spatially-dependent variation from the residuals, which suggests that any factors unaccounted for vary at short scales in the landscape relative to the sampling interval of approximately 30 m.

The soil under arable land use also shows statistically significantly smaller rates of CO_2 emission than does soil under other land uses. Note again that SOC is already in the model, so this factor, while also differing between soils under different land use, cannot be the explanatory factor. Much of the arable land on the transect had autumn-sown crops, and so had been cultivated within the previous 6 months. This disturbance, as well as long-term effects of arable land use, might account for the difference between soil under arable production and the other soil on the transect. Note that the difference between the other two land-use classes was not significant.

Although the effect of SOC is not very strong in the model, it is not surprising that it is a significant factor. It is interesting to note the effect of ammonium, which suggests that the link between the carbon and nitrogen cycles in soil at the mineralization stage might be expressed here.

In summary, the organic carbon content of the soil, differences between arable and non-arable land use and differences between soil associations over two Cretaceous bed rock units with contrasting lithology appear to account for about half the observed variation in rates of CO_2 emission on this transect; the remaining variation appears to be largely spatially independent at scales coarser than 30 m. Addition of the effect of soil association had the largest effect on the relative importance of the spatially correlated term in the random variation. Ammonium concentration in the soil appears as a significant factor in the model, possibly indicating linkages between two nutrient cycles.

The key point about this process is that, having tested a set of predetermined hypotheses with control of the marginal false discovery rate, we can have confidence that we have identified effects for which our data provide robust evidence. This would not be the case if we had used statistical testing *post hoc* to bolster our interpretation of effects discovered by visual inspection of plots or tables of means, or from some automated data mining procedure.

It might be argued that this is a counsel of perfection, and that the application of data mining and visualization to large datasets has considerable potential for the discovery of new knowledge about processes in soil. This process is sometimes called 'hypothesis discovery' (Payne, 2014). This might be useful, provided we remember that the hypothesis discovered remains to be tested on new

data from an appropriately designed experiment or sample survey. I am not aware of any studies in soil science where a data mining exercise has been followed by hypothesis testing on a genuinely independent set of data. It is more common to rely on a jackknifing of the dataset into separate subsets for modelling and validation, but, as Brus *et al.* (2011) point out, it is often not clear how to estimate unbiased statistics for model validation in these circumstances. The extent to which the two datasets can be regarded as independent will be severely limited by the original sample design. While there is some literature on jackknifing in the presence of dependence (e.g. Lele, 1991; Zhang *et al.*, 2012), the problem of forming robust jackknife subsets from spatially-dependent samples does not appear to have received adequate attention since the difficulties of the spatial case were observed by Cressie (1993).

In this study the α-investment procedure is used to test a sequence of hypotheses about controls on the rate of a process in the soil. Another setting in which a controlled procedure is needed to identify statistically significant relations between a soil property of interest and some subset of available covariates is digital soil mapping (DSM). In DSM the covariates may include remote sensor measurements and variables derived from a digital elevation model. Central to the success of the α-investment procedure is the preparation of a sequence of tests. As shown in the case study, and in Supporting Information, the selection of a sequence of tests does not depend only on the identification of plausible and interesting mechanistic effects. In a statistical model the value of a proposed covariate depends also on its correlation with covariates already in use. This is why the correlation between covariates was considered in the case study to avoid the use of more than one from any subset that is strongly mutually correlated, at least in early investment decisions. The process of assessing both the prior scientific plausibility of a mechanistic relation between a covariate and a target soil variable and the cross-correlations between available covariates might be more challenging in the DSM case than in studies that focus on soil processes. This is because the covariates in DSM are often proxies for underlying processes with tenuous or poorly understood mechanistic relations to the target soil variable. In this case, knowledge of the processes might be of limited value in the selection of a sequence of tests. This is a problem that requires further research,

The statistical literature on α-investment raises another question of interest. With the passage of time, the value of a dataset for testing prespecified hypotheses that are genuinely independent of past tests and analyses is likely to deteriorate. This is because reports of analyses, including visualizations of data, mean that the scientific community that might use those data learn the general patterns of joint variation that the data exhibit. Therefore, the hypotheses that are formulated are, to a greater or lesser extent, conditioned by those data that are then used for validation. This problem has been studied formally by Aharoni & Rosset (2014), who use the α-investing procedure of Foster & Stine (2008) to develop the concept of 'quality-preserving databases' (QPDs). In the management of a QPD all the hypotheses tested on that database are treated as a single stream, and it is recognized that maintaining the power of tests within this stream requires that additional independent data are added to the database over time. The α-investment approach is used to manage the procedure optimally. It is then possible to calculate the direct costs of maintaining the QPD (in terms of the field, laboratory and data management costs of obtaining new data), which can be passed directly to a data-user as the costs of maintaining the QPD after hypothesis tests that he or she has conducted on the data. It might be less straightforward to maintain a soil QPD just by adding new data, particularly if sampling is not carried out according to a probability design, and for soil properties subject to temporal change. It might also be difficult to convince funders of the need to add new data to a database simply to 'stand still' in terms of its scientific utility, especially in an era of limited resources in which the 'measure once, use many times' philosophy of data management has obvious attractions in many areas of environmental science (e.g. Knol, 2010; Lindstrom *et al.*, 2012). Nonetheless, it is important to be aware that the continued multiple uses of common public databases for scientific hypothesis testing is not unproblematic statistically, and α-investment is one approach that can be used to deal robustly with these problems while maintaining statistical power.

Conclusions

This research has shown how the α-investment procedure can be used to control the process of statistical hypothesis testing on a soil dataset, by the investigation of prespecified hypotheses and maintenance of a selected marginal false discovery rate. This approach can enable us to avoid problems entailed by multiple testing in a disciplined, hypothesis-driven approach to inference from data, which both maintains statistical power and guards against the perils of '*P*-hacking'. In this particular case study I found that the rate of CO_2 emission from incubated intact soil cores collected on a transect in eastern England was related to the organic carbon content of the soil and the measured concentration of ammonium nitrogen. The rates of emission also differed between arable and non-arable land use (smaller in arable soil), but not between woodland soil and that under grass. Rates of emission differed between soil formed over the Lower Greensand and that formed over the Gault Clay (larger in the latter), although no significant relations were found with some individual soil properties (bulk density, water-filled pore space fraction and organic carbon to total nitrogen ratio).

Supporting Information

The following supporting information is available in the online version of this article:

Table S1. Correlations between soil variables.

Table S2. Correlations between residuals from land use mean for each variable.

Figure S1. Values of each eigenvalue as a proportion of the trace.

Figure S2. Correlations of each variable with principal components 1–3

Acknowledgement

This paper is published with the permission of the Executive Director of the British Geological Survey (NERC).

References

Aciego Pietri, J.C. & Brookes, P.C. 2008. Relationships between soil pH and microbial properties in a UK arable soil. *Soil Biology & Biochemistry*, **40**, 1856–1861.

Aharoni, E. & Rosset, S. 2014. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society B*, **76**, 771–794.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.

Benjamini, Y. & Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Brus, D., Kempen, B. & Heuvelink, G.B.M. 2011. Sampling for validation of digital soil maps. *European Journal of Soil Science*, **62**, 394–407.

Bulmer, M.G. 1979. *Principles of Statistics*. Dover, Mineola, NY.

Cressie, N.A.C. 1993. *Statistics for Spatial Data*. John Wiley & Sons, New York.

Diggle, P.J. & Ribeiro, P.J. 2007. *Model-Based Geostatistics*. Springer, New York.

Foster, D.P. & Stine, R.A. 2008. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society B*, **70**, 429–444.

Haskard, K.A., Welham, S.J. & Lark, R.M. 2010. Spectral tempering to model non-stationary covariance of nitrous oxide emissions from soil using continuous or categorical explanatory variables at a landscape scale. *Geoderma*, **159**, 358–370.

IUSS Working Group WRB. 2006. *World Reference Base for Soil Resources 2006*. 2nd ed. World Soil Resources Rep. 103. FAO, Rome.

Karp, N.A., Heller, R., Yaacoby, S., White, J.K. & Benjamini, Y. 2016. Improving the identification of phenotypic abnormalities and sexual dimorphism in mice when studying rate event categorical characteristics. *Genetics*. doi: 10.1534/genetics.116.195388in press.

King, D.W. 1969. *Soils of the Luton and Bedford District*. Special Survey No 1. Soil Survey of England and Wales. Lawes Agricultural Trust, Harpenden.

Knol, O. 2010. Successful biodiversity monitoring in the Netherlands: the network ecological monitoring (NEM). In: *EnviroInfo 2010, Integration of Environmental Information in Europe* (eds G. von Claus & A.B. Cremers), pp. 457–461. Shaker Verlag, Herzogenrath.

Koenig, F., Slattery, J., Groves, T., Lang, T., Benjaminia, Y., Day, S. *et al.* 2015. Sharing clinical trial data on patient level: opportunities and challenges. *Biometrical Journal*, **57**, 8–26.

Lark, R.M. & Cullis, B.R. 2004. Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, **55**, 799–813.

Lark, R.M. & Milne, A.E. 2016. Boundary line analysis of the effect of water-filled pore space on nitrous oxide emission from cores of arable soil. *European Journal of Soil Science*, **67**, 148–159.

Lark, R.M. & Scheib, C. 2013. Land use and lead content in the soils of London. *Geoderma*, **209–210**, 65–74.

Lark, R.M., Bishop, T.F.A. & Webster, R. 2007. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. *Geoderma*, **138**, 65–78.

Lele, S. 1991. Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes. *Journal of the Royal Statistical Society, B*, **53**, 253–267.

Lindstrom, E., Gunn, J., Fischer, A., McCurdy, A. & Glover, L.K. 2012. *A Framework for Ocean Observing*. IOC/INF-1284. UNESCO, Paris. doi: 10.5270/OceanObs09-FOO.

Lin, D., Foster, D. & Ungar, L. 2011. VIF regression: a fast algorithm for large data. *Journal of the American Statistical Association*, **106**, 232–247.

Linn, D.M. & Doran, J.W. 1984. Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and nontilled soils. *Soil Science Society of America Journal*, **48**, 1267–1272.

Minasny, B., McBratney, A.B. & Bristow, K.L. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma*, **93**, 225–253.

Moyano, F.E., Vasilyeva, N., Bouckaert, L., Cook, F., Craine, J., Curiel Yuste, J., *et al.* 2012. The moisture response of soil heterotrophic respiration: interaction with soil properties. *Biogeosciences*, **9**, 1173–1182.

Moyano, F.E., Manzoni, S. & Chenu, C. 2013. Responses of soil heterotrophic respiration to moisture availability: an exploration of processes and models. *Soil Biology & Biochemistry*, **59**, 72–85.

Payne, P.R.O. 2014. In silico hypothesis discovery. In: *Translational Informatics* (eds P.R.O. Payne & P.J. Embi), pp. 129–151. Springer-Verlag, London.

R Core Team 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. [WWW document]. URL URL http://www.R-project.org/ [accessed on 5 January 2017].

Saiz, G., Green, C., Butterbach-Bahl, K., Kiese, R., Avitabile, V. & Farrell, E.P. 2006. Seasonal and spatial variability of soil respiration in four Sitka spruce stands. *Plant & Soil*, **287**, 161–176.

Tukey, J.W. 1991. The philosophy of multiple comparisons. *Statistical Science*, **6**, 100–116.

Turner, T.L., von Wettberg, E.J. & Nuzhdin, S.V. 2008. Genomic analysis of differentiation between soil types reveals candidate geners for local adaptation in *Arabidopsis lyrata*. *PLoS One*, **3**, e3183.

Verbeke, G. & Molenberghs, G. 2000. *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Wasserstein, R.L. & Lazar, N.A. 2016. The ASA's Statement on *p*-Values: Context, process, and purpose. *The American Statistician*, **70**, 129–133.

Webster, R. & Oliver, M.A. 2007. *Geostatistics for Environmental Scientists*, 2nd edn. John Wiley & Sons, Chichester.

Williams, D.E. 1949. A rapid manometric method for the determination of carbonate in soil. *Soil Science Society of America Proceedings*, **25**, 248–250.

Zhang, R., Peng, L. & Qi, Y. 2012. Jackknife-blockwise empirical likelihood methods under dependence. *Journal of Multivariate Analysis*, **104**, 56–72.