Supplement

1 Theory

1.1 Conditional Probabilities

Calculating the joint probability that a location requires an intervention, and that the kriged estimate does not indicate this

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:

$$\int_{-\infty}^{-\infty} P\left(\tilde{Z} > z_t | z = Z\right) f(Z) \mathrm{d}\, Z, \tag{1}$$

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:

$$\varepsilon_z \; = \; z - \tilde{Z}, \tag{2}$$

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

$$\mathrm{Cov}\left[z(\mathbf{x}_o), \varepsilon_z(\mathbf{x}_0)\right] \; = \; \mathrm{Var}\left[Z(\mathbf{x}_0)\right] - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{c}, \tag{3}$$

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 (2)

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

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

$$\varepsilon_z = z - z_t$$

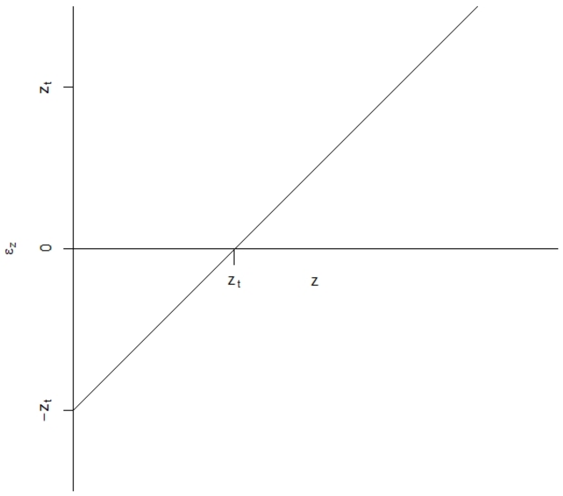


Figure S1. Plot of error (positive or negative) against the true value of z.

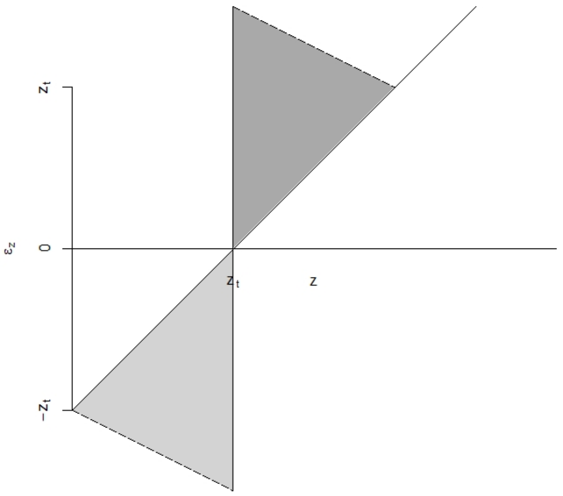


Figure S2. Plot of error against the true value of z.

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

$$P\left(z(\mathbf{x}_0) < z_t, \varepsilon_z < z(\mathbf{x}_0) - z_t\right) = \int\int f_{z,\varepsilon_z}(z, \varepsilon_z) \mathrm{dz}\ \mathrm{d}\varepsilon_{\mathrm{z}}, \tag{4}$$

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

$$P\left(z(\mathbf{x}_0) < z_t\right) = \int_{-\infty}^{z_t} f_z(Z) \mathrm{dz}, \tag{5}$$

and the desired conditional probability

$$P\left(\varepsilon_z < z(\mathbf{x}_0) - z_t | z(\mathbf{x}_0) < z_t\right) = \frac{P\left(z(\mathbf{x}_0) < z_t, \varepsilon_z < z(\mathbf{x}_0) - z_t\right)}{P(z(\mathbf{x}_0) - z_t)}. \tag{6}$$

1.2 Implicit loss function

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:

$$\mathrm{C(n)} = \omega + \mathrm{vAr} + \beta \mathrm{At_r}, \tag{7}$$

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

$$\mathcal{D} \;=\; k\sqrt{n}. \tag{8}$$

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

$$\mathcal{D}_A \;=\; k\sqrt{An}, \tag{9}$$

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

$$\mathcal{D}_n \;=\; k\sqrt{\frac{n}{A}}. \tag{10}$$

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

$$\mathcal{T}_t \;=\; \tau_1\sqrt{\frac{n}{A}}. \tag{11}$$

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

$$\mathcal{T}_s \;=\; \tau_2\frac{n}{A}. \tag{12}$$

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

$$\mathcal{T}_o \;=\; \beta_1\sqrt{\frac{n}{A}} + \beta_2\frac{n}{A} + \beta_0 + T + \varepsilon, \tag{13}$$

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.

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 [13] above, for each team–day.
- 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

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

This shows significant effects of both powers of sample density.

The estimated model coefficients are as follows

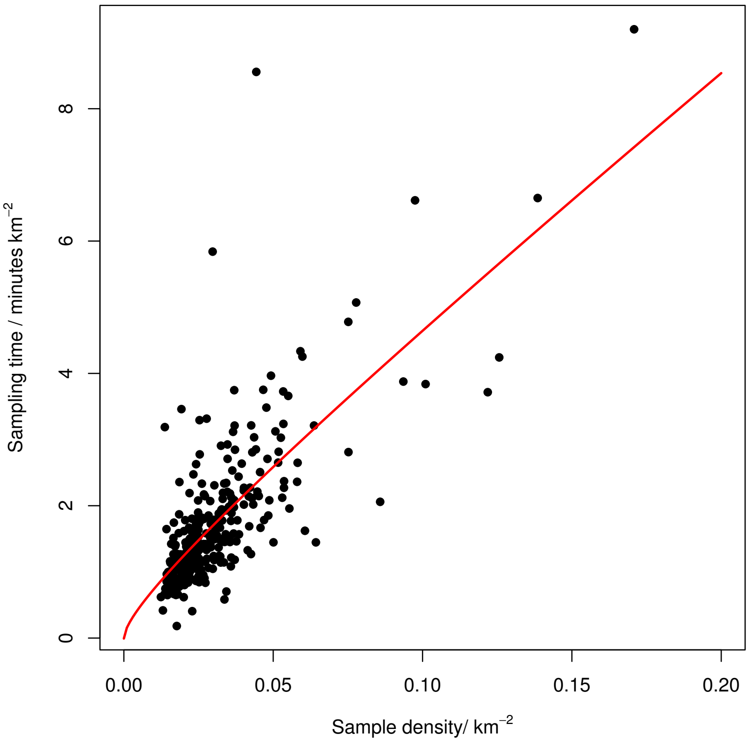


Figure S3. Scatter plot showing the data and fitted model.

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

The data and fitted model are shown bon Figure S3.

Worked example

Rumphi district: Area 4,769 km^2

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

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

2 Test methods: charts presented to the stakeholders

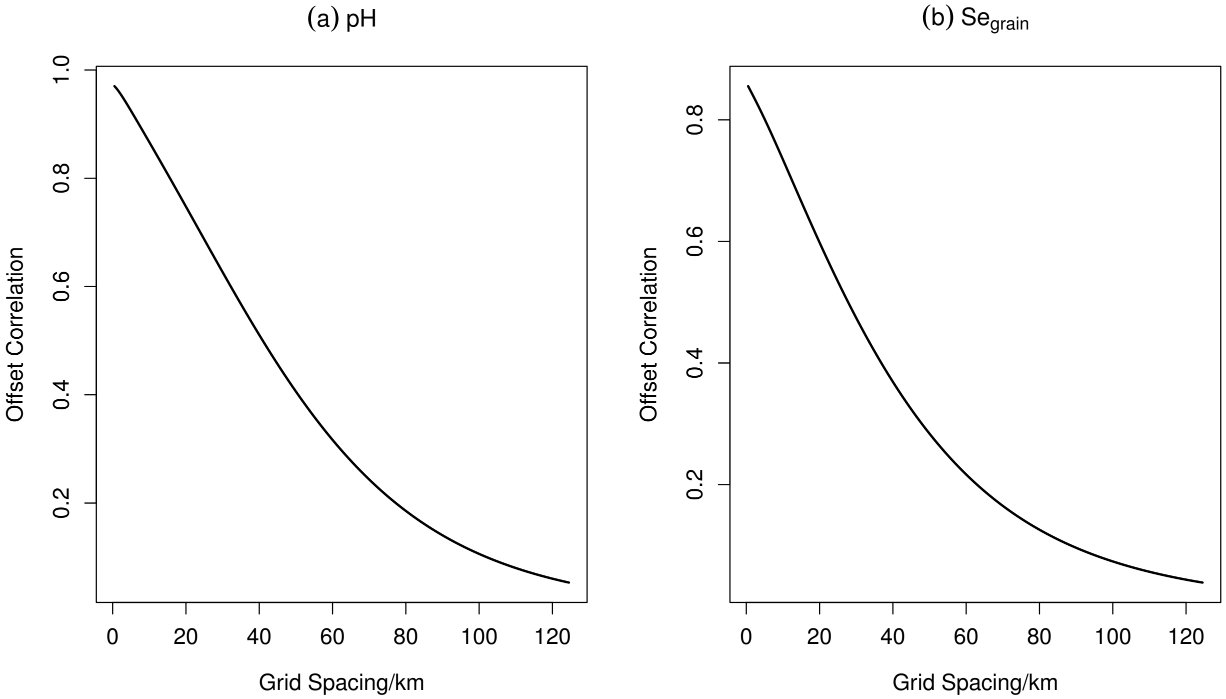


Figure S4. A plot of offset correlation and grid spacing for (a) soil pH and (b) $\mathrm{Se}_{\mathrm{grain}}$ in Malawi.

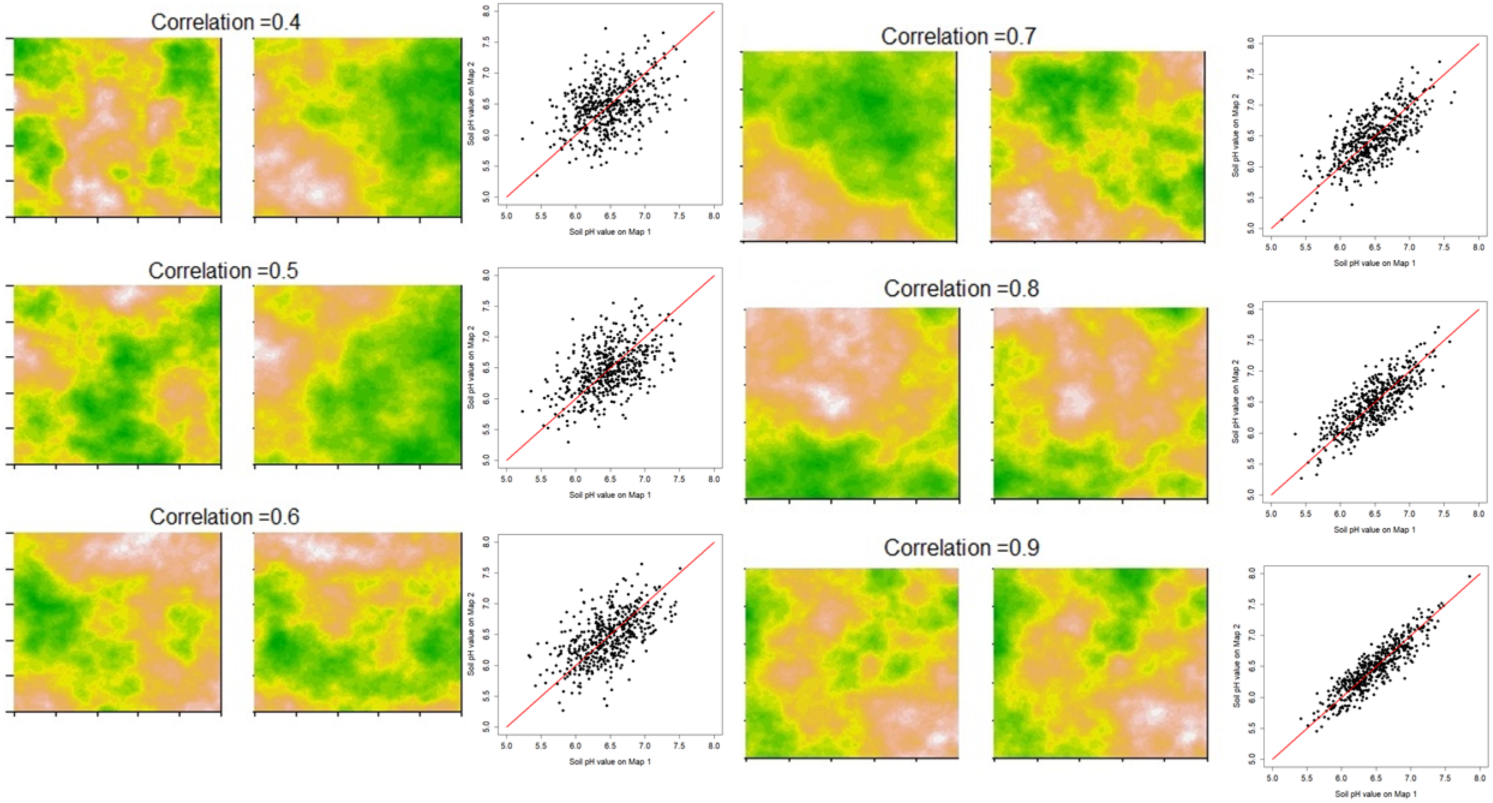


Figure S5. The pairs of example maps of soil pH, each pair being based on a different grid spacing, with a different offset correlation and corresponding scatter plots that illustrated the strength of the correlation.

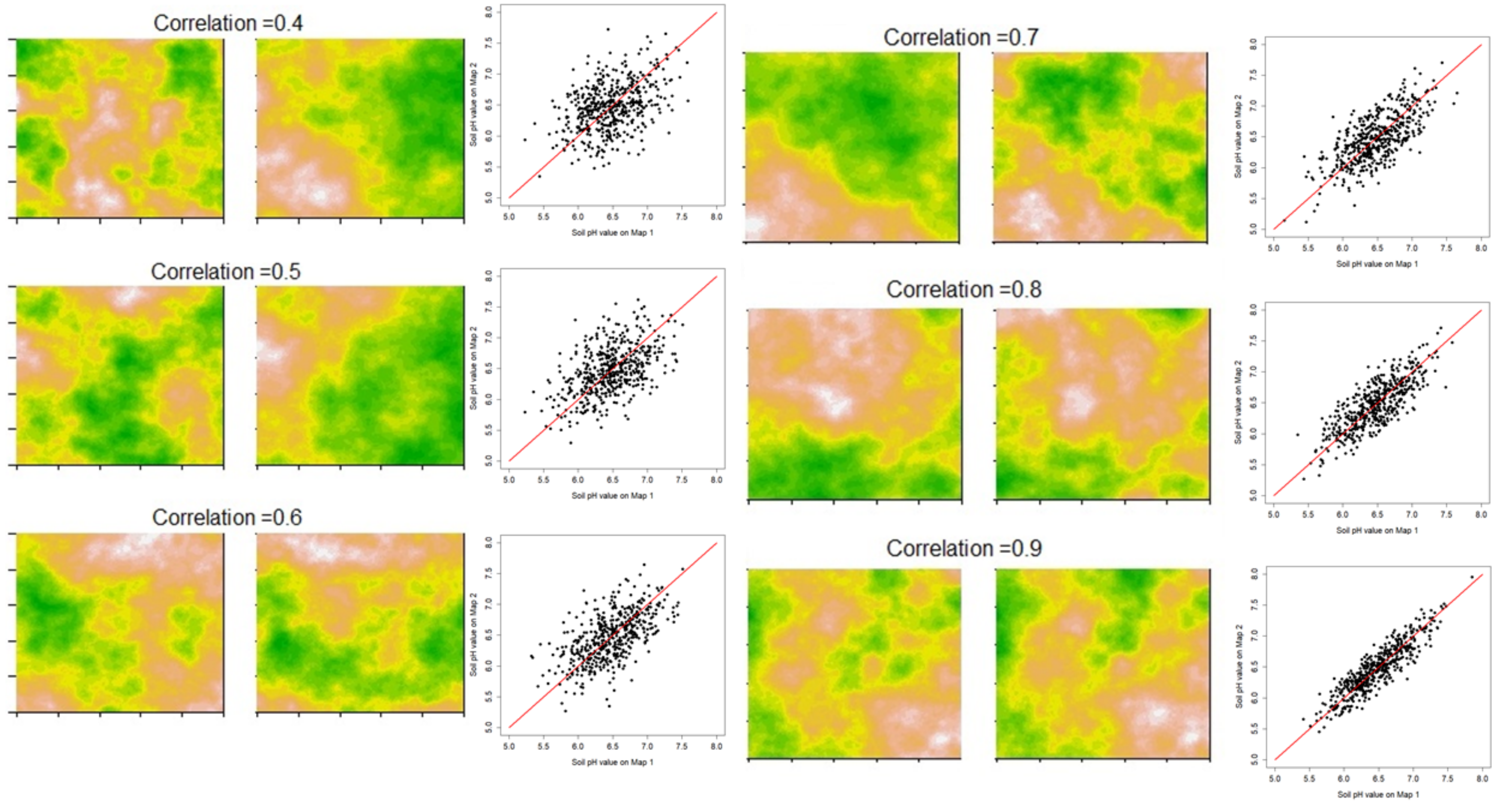


Figure S6. The pairs of example maps of Se concentration in grain, each pair being based on a different grid spacing, with a different offset correlation and corresponding scatter plots that illustrated the strength of the correlation.

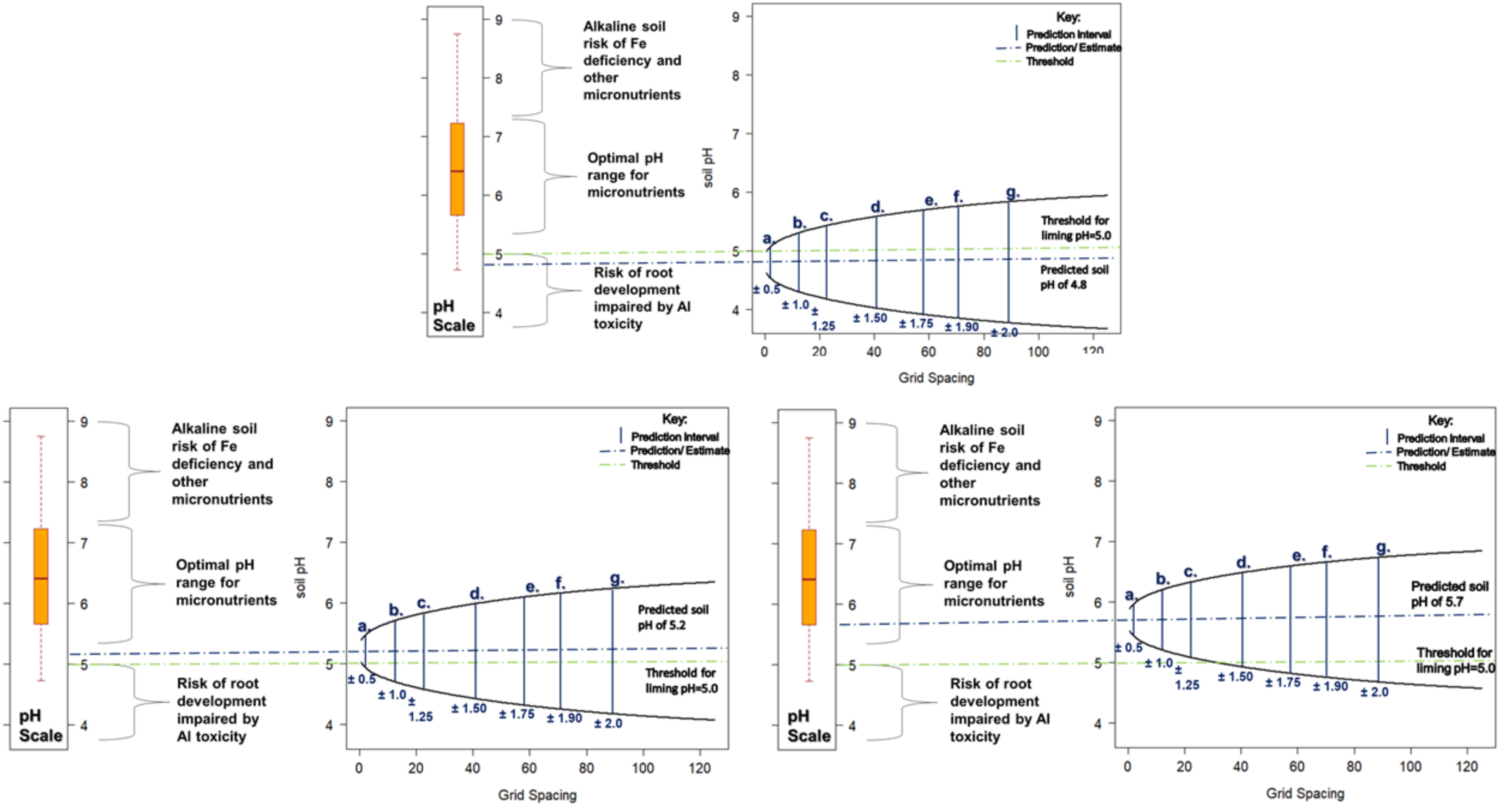


Figure S7. Chart consisting of box plot of the distribution of the soil pH, a graph of the lower and upper prediction intervals for the prediction for grid spacings from 0 to 120 km. With a blue line corresponding to the prediction and the green one for the threshold value.

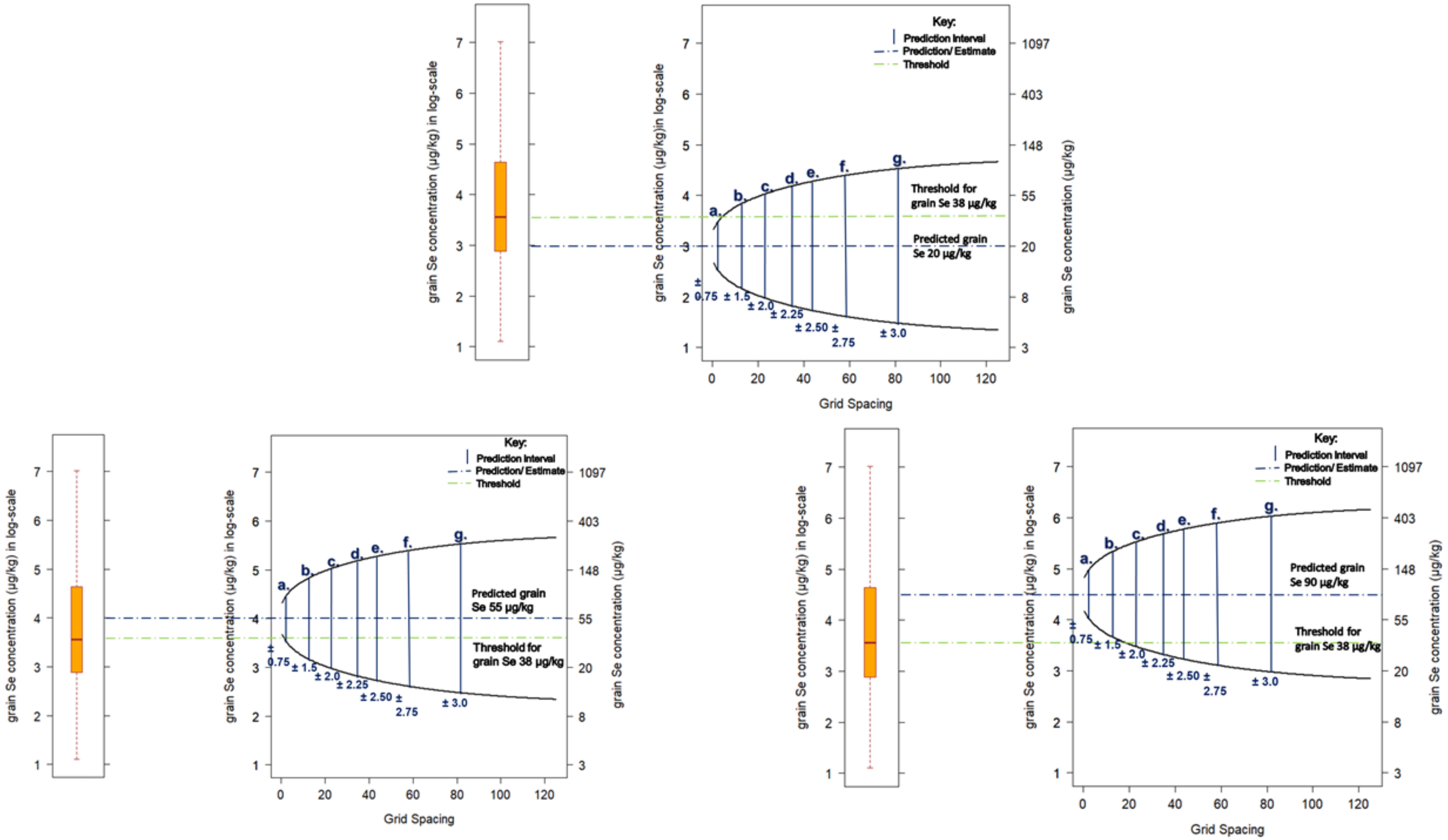


Figure S8. Chart consisting of box plot of the distribution of the Se concentration in grain, a graph of the lower and upper prediction intervals for the prediction for grid spacings from 0 to 120 km. With a blue line corresponding to the prediction and the green one for the threshold value.

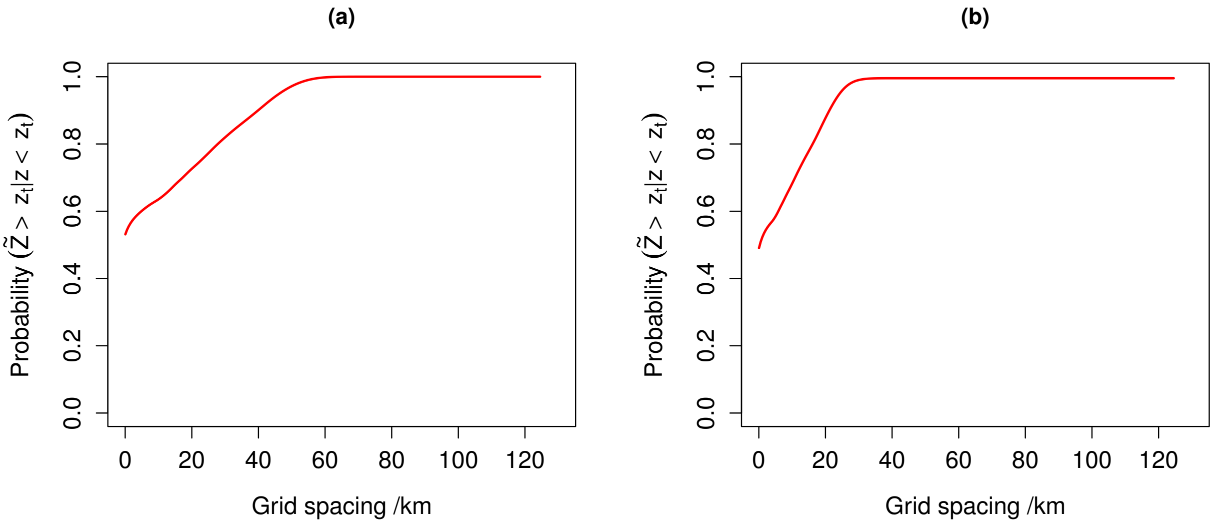
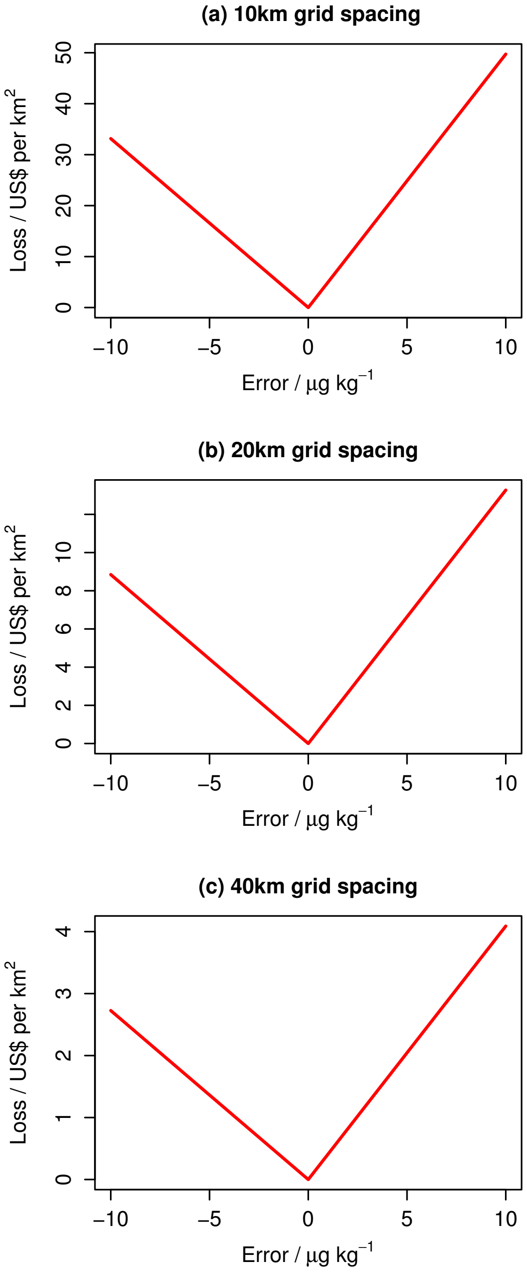


Figure S9. Graph showing the probability, given that an intervention is required at $\mathbf{x}_o$ that, due to error in prediction, the mapped variable does not show this. z_t is the threshold of interest. (a) is for soil pH and (b) for $\mathrm{Se}_{\mathrm{grain}}$ concentration.


(a) 10km grid spacing

(b) 20km grid spacing

(c) 40km grid spacing

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

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.

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.