Supplement of Geosci. Commun., 8, 267–284, 2025
https://doi.org/10.5194/gc-8-267-2025-supplement

Supplement of

Communicating expected uncertainty in a geostatistical survey to support co-design with users of information

Christopher Chagumaira et al.

Correspondence to: Christopher Chagumaira (christopher.chagumaira@afbini.gov.uk, christopher.chagumaira@nottingham.ac.uk)

Supplement

S1 Theory

S1.1 Offset correlation

The expected correlation between the kriging predictions, $\tilde{Z}_1(\mathbf{x}_0)$, made from data collected on a square grid, of interval ζ, and predictions, $\tilde{Z}_2(\mathbf{x}_0)$, made from a second grid, a translation of the first grid by $\zeta/2$ in both directions is known as the offset correlation. The correlation of the two kriging predictions can be computed by:

$$\rho_{\tilde{Z}_1,\tilde{Z}_2} = \frac{\mathbf{C}_{\tilde{Z}_1,\tilde{Z}_2}(\mathbf{x}_0)}{\sqrt{\sigma^2_{\mathrm{K}_{\tilde{Z}_1}}\sigma^2_{\mathrm{K}_{\tilde{Z}_2}}}}, \tag{S1}$$

where $\mathbf{C}_{\tilde{Z}_1,\tilde{Z}_2}(\mathbf{x}_0)$ is the covariance $\tilde{Z}_1(\mathbf{x}_0)$ and $\tilde{Z}_2(\mathbf{x}_0)$. $\sigma^2_{\mathrm{K}_{\tilde{Z}_1}}$ and $\sigma^2_{\mathrm{K}_{\tilde{Z}_2}}$ are the kriging variances of the predictions from the first and second grid, respectively.

The offset correlation depends on $\mathbf{x}_0$, and is smallest at the location furthest from points on either grid. This minimum offset correlation is used to evaluate predictions from a grid spacing ζ. As the uncertainty in the map, attributable to sample density, decreases, the offset correlation increases. The denser the grid the more consistent the maps and the offset correlation will be 1 if the maps are identical and 0 if they are entirely unrelated to each other. The offset correlation is bounded on the interval [0,1], and ranges from zero (when the maps produced from the two grids are independent of each other (at a coarse spacing) and approach 1 as the grid becomes finer and the two maps become increasingly similar. Lark and Lapworth (2013) describes the offset correlation in greater detail.

S1.2 Prediction interval

Some unknown quantity at a location (e.g. soil pH or $\mathrm{Se}_{\mathrm{grain}}$) is characterised by a prediction distribution conditional on the data and statistical model. The kriging prediction is a weighted average of the data

$$\tilde{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda z(\mathbf{x}_i), \tag{S2}$$

where $z(\mathbf{x}_i)$ is the data and λ are the kriging weights (Webster and Oliver, 2007). The kriging variance, σ^2_{K} is defined as:

$$\sigma^2_{\mathrm{K}} = \mathrm{E}[\{Z(\mathbf{x}_0) - \tilde{Z}(\mathbf{x}_0)\}^2]. \tag{S3}$$

Cross-validation predictions of the statistical model need to be examined by exploratory analysis of the kriging error, $\varepsilon(\mathbf{x}_0)$, defined as $\varepsilon(\mathbf{x}_0) = \{z(\mathbf{x}_0) - \tilde{Z}(\mathbf{x}_o)\}$ to check if the assumption of the normality holds. The kriging predictor is unbiased and the mean of the errors is zero, and their standard deviation is equal to the kriging standard deviation, σ_{K}, from kriging. Based on this, a 95% prediction interval can be computed as:

$$\left[\tilde{Z}(\mathbf{x}_0) - 1.96\sigma_{\mathrm{K}}(\mathbf{x}_0), \tilde{Z}(\mathbf{x}_0) + 1.96\sigma_{\mathrm{K}}(\mathbf{x}_0)\right]. \tag{S4}$$

The prediction distribution may also be obtained on a block support–for example if predictions are required at the scale of a farm mean or a mean for an administrative region. The same approach holds to the derivation of a prediction interval.

S1.3 Conditional Probability

Consider a situation where the mean value of the variable across the region of interest is above the threshold below which some intervention is indicated. The value of spatial information in this setting is for identification of those locations where the intervention is required. In this case the probability that the predicted value of the variable at some location,$\mathbf{x}_0$, exceeds the threshold conditional on the true value's being below the threshold indicates the risk of a false negative conclusion at that location. We may expect this probability, which we denote by $p_{\mathrm{e|b}}$ to depend on the sampling density over some range of possible grid spacings. As the grid spacing becomes coarser so the the predicted value tends to the overall mean and $p_{\mathrm{e|b}}$ tends to 1.0.

At some location the true value of a property, z, might or might not indicate that an intervention is required. For purposes of this argument we assume that an intervention is required if $z \leq z_t$, a threshold value. We wish to compute the joint probability that a random location (a) requires the intervention (i.e. $z \leq z_t$), and (b) that the prediction, $\tilde{Z}$ indicates otherwise, (i.e. $\tilde{Z} > z_t$). If the kriging error, $z - \tilde{Z}$, were independent of z, then we might consider, assuming normal kriging errors and a known kriging variance, the probability that $\tilde{Z} > z_t$, given a value $Z = z$, $P\left(\tilde{Z} > z_t | z = Z\right)$, and then compute its expected value over the distribution of Z:

$$\int_{-\infty}^{-\infty} P\left(\tilde{Z} > z_t | z = Z\right) f(Z) \mathrm{d}\, Z, \tag{S5}$$

where $f(Z)$ denotes the PDF of Z. However, this independence does not hold. The kriging predictor, like any smoothing estimator, is conditionally biased in the sense that the error:

$$\varepsilon_z \;=\; z - \tilde{Z}, \tag{S6}$$

is likely to be positive for large z and negative for small z.

We can write the covariance of $z(\mathbf{x}_0)$ and $\varepsilon_z(\mathbf{x}_0)$ at some location $\mathbf{x}_0$ as

$$\mathrm{Cov}\left[z(\mathbf{x}_o), \varepsilon_z(\mathbf{x}_0)\right] \;=\; \mathrm{Var}\left[Z(\mathbf{x}_0)\right] - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{c}, \tag{S7}$$

where $\boldsymbol{\lambda}$ denotes the vector of n_n kriging weights for observations used to make the prediction, and $\mathbf{c}$ denotes the vector of covariances between each of these observations and $Z(\mathbf{x}_0)$. From Eq (S6)

$$\tilde{Z} = z - \varepsilon_z \therefore \tilde{Z} > z_t \Leftrightarrow z - \varepsilon_z > z_t \Leftrightarrow \varepsilon_z < z - z_t$$

Figure S1 shows a plot of error (positive or negative) against the true value of z. The line is the function

$$\varepsilon_z = z - z_t$$

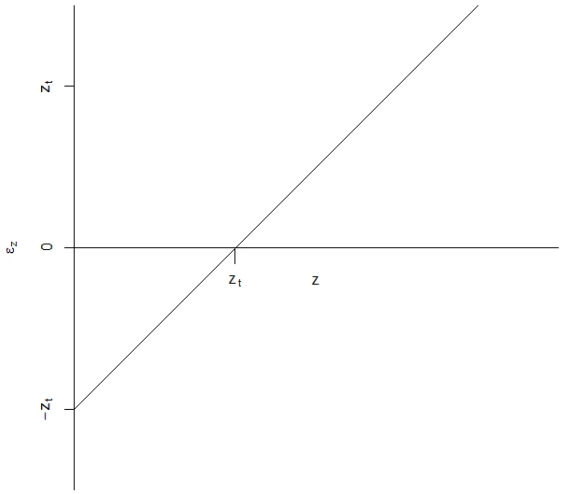

Figure S1. Plot of error (positive or negative) against the true value of z.

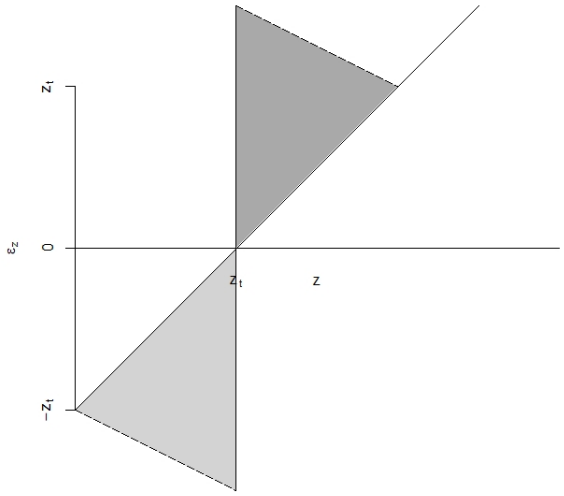

Figure S2. Plot of error against the true value of z.

In Figure S2 the light-grey shaded region, unbounded where the line is dashed, corresponds to where

$$z \leq z_t$$

and

$$\varepsilon_z < z - z_t,$$

i.e. to where the intervention is indicated if z is known without error, but $\tilde{Z} > z_t$. The other error condition is that $z > z_t$ and $\tilde{Z} \leq z_t$. This is represented by the dark grey space in Figure 2.

Table S1. Parameters of the joint distribution of Z and ε_z.

Mean of Z	Population mean of the variable
Variance of Z	*A priori* variance of the variable, i.e. $c_0 + c_1$.
Mean of ε_z	0, as kriging is unbiased
Variance of ε_z	Kriging variance
Covariance of ε_z and Z	$\mathrm{Var}\left[Z(\mathbf{x}_0)\right] - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{c}$

we may therefore, compute the joint probabilities that $z(\mathbf{x}_0) \leq z_t$ and $\varepsilon_z < z - z_t$ by

$$P\left(z(\mathbf{x}_0) < z_t, \varepsilon_z < z(\mathbf{x}_0) - z_t\right) = \int\int f_{z,\varepsilon_z}(z, \varepsilon_z)\mathrm{dz}\ \mathrm{d}\varepsilon_\mathrm{z}, \tag{S8}$$

where $f_{z,\varepsilon_z}(z,\varepsilon_z)$ is the joint normal distribution of $z(\mathbf{x}_0)$ and ε_z with parameters in Table S1 and the corresponding probability that $z(\mathbf{x}_0) < z_t$ is

$$P\left(z(\mathbf{x}_0) < z_t\right) = \int_{-\infty}^{z_t} f_z(Z)\mathrm{dz}, \tag{S9}$$

and the desired conditional probability

$$P\left(\varepsilon_z < z(\mathbf{x}_0) - z_t | z(\mathbf{x}_0) < z_t\right) = \frac{P\left(z(\mathbf{x}_0) < z_t, \varepsilon_z < z(\mathbf{x}_0) - z_t\right)}{P(z(\mathbf{x}_0) - z_t)}. \tag{S10}$$

As noted above, for brevity we denote the conditional probability defined in Equation (S10) informally by $p_{\mathrm{e|b}}$.

S1.4 Implicit loss function

The loss is a function of the error, if $\tilde{Z}$ is the predicted value and the true value is z, then the error is $\mathcal{L}(\tilde{Z} - z)$. If the value of z is equals to 0, then the error is equal to $\tilde{Z}$. The loss function is explained in greater detail by Journel (1984), Goovaerts (1997) and Lark and Knights (2015). Journel (1984) defined a general linear loss function as:

$$\begin{aligned}\mathcal{L}(\tilde{Z} - z) &= \alpha_1 |\tilde{Z} - z| \text{ if } \tilde{Z} < z \\ &= \alpha_2 |\tilde{Z} - z| \text{ if } \tilde{Z} \geq z.\end{aligned} \tag{S11}$$

The parameters α_1 and α_2 have positive real values. The coefficient α_2 is the loss per unit error of underestimation and α_1 is the loss per unit of error of overestimation. The slopes, α_1 and α_2 define the asymmetry of the loss function. The loss function can be symmetrical, i.e. penalizing overestimation and underestimation equally; or can be asymmetrical because over-and-underestimation have different consequences. The asymmetry of the loss function is the ratio of the loss per unit value by which a quantity is underestimated to the loss per unit value of an overestimation (Lark and Knights, 2015). The asymmetry, a, is obtained by

$$a = \frac{\alpha_2}{\alpha_1}, \tag{S12}$$

i.e., is independent of the absolute value of z. If the loss function depends only on the estimation error, then z can be set to zero, without loss of generality and the expected loss can be computed as a function of the error variance, and so of the sample size (Lark and Knights, 2015). Increasing sample size reduces the minimum expected loss in so far as it reduces the error variance. Therefore, the cost of obtaining n samples can be measured at which the marginal cost of an additional sample point is equal to the reduction in expected loss that single sample achieves (Goovaerts, 1997). However, it maybe difficult to define a loss function prior to making decisions about sampling. The losses may not be easy to quantify, e.g. social costs of failing to intervene, costs of unnecessary interventions, loss of confidence in the decision-making organisation. information users can be helped to reflect on possible loss functions through the implicit loss function. It is a loss function that makes a

specified sample size, n, a rational choice, given the marginal costs. That is to say, it is the loss function implied by a choice of $\bar{n}$, assuming this is rational. The implicit loss function is conditional on a logistic model (described in section below), that expresses the marginal costs of the sampling exercise and the conditional distribution of z as a function of effort (Lark and Knights, 2015) and is obtained by finding $\bar{\alpha}_1$ (given asymmetry), such that

$$\breve{\mathcal{L}}(\bar{n}-1|\bar{\alpha}_1,\bar{\alpha}_2,) - \breve{\mathcal{L}}(\bar{n}|\bar{\alpha}_1,\bar{\alpha}_2,) = \mathrm{C}(\bar{\mathrm{n}}) - \mathrm{C}(\bar{\mathrm{n}}-1), \tag{S13}$$

where $\bar{n}$ is the specified number of samples, $\mathrm{C(n)}$ is the function that returns the cost of n samples and is a vector of variogram parameters, so kriging variance is a contributor. The asymmetry can be set at different values, or inferred from other elicited opinions of the information user group (Lark and Knights, 2015). The expected loss can be minimised at a location given some prediction distribution of $\tilde{Z}$ for the variable of interest by specifying the value of variable corresponding to a given probability (P_0), i.e.,

$$\tilde{Z} = F^{-1}(P_0). \tag{S14}$$

Where, F^{-1} denotes the quantile of the prediction distribution for a probability P_0 obtained from

$$P_0 = \frac{\alpha_2}{\alpha_1+\alpha_2}, \tag{S15}$$

(Journel, 1984). Lark and Knights (2015) suggested that a information user group might consider an implicit loss function for different $\bar{n}$ as starting points in the elicitation of a sample size, or compare implicit loss functions for different projects given different partitions of a total budget between them. No attempt has been made to elicit opinions from information users on implicit loss function, so we tried it in this study.

S1.4.1 Logistical cost model

In this section we describe how the function defined in Lark and Knights (2015) to return the costs of n samples over an area A km^2, with a sample density of $r = N/A$ samples per km^2:

$$\mathrm{C(n)} = \omega + \mathrm{vAr} + \beta\mathrm{At_r}, \tag{S16}$$

where ω are the fixed costs, v cost of laboratory analysis per unit, and β the field costs per work day per team. The variable t_r is time taken to sample per km^2 at a density of r per km^2.

Consider a unit area containing the n sample locations. Following Beardwood et al. (1959), the expected distance to travel between sample points can be written as

$$\mathcal{D} \;=\; k\sqrt{n}. \tag{S17}$$

If we change the area in which the sample points are distributed to some value A, then the distance travelled is scaled by $\sqrt{A}$ and so

$$\mathcal{D}_A \;=\; k\sqrt{An}, \tag{S18}$$

and so we may write the distance travelled to sample n points per unit area as

$$\mathcal{D}_n \;=\; k\sqrt{\frac{n}{A}}. \tag{S19}$$

Assuming that the rate of travel is a random variable independent of sample density, we can therefore conclude that the time taken per unit area to travel between sample points is proportional to the square root of sample density

$$\mathcal{T}_t \;=\; \tau_1\sqrt{\frac{n}{A}}. \tag{S20}$$

Similarly, assuming that the sampling time is a random variable independent of sample density (time at a sample site), sampling time per unit area is proportional to sample density

$$\mathcal{T}_s \;=\; \tau_2\frac{n}{A}. \tag{S21}$$

Given these results, we may propose as a model for total sampling time per unit area

$$\mathcal{T}_o \;=\; \beta_1\sqrt{\frac{n}{A}} + \beta_2\frac{n}{A} + \beta_0 + T + \varepsilon, \tag{S22}$$

where β_0 is a constant to allow for fixed time requirements, T is a random effect of mean zero for between-team variation in sampling time and ε is a random effect of mean zero for the between-day (residual) variation.

S1.5 Fitting to data

In order to compute the variable t_r, we extracted the required data from the geostatical survey conducted in Malawi for the GeoNutrition project (Gashu et al., 2021). There were 8 teams that collected a total of 1812 sites of soil and crop samples were visited, this is described in detail by Gashu et al. (2021), Botoman et al. (2022) and Kumssa et al. (2022). For each team-day from the GeoNutrition survey of Malawi we have extracted the following:

- Number of points sampled.
- Mean time spent travelling per sample, removing the maximum inter-sample interval each day due to 'lunch break effect'. The units were in minutes.
- Mean time spent at a sample site. The units were in minutes.
- Length of the sampling day. The units were in minutes. The mean value is 331.
- The total area sampled that day. This is defined as the area of the sample domain which is in the Voronoi cell for the day's sample points. Unit were in square kilometres (km^2).

These variables are combined. We then compute the following:

- The total time spent sampling per unit area, $\mathcal{T}_o$ in Eq [S22] above, for each team–day.

Table S2. The anova table for the model

Effect	num DF	denom DF	F-ratio	P
Square root of Sampling density	1	294	347.21	<0.0001
Sampling Density	1	294	9.12	0.0027

Table S3. The estimated model coefficients

Coefficient	Estimate	SE
β_0	−0.007	0.51
β_1	4.08	4.89
β_2	33.6	11.12

- Sample density, $\frac{n}{A}$, for each team–day.
- The square root of sample density.

We can then fit a linear mixed model for $\mathcal{T}_o$ in which the fixed effects are $\sqrt{\frac{n}{A}}$ and $\frac{n}{A}$ and in which team is a random effect. The anova table for the model is as follows in Table S2

This shows significant effects of both powers of sample density.
The estimated model coefficients are shown in Table S3

The data and fitted model are shown on Figure S3.

S1.6 Worked example

Rumphi district: Area 4,769 km^2

*Given total area of Rumphi and assuming a mean sampling day of 331 minutes (Table S4)

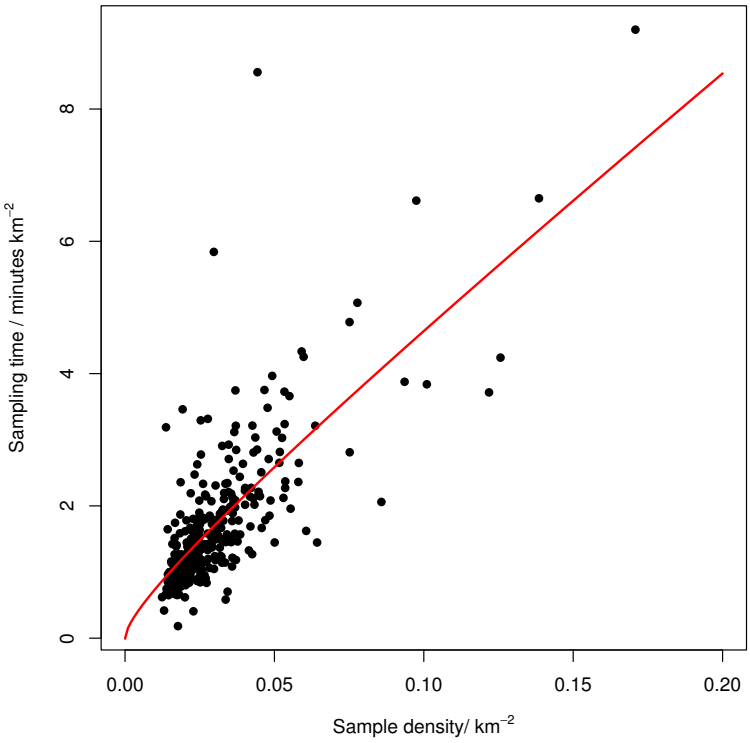

Figure S3. Scatter plot showing the data and fitted model.

Table S4. Worked example for Rumphi district

Sample size	Sample Density /km^{-2}	Predicted sample effort /min km^{-2}	Total sample effort / team–days*
200	0.0419	2.238	35.6
500	0.1048	4.837	76.9
1000	0.2097	8.907	141.6

S2 Test methods: charts presented to the stakeholders

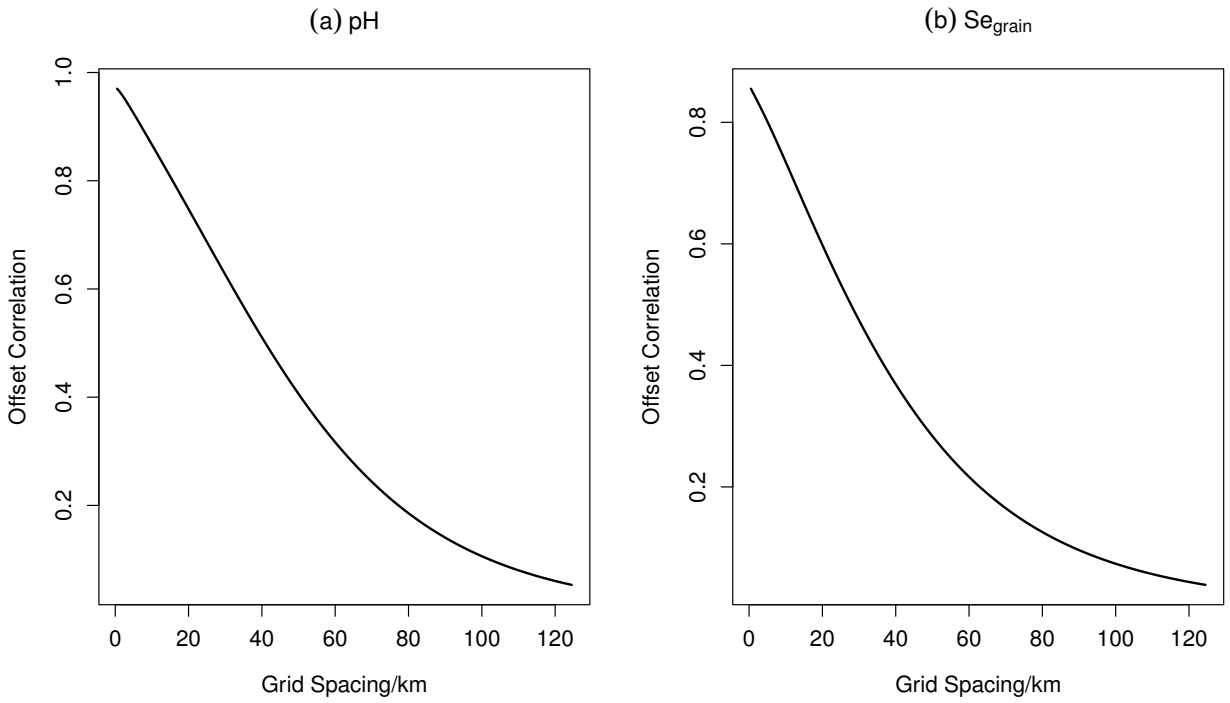

Figure S4. A plot of offset correlation and grid spacing for (a) soil pH and (b) $\mathrm{Se}_{\mathrm{grain}}$ in Malawi.

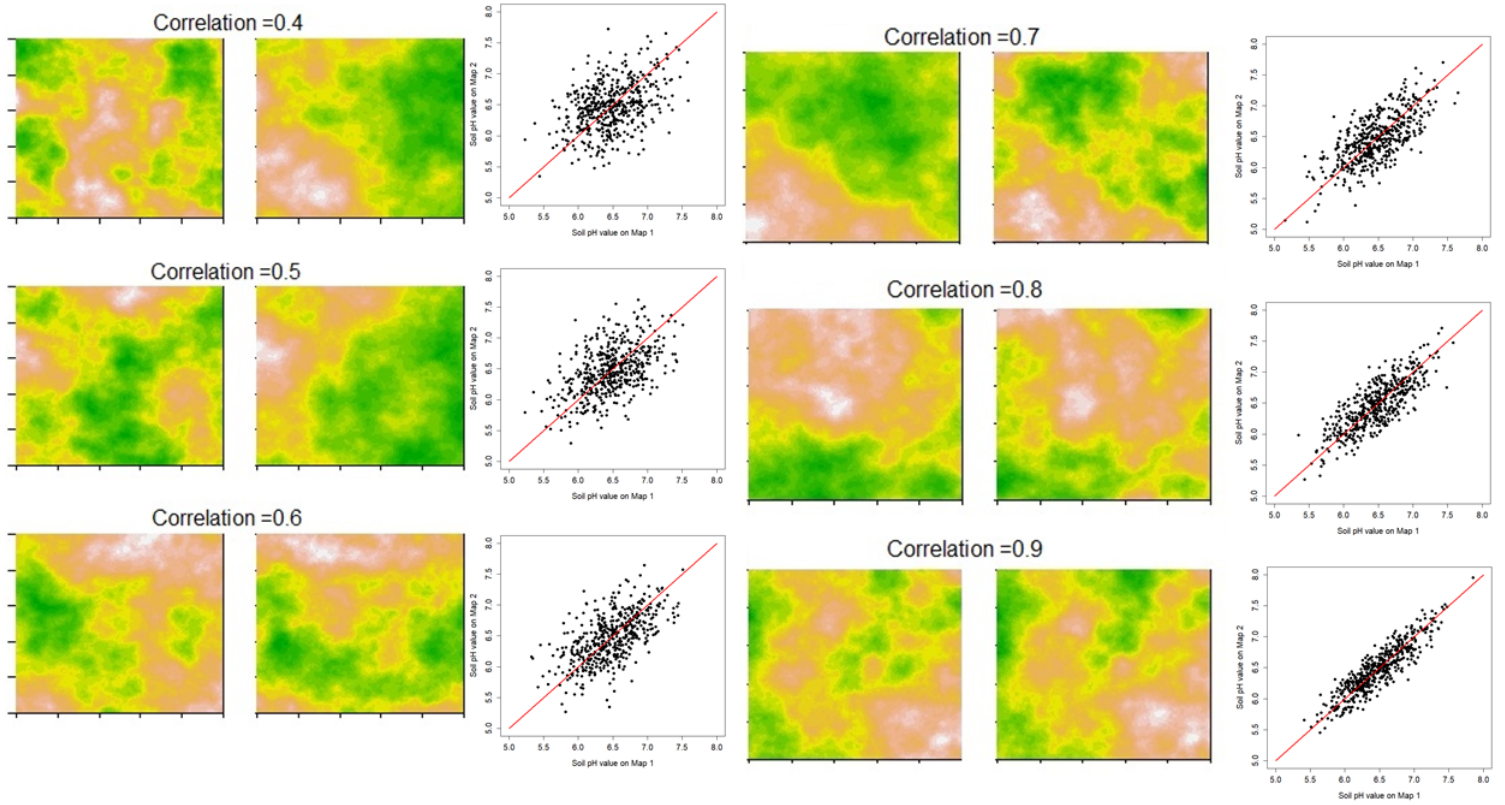

Figure S5. The pairs of example maps of soil pH, each pair being based on a different grid spacing, with a different offset correlation (Q1) and corresponding scatter plots that illustrated the strength of the correlation.

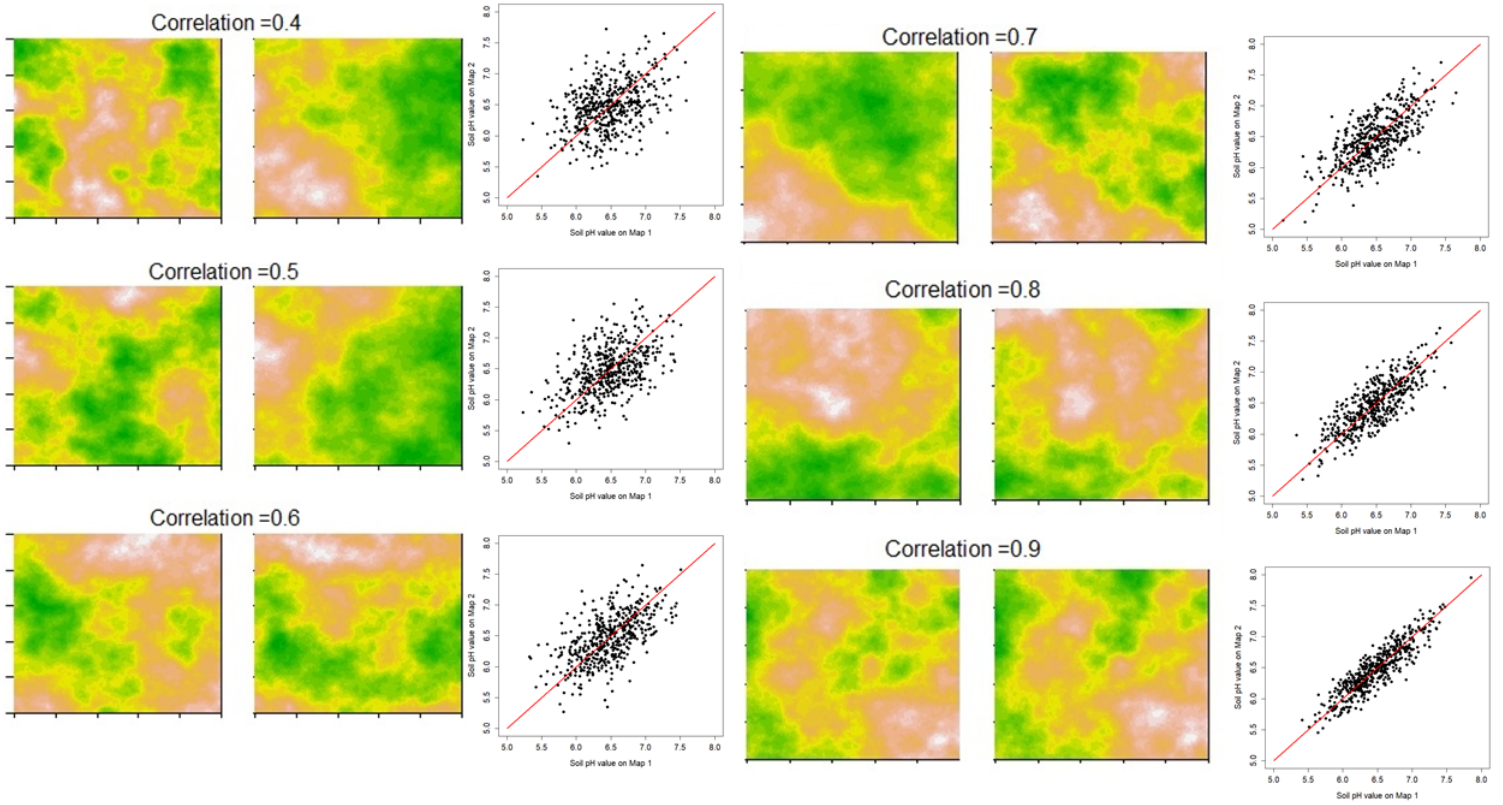

Figure S6. The pairs of example maps of Se concentration in grain, each pair being based on a different grid spacing, with a different offset correlation (Q1) and corresponding scatter plots that illustrated the strength of the correlation.

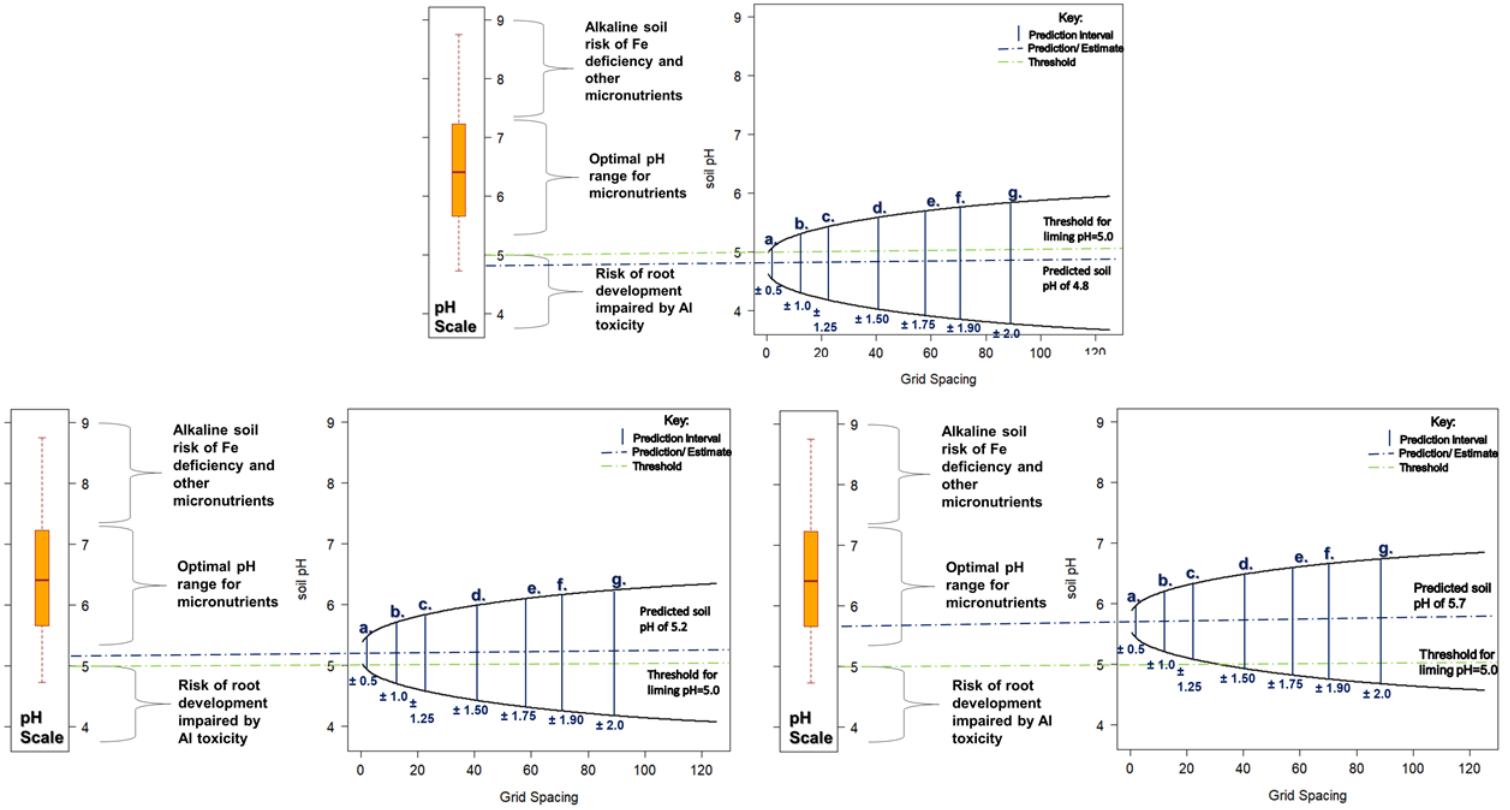

Figure S7. Chart consisting of box plot of the distribution of the soil pH, a graph of the lower and upper prediction intervals (Q2) for the prediction for grid spacings from 0 to 120 km. With a blue line corresponding to the prediction and the green one for the threshold value.

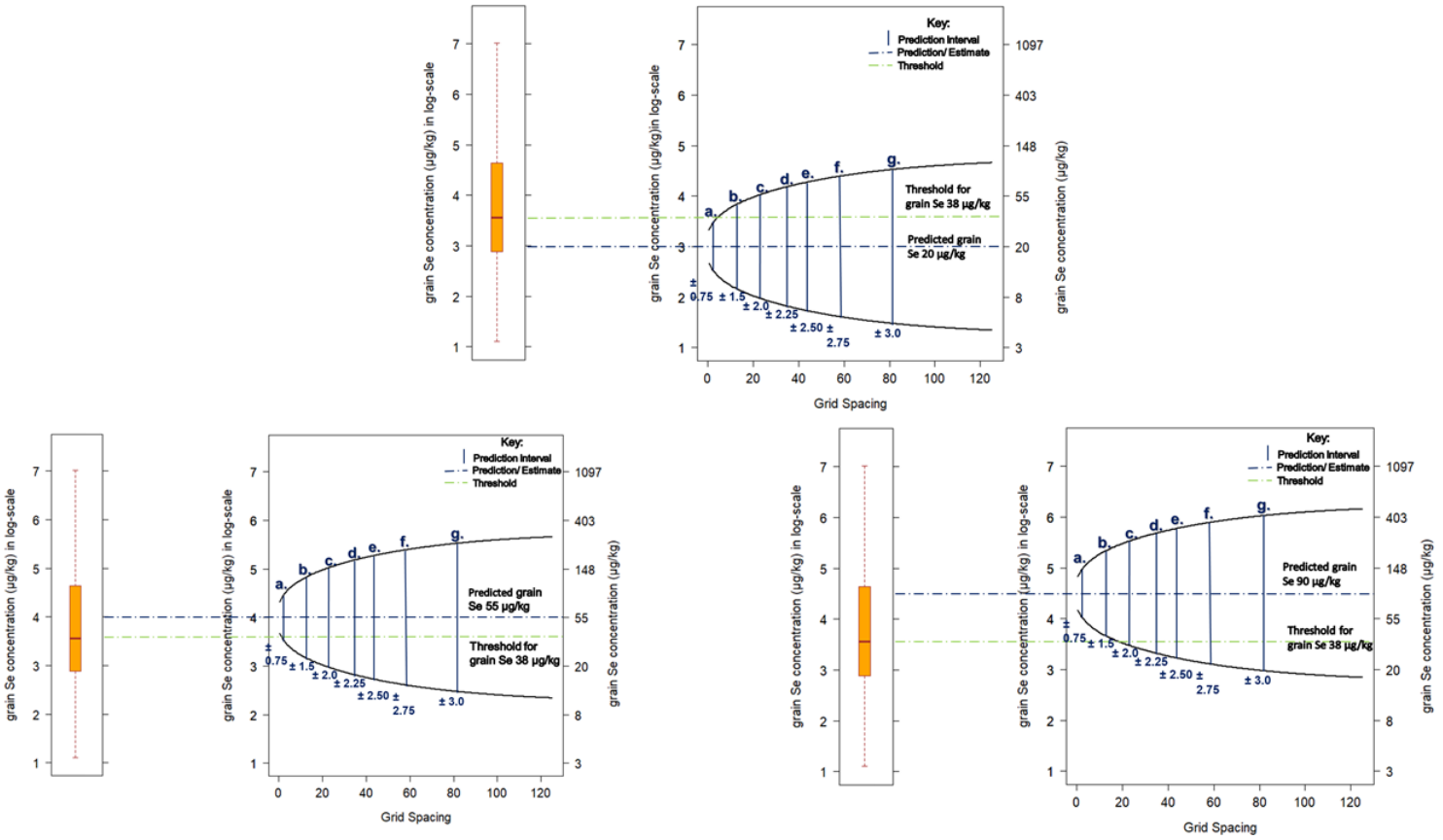

Figure S8. Chart consisting of box plot of the distribution of the Se concentration in grain, a graph of the lower and upper prediction intervals (Q2) for the prediction for grid spacings from 0 to 120 km. With a blue line corresponding to the prediction and the green one for the threshold value.

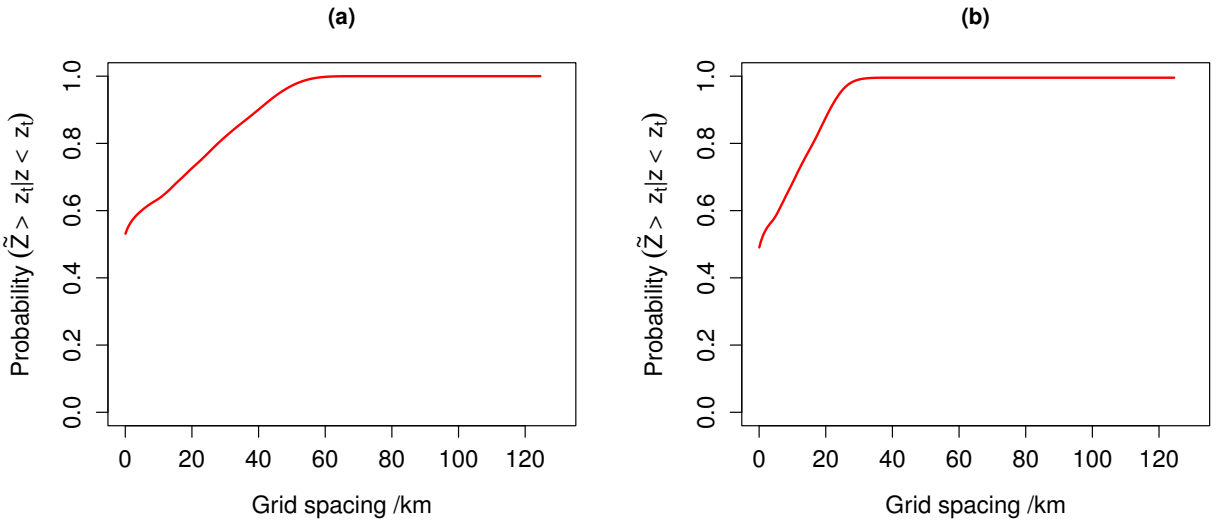

Figure S9. Graph showing the probability (Q3), given that an intervention is required at $\mathbf{x}_o$ that, due to error in prediction, the mapped variable does not show this. z_t is the threshold of interest. (a) is for soil pH and (b) for $\mathrm{Se}_{\mathrm{grain}}$ concentration.

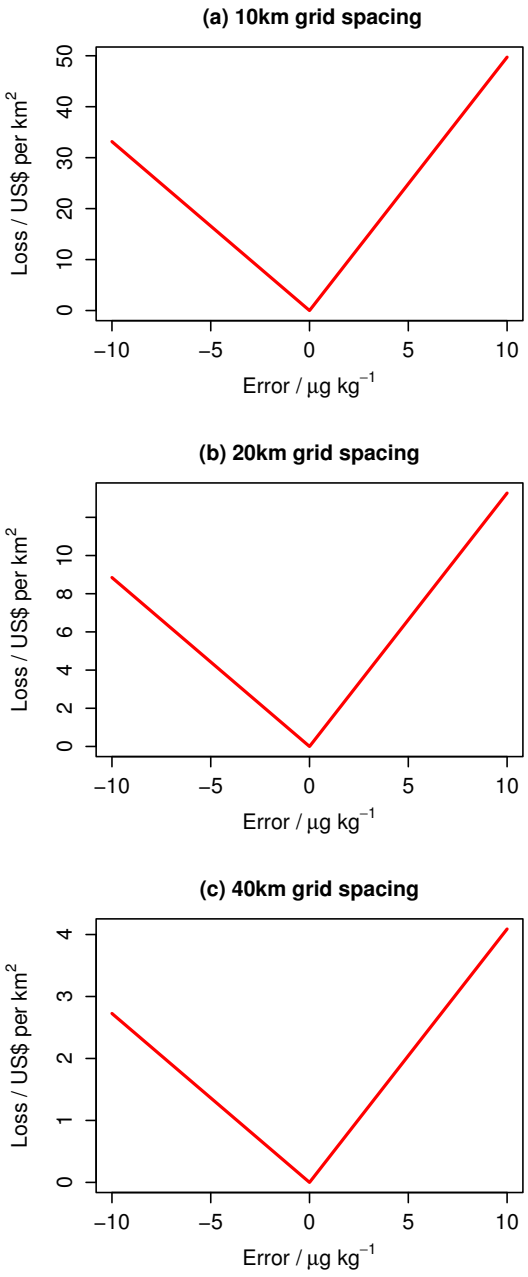

Figure S10. Three specified implicit loss functions (Q4) for predictions Se concentration in grain an administrative district in Malawi presented to the participants.

S3 Composition of Participants

Figure S11 shows compositions of participants by (a) location, (b) level of mathematical education, (c) level of use of statistics and (d) professional group

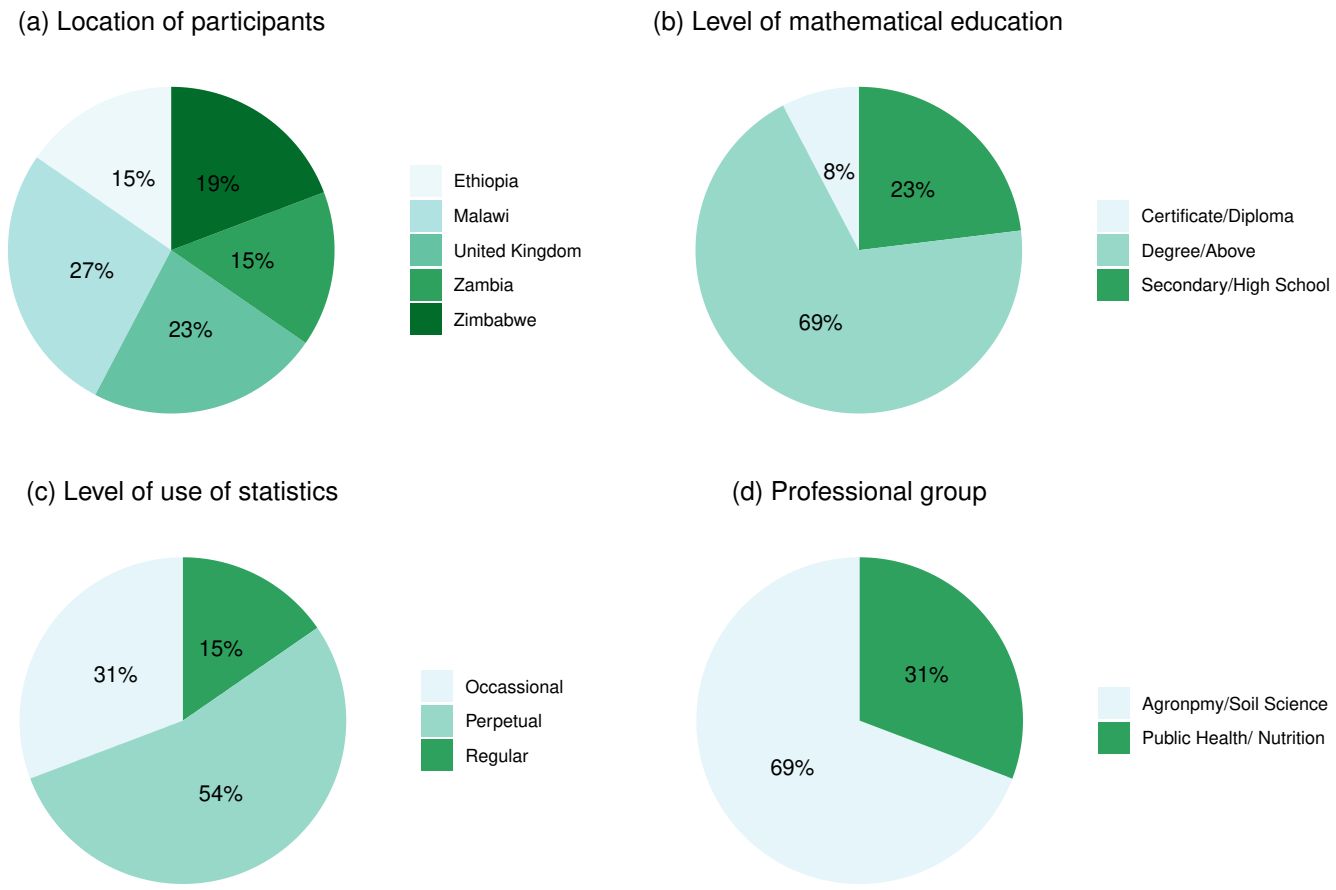

Figure S11. Pie charts showing the percentage of participants by (a) location, (b) level of mathematical education, (c) level of use of statistics and (d) professional group.

S4 Contingency tables

In this section we described how a contingency table can be partitioned to evaluate whether there are differences in the responses of the participants based on (i) variable used in the test method, (ii) professional group and (iii) by frequency of use of statistics. In Table S5, we illustrate how the contingency table can be partitioned. The table can be partition into components corresponding to pooled table and subtables of the full table.

Table S5. An illustration of how the log-likelihood ratio was used to partition full table into subtables and pooled tables.

Deviance = L_f
Response = $O_{i,j}$
degrees of freedom = $DF_F = (2-1)\times(12-1) = 11$

Full table

	Soil pH						Se_{grain} concentration					
	Agronomy or soil science (AGS)			Public health or nutrition (PHN)			Agronomy or soil science (AGS)			Public health or nutrition (PHN)		
Response	**Occasionally**	**Regularly**	**Perpetually**	**Occasionally**	**Regularly**	**Perpetually**	**Occasionally**	**Regularly**	**Perpetually**	**Occasionally**	**Regularly**	**Perpetually**
Spacing 1	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,4}$	$O_{1,5}$	$O_{1,6}$	$O_{1,7}$	$O_{1,8}$	$O_{1,9}$	$O_{1,10}$	$O_{1,11}$	$O_{1,12}$
Spacing 2	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,4}$	$O_{2,5}$	$O_{2,6}$	$O_{2,7}$	$O_{2,8}$	$O_{2,9}$	$O_{2,10}$	$O_{2,11}$	$O_{2,12}$

Pooled table 1

Pooled table (Professional groups pooled within variable):
Deviance = L_{p1}
degrees of freedom = $DF_{P1} = (2-1)\times(2-1) = 1$

Response	Soil pH	Se_{grain} concentration
Spacing 1	$O_{1,1}+O_{1,2}+O_{1,3}+O_{1,7}+O_{1,8}+O_{1,9}$	$O_{1,4}+O_{1,5}+O_{1,6}+O_{1,10}+O_{1,11}+O_{1,12}$
Spacing 2	$O_{2,1}+O_{2,2}+O_{2,3}+O_{2,7}+O_{2,8}+O_{2,9}$	$O_{2,4}+O_{2,5}+O_{2,6}+O_{2,10}+O_{2,11}+O_{2,12}$

Subtable 1

Subtable 1 (Soil pH):
Deviance= L_{s1}
degrees of freedom = $DF_{S1} = (2-1)\times(6-1) = 5$

	Soil pH					
	Agronomy or soil science (AGS)			Public health or nutrition (PHN)		
Response	**Occasionally**	**Regularly**	**Perpetually**	**Occasionally**	**Regularly**	**Perpetually**
Spacing 1	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,4}$	$O_{1,5}$	$O_{1,6}$
Spacing 2	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,4}$	$O_{2,5}$	$O_{2,6}$

Subtable 2

Subtable 2 (Se_{grain} concentration):
Deviance = L_{s1}
degrees of freedom= $DF_{S2} = (2-1)\times(6-1) = 5$

	Se_{grain} concentration					
	Agronomy or soil science (AGS)			Public health or nutrition (PHN)		
Response	**Occasionally**	**Regularly**	**Perpetually**	**Occasionally**	**Regularly**	**Perpetually**
Spacing 1	$O_{1,7}$	$O_{1,8}$	$O_{1,9}$	$O_{1,10}$	$O_{1,11}$	$O_{1,12}$
Spacing 2	$O_{2,7}$	$O_{2,8}$	$O_{2,9}$	$O_{2,10}$	$O_{2,11}$	$O_{2,12}$

Deviance partition:
$L_f = L_{p1}+L_{s1}+L_{s2}$

Degrees of freedom partition:
$DF_f = DF_{p1}+DF_{s1}+DF_{s2}$

The full table in Table S5, was partitioned into components corresponding to subtables for soil pH (Subtable 1 in Table S5), and $\mathrm{Se}_{\mathrm{grain}}$ concentration (Subtable 2 in Table S5). Then the pooled table completes the partition. The degrees of freedom and deviances for the three table sum to the degrees of freedom and deviance of the full table. Using the contingency table, we could conclude if there are differences in responses for the two variables. The full table in can further be partitioned, in a similar way, by the background of the respondents i.e., professional group and frequency of use of statistics.

The full contingency table for Q1, for offset correlation, is presented as Table S6, in the appendix. The table shows how many individuals selected the given responses for offset correlation. This table is according to variable used (soil pH vs. $\mathrm{Se}_{\mathrm{grain}}$), professional group and frequency of use of statistics. Table S7 shows how many individuals selected a given response to Q1, for offset correlation, when columns are pooled within variable used, soil pH or $\mathrm{Se}_{\mathrm{grain}}$ concentration. Table S8 shows the pooled counts of the responses for Q1.

Table S6. The full contingency table for Q1 for offset correlation, showing how many respondents selected the given responses for offset correlation. The table is according to variable used, professional group and frequency of use of statistics in job role. The figures in parentheses are the expected numbers, $e_{i,j}$ a product of row and column totals dived by the total number of responses. The acronyms represent the professional groups (AGS–agronomist or soil scientist; PHN– public health or nutrition specialists), and frequency of use of statistics in job role (Pep– perpetual use of statistics; Occ–Occasional use of statistics and Reg– regular use of statistics).

Response	soil pH						$\mathrm{Se}_{\mathrm{grain}}$					
	AGS			PHN			AGS			PHN		
	Pep	Occ	Reg	Pep	Occ	Reg	Pep	Occ	Reg	Pep	Occ	Reg
Offset=0.4	0(0.23)	0(0.31)	1(0.85)	0(0.08)	0(0.31)	0(0.23)	0(0.23)	2(0.31)	1(0.31)	0(0.08)	0(0.31)	0(0.23)
Offset=0.5	0(0.17)	1(0.23)	0(0.63)	0(0.06)	0(0.23)	1(0.17)	0(0.17)	0(0.23)	1(0.23)	0(0.06)	0(0.23)	0(0.17)
Offset=0.6	1(0.40)	0(0.54)	1(1.48)	0(0.13)	1(0.54)	0(0.40)	1(0.40)	0(0.54)	0(0.54)	0(0.13)	1(0.54)	2(0.40)
Offset=0.7	1(0.92)	2(1.23)	2(3.38)	0(0.31)	3(1.23)	2(0.92)	1(0.92)	1(1.23)	3(1.23)	0(0.31)	1(1.23)	0(0.92)
Offset=0.8	0(0.87)	0(1.15)	5(3.17)	1(0.29)	0(1.15)	0(0.87)	0(0.87)	1(1.15)	5(1.15)	1(0.29)	1(1.15)	1(0.87)
Offset=0.9	1(0.40)	1(0.54)	2(1.48)	0(0.13)	0(0.54)	0(0.40)	1(0.40)	0(0.54)	1(0.54)	0(0.13)	1(0.54)	0(0.40)

Table S7. A subtable showing how many individuals selected a given response to Q1, for offset correlation, when columns are pooled within variable used (soil pH vs. $\mathrm{Se}_{\mathrm{grain}}$ concentration).

Response	soil pH	$\mathrm{Se}_{\mathrm{grain}}$
Offset=0.4	1	3
Offset=0.5	2	1
Offset=0.6	3	4
Offset=0.7	10	6
Offset=0.8	6	9
Offset=0.9	4	3

Table S8. Pooled responses given to the question on offset correlation.

Response	Pooled counts
Offset=0.4	4
Offset=0.5	3
Offset=0.6	7
Offset=0.7	16
Offset=0.8	15
Offset=0.9	7

References

Beardwood, J., Halton, J. H., and Hammersley, J. M.: The shortest path through many points, in: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 55, pp. 299–327, Cambridge University Press, 1959.

Botoman, L., Chagumaira, C., Mossa, A. W., Amede, T., Ander, E. L., Bailey, E. H., Chimungu, J. G., Gameda, S., Gashu, D., Haefele, S. M., Joy, E. J. M., Kumssa, D. B., Ligowe, I. S., Mcgrath, S. P., Milne, A. E., Munthali, M., Towett, E., Walsh, M. G., Wilson, L., Young, S. D., Broadley, M. R., Lark, R. M., and Nalivata, P. C.: Soil and landscape factors influence geospatial variation in maize grain zinc concentration in Malawi, Scientific Reports, https://doi.org/10.1038/s41598-022-12014-w, 2022.

Gashu, D., Nalivata, P. C., Amede, T., Ander, E. L., Bailey, E. H., Botoman, L., Chagumaira, C., Gameda, S., Haefele, S. M., Hailu, K., Joy, E. J. M., Kalimbira, A. A., Kumssa, D. B., Lark, R. M., Ligowe, I. S., McGrath, S. P., Milne, A. E., Mossa, A. W., Munthali, M., Towett, E. K., Walsh, M. G., Wilson, L., Young, S. D., and Broadley, M. R.: The nutritional quality of cereals varies geospatially in Ethiopia and Malawi, Nature, 594, 71–76, https://doi.org/10.1038/s41586-021-03559-3, 2021.

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

Journel, A.: Mad and conditional quantile estimators, in: Geostatistics for natural resources characterization, pp. 261–270, Springer, 1984.

Kumssa, D. B., Mossa, A. W., Amede, T., Ander, E. L., Bailey, E. H., Botoman, L., Chagumaira, C., Chimungu, J. G., Davis, K., Gameda, S., Haefele, S., Hailu, K., Joy, E. J. M., Lark, R. M., Ligowe, I. S., Muleya, P., McGrath, S. P., Milne, A. E., Munthali, M., Towett, E., Walsh, M. G., Wilson, L., Young, S. D., Haji, I. R., Broadley, M. R., Gashu, D., and Nalivata, P. C.: Cereal grain mineral micronutrient and soil chemistry data from GeoNutrition surveys in Ethiopia and Malawi [Dataset], https://doi.org/10.6084/m9.figshare.15911973.v1, 2022.

Lark, R. M. and Knights, K. V.: The implicit loss function for errors in soil information, Geoderma, 251-252, 24–32, https://doi.org/10.1016/j.geoderma.2015.03.014, 2015.

Lark, R. M. and Lapworth, D. J.: The offset correlation, a novel quality measure for planning geochemical surveys of the soil by kriging, Geoderma, 197-198, 27–35, https://doi.org/10.1016/j.geoderma.2012.12.020, 2013.

Webster, R. and Oliver, M. A.: Geostatistics for Natural Environmental Scientists, John Wiley & Sons Chichester, 2nd edn., https://doi.org/10.2136/vzj2002.0321, 2007.