Author Responses to Referee Comments:

We would like to thank the referees for their thorough review of our manuscript. We have revised the manuscript based on their suggestions and comments. We reply to each of the comments below. Our changes in the paper are in blue below, and the line numbers and sections refer to the revised manuscript.

Referee Comments 1 (RC1):

Referee Comment	Author Comment
(i) The shift from the goal of science communication towards complex mathematical modelling leaves me perplexed regarding several key aspects.	(i) It is precisely this shift between complex model output and communication which is the concern of this paper. Other papers we have published (Chagumaira et al., 2021, 2022) have addressed the question of how the output of such models, which quantifies the uncertainty of spatial predictions, can be communicated to users of the information. In this paper we recognize that one of the strengths of the geostatistical modelling process which statisticians have exploited is that the uncertainty of predictions can be proposed a priori for different sampling intensities, which can help when deciding how much effort to put into field work. However, this only works if the criteria for information quality can be communicated effectively to all stakeholders as the decision on survey effort is one which must be made collaboratively, the final decision resting with the survey sponsor or science lead who typically might not have statistical expertise. That is what we address in this paper. When translated effectively, mathematical models are powerful tools for engaging diverse audiences, and explore different scenarios and understand the cause-and-effect relationships within geosystems.

	To clarify this for the reader we edited the introduction from L22 to 96 (revised manuscript). This contains additional material included to address the reviewer's point about technical detail in the paper. *1.1 Mapping to support decisions: importance of spatial mapping* Spatial information is needed to support decisions at different spatial scales. Many approaches can be used to predict soil or grain properties at unsampled locations (e.g. machine learning and geostatistical methods). These methods make predictions based on a set of point observations configured on a systematic grid or spatial coverage sampling design. Geostatistical methods capture the spatial dependence by modelling the variation as an outcome of a random process (Webster, 2000), whereas machine learning methods do not entail a statistical assumption about the distribution of soil or grain property. Therefore, geostatistical methods offer an approach to sampling because they leverage on the statistical model that provides a basis for planning sampling given a statistical model. Spatial information is usually derived from field data obtained in surveys. These surveys have costs: travel and logistics, staff costs, time for community engagement, management costs and analytical costs for processing material collected in the field. The denser the sampling the higher the quality of the resulting information (in the sense that the uncertainty attached to spatial predictions is reduced). However, there are diminishing returns to increasing survey effort, and so there is an optimal survey effort where the marginal costs of the survey match the marginal improvement in the resulting information (Lark et al., 2022). The dependence of the quality of spatial information on survey effort has been studied by geostatisticians. In a geostatistical model the value of a variable at an unsampled location has a prediction distribution, conditional

on the model and the data. The variance of this distribution, however, the kriging variance, is conditional on the model only and so can be calculated from the model for any posited set of observations. McBratney et al. (1981) showed how, given a variogram model, ordinary kriging variances could be computed at the cell centres of square sampling grids of different spacing. A plot of kriging variance against spacing could be used to select a sampling grid spacing if a target kriging variance can be specified. A technical challenge is how to obtain the variogram before sampling. One might undertake a reconnaissance survey (particularly when a large final survey is envisaged) to estimate the variogram and use a Bayesian approach to account for its uncertainty (Lark et al., 2017), use a variogram from a cognate area (Alemu et al, 2022), use an average variogram for the variable derived from published studies (Paterson et al., 2018), use a variogram elicited from experts (Truong et al, 2013) or use an adaptive sampling strategy with several phases in which the spatial model is the primary output from early phases (Marchant and Lark, 2006). The general approach of sampling design for ordinary kriging, which McBratney et al. (1981) developed can also be extended to the more general case of spatial prediction from a linear mixed model with spatially correlated random effects and fixed effects which include covariates such as measurements from remote sensors, variables derived from digital terrain models and factorial covariates such as soil maps (Brus and Heuvelink, 2007).

1.2 Communicating the uncertainty of spatial information from proposed survey designs.

Despite this effort to address the statistical component of survey planning, the generation of measures of uncertainty for particular proposed designs, there has not been a complementary effort on how these measures are understood by stakeholders, such as survey sponsors, who might have the final responsibility of setting a survey budget, and so determining the quality of the resulting information. Previous studies, including that by

Chagumaira et al (2021), have shown that non-statisticians commonly do not find the kriging variance a meaningful measure of uncertainty to interpret spatial predictions, so it is unlikely they would find it useful as a measure of the quality of survey outputs to balance against costs.

In this study we worked with stakeholder groups (soil science, agronomy, nutrition, and public health) to examine how they interpret measures of survey quality, and whether they regard them as suitable for guiding a decision on the density of samples to be required for a survey. The measures we considered were all ones which could be derived from an initial variogram of the target variable, and we outline them briefly here, more detail is given in the Appendix A (A1—A4).

1.3 Proposed methods for communicating information quality.

We consider two measures derived from the kriging variance as measures of information quality. The first is the *prediction interval*, the interval which includes the unsampled value with some specified probability. Prediction intervals for surveys on grids of different spacing were proposed, in visual form which allowed the user to evaluate them relative, for example, to differences between critical values of the target variable for management purposes. The second measure was based on the *joint probability* that a location requires some intervention (because the surveyed variable exceeds or falls below a threshold) and that the spatial prediction at that location indicates the contrary. This was proposed because we found that stakeholders were generally receptive to presentations of uncertain information based on the probability that the mapped variable falls above or below a significant threshold (Chagumaira et al., 2021).

The third measure which we considered is based on value of information theory (Journel, 1984; Lark et al., 2022). It is the *implicit loss function* (Lark and Knights, 2015). A loss function represents the loss incurred when a

decision is based on spatial information which is correct (loss = 0) or in error (loss ≥ 0). This is used to analyse quality of information in cases where losses are reasonably straight forward to specify for different scenarios (e.g. Ramsey et al., 2002). Lark and Knights (2015) proposed that, for more complex cases, the implicit loss function might be used in critical assessment of a specified level of survey effort, based for example, on a fixed budget. An implicit loss function is one which, given a model of survey logistics, and statistical information (such as a variogram when the information is obtained by geostatistical prediction) makes a specified survey density the rational choice, i.e. the choice under which a marginal increase in survey cost is equal to the marginal reduction in expected loss when decisions are based on the resulting information. Lark and Knights (2015) proposed that reflection on the implicit loss function would help stakeholders to decide whether a proposed survey budget is consistent with stakeholders' views on the implications of making decisions with uncertain information, and we evaluated that here.

The fourth measure which we considered is the *offset correlation*. This is a measure of the consistency of spatial information produced when surveying at a particular grid spacing. Lark and Lapworth (2013) considered a hypothetical case in which a variable is mapped by ordinary kriging from data on a sample grid of spacing ζ, a second map is then made of the same variable and from a grid of the same spacing, but in which the origin is shifted from the original grid by $\zeta/2$ in each direction. They showed that, for a specified variogram, the correlation of the mapped values at some location increased as the sampling grid became denser. We suggested that this minimum offset correlation (at a location furthest from a sample point in either grid) is an intuitive measure of the quality of a survey output, it shows the extent to which the mapped value of the variable is robust to the location selected as the origin of the survey grid.

	Our primary interest was how far stakeholders felt confident in using these measures of the quality of spatial information as a basis for selecting the sample spacing (and hence the cost) of a hypothetical survey which they were engaged in planning. This is not a sufficient basis for deciding whether a criterion is robust and useful, but it is a necessary basis, since unless stakeholders feel that they understand a method and its implications they cannot be expected to use it. We also examined how far the stakeholders interpretations of the criteria were internally consistent (i.e. with the definitions of each criterion) and examined how far they resulted in consistent selections of grid spacing for a particular variable. Marchant, B.P & Lark, R.M. (2006). Adaptive sampling for reconnaissance surveys for geostatistical mapping of the soil, European Journal of Soil Science. 57, 831–845 Brus, D, J & Heuvelink, G.B.M (2007). Optimization of sample patterns for universal kriging of environmental variables, Geoderma, 138(1), 86-95. Ramsey, M.H., Taylor, P.D., Lee, J.C., 2002. Optimized contaminated land investigation at minimum overall cost to achieve fitness-for-purpose, Journal of Environmental Monitoring, 4, 809–814. Webster, R., 2000. Is soil variation random? Geoderma, 97, 149-163.
(ii) For example, the process by which maps "were presented to a group of stakeholders, who were asked to use them in turn to select a sampling density" is vague. Apart from the presentation style to the participants, it's crucial to recognize that the map quality presented to stakeholders is influenced by multiple variables, such as the	(ii) The reviewer refers to hypothetical map pairs designed to allow the respondent to visualize what two spatial variables with a certain correlation might look like. This is in the specific context of the offset correlation measure. We edited the text at L173 (in the revised manuscript) to:

initial sampling density, data quality, sampling design, and robustness of the geostatistical models themselves. The paper's direction, asking stakeholders to rank methods based on their effectiveness, seems ill-defined and lacking theoretical grounding. Furthermore, the framing of the quality of input data solely as a function of sampling density is simplistic and ignores other essential qualitative considerations, such as the type of data collected. While the decision on sampling density is undoubtedly vital, the assumption that it should be taken a priori disregards other vital factors like the scale, sampling strategy (design), and the level of detail of the phenomenon under study. These become paramount when assessing the overall quality of the required output.	We presented the participants with correlated pairs of hypothetical maps of soil pH and ${\rm Se_{grain}} $, with differing correlations, so that the extent to which maps might differ as a result of the grid offset could be visualized
A particularly concerning aspect of this paper is the scant information provided about the stakeholders' selection, background, and representation. Grouping of soil scientists and agronomists together, for instance, and juxtaposing them with public health experts and nutritionists, lacks clear justification.	Stakeholders with different disciplines need to work together on complex problems such as MND, and it makes sense to involve them all in the process. As the elicitation obtained individual responses, the background of each was recorded and accounted for in the analysis (see, for example, Table 2) we were able to assess any differences in understanding associated with educational background, training, and experience in different disciplines. This is an essential element of understanding for our work. We made this change on L98 to L118, to clarify this: This study was conducted with information-users who have been involved with the GeoNutrition project (http://www.geonutrition.com/), which examined strategies to alleviate micronutrient deficiencies (MNDs) in Ethiopia and Malawi and included surveys to provide baseline information on MN concentrations in staple crops and soils, and soil properties (such as pH) which influence soil to plant transfers of MN. The GeoNutrition project had teams from multiple disciplines (agriculture, soil science, human nutrition, and public health). It has been shown that concentration of micronutrients in staple crops and in soils vary spatially, as do biomarkers for MN status and so interventions to address the deficiencies

	should be based on spatial information on all these variables (Gashu et al., 2021; Botoman et al., 2022). The spatial information therefore has to be interpreted by information users from this broad set of disciplines, and all of them might also contribute to decisions on the amount of effort to be expended on field survey. It is plausible that experts with training in different disciplines might find different quantitative methods to express uncertainty in information useful for decision-making, and so we recruited a panel for elicitation which spanned these disciplines. We recruited the panel from institutions which were partners of the GeoNutrition project research team and the allied Translating GeoNutrition project in Zimbabwe (ZimGRTA) and the University of Zambia. The participants were volunteers, recruited from national-level institutions with responsibility for interventions and policy working on agricultural research and extension services, public health bodies and nutritional research. Soil scientists from the UK were also included. Panel members were invited by email from the local GeoNutrition/ZimGRTA lead. Most of the participants were familiar with the GeoNutrition project. They had an interest in contributing to the planning, execution, and interpretation of surveys to address micronutrient deficiencies, and so were aware of the importance of being able to engage with the process. Ethical approval to conduct this study was granted by the University of Nottingham, School of Biosciences Research Ethics Committees (SBREC202122022FEO) and participants gave informed consent to their participation and subsequent use of their responses. No remuneration was offered, but all participants in African countries who were not able to participate from institutional offices were provided with a one-day data bundle to allow them to join online.
The omission of machine learning and AI algorithms, especially in an era defined by big data, further complicates the study's scenario of understanding uncertainties.	Machine learning and AI are important topics in digital soil mapping. However, because they do not entail a statistical model, they provide no basis for rational decisions on sampling intensity. Some work has been done on sample design for ML-based mapping, but these are purely heuristic methods which do not allow sample density to be linked to the

	quality of the resulting predictions. As we note at L26 to L29 (revised manuscript) our approach is entirely compatible with statistical methods for spatial prediction which use any of the “big data” sources deployed in digital soil mapping
Abstract: The abstract begins with an unconventional approach, dedicating over five lines to general statements (L1-5). This choice leads to a lack of specificity in addressing the real problem of communicating uncertainty during the planning stage of a geostatistical survey. While there is no scientific dispute that sampling density correlates with prediction uncertainty, this paper fails to elucidate what sets it apart from existing knowledge. There's an opportunity to articulate unique perspectives or new insights on uncertainty, but the paper does not seize it. The introduction of four different ways in which "the relationship between sample density and the uncertainty of predictions can be related" falls short of justifying this research, as no novel insights or values are identified. The abstract would benefit from a more concise focus on the specific problem at hand and a clear rationale for why the chosen methodology is innovative or necessary. Without these clarifications, the abstract's approach feels redundant and fails to engage the reader in a meaningful way. L8-9: “All four of these methods were investigated using information on soil pH and Se concentration in grain in Malawi” à Investigated in what sense? Please be specific. Additionally, the term "stakeholders" is used in an overly generic way, particularly in the abstract. This lack of specificity leaves the reader wondering who exactly these stakeholders are. Without understanding their roles, experiences, backgrounds, and locations, it's challenging to gauge the relevance and applicability of their opinions and decisions in the context of the research. The paper would benefit greatly from identifying these stakeholders more precisely. Are they soil scientists, agronomists, public health experts, or nutritionists? What qualifies them to contribute to this particular study? The clarity on	We thank the referee for their comments and suggestions. We have restructured the Abstract (L1 to L21): Much research has examined communication about uncertainty in spatial information to users of that information, but an equally challenging task is enabling those users to understand measures of uncertainty for surveys of different intensity (and so cost) at the planning stage. While statisticians can relate sampling density to measures of uncertainty such as prediction error variance, these do not necessarily help stakeholders (e.g., agronomists, soil scientists, policy makers and health experts) to make rational decisions about how much budget should be assigned to field sampling to produce information of adequate quality. In this study, we considered four ways to communicate uncertainty associated with predictions made based on data from a geostatistical survey, to determine an appropriate sampling density to meet stakeholders expectations. These include two methods based on the conditional prediction distribution: the width of prediction intervals, and the joint probability that a particular intervention is required at a random location, but the spatial information indicates the contrary. A third method, the offset correlation is a measure of the consistency of kriging predictions made from data on sample grids with the same spacing but different origins. The implicit loss function is a method which allows the user to reflect on the valuation of losses from decisions based on uncertain information implicit in selecting some arbitrary sampling density. Evaluation of the four communication methods was done through a questionnaire by eliciting opinions of participants with experience in planning surveys, about the method’s comprehensibility and effectiveness and the sampling density that they would select based on that method. Our results show significant differences in how the

these questions would not only strengthen the abstract but also establish a solid foundation for the rest of the paper. My concerns regarding the selection, experience, and location of these stakeholders will be elaborated further in the subsequent sections of this review.	participants responded to the methods, with the joint probability and implicit loss function approaches being not well understood, and offset correlation was the most understood. During feedback sessions, the stakeholders highlighted that they were more familiar with the concept of correlation, with a closed interval of [0,1] and this explains the more consistent responses under this method. The offset correlation will likely be more useful to stakeholders, with little or no statistical background, who are unable to express their requirements of information quality based on other measures of uncertainty.
1 Introduction: The introduction of the paper provides a broad overview of the study's themes, but it appears superficial and lacks a specific focus on the paper's actual subject. Instead of honing in on the unique issue this research aims to address - namely, how stakeholders deal with uncertainty in planning mapping surveys - the section tends to wander through various unrelated topics. For instance, the introduction's first paragraph begins with a discussion of MND in sub-Saharan Africa, a detail that seems incongruent with the non-location-specific context of the research. A considerable portion of the content here is overly generic and fails to pinpoint the problem the study seeks to explore. Furthermore, the paper makes a convoluted attempt to rationalize the methods (such as offset correlation, implicit loss function, kriging variance, conditional probability) presented to the stakeholders. These methods are described in laborious detail, yet the rationale for their comparison remains unclear. The text also fails to address whether there is scientific consensus on the superiority of any one method. Much of this section is bogged down with technical details that may be unengaging for a wider audience. Simplifying some of these concepts would make them more accessible and align better with the journal's target readership. For example, the sentence '… an implicit loss function,	We thank the referee for the opinion about the structure of the introduction. In response to the referee's comment on the first paragraph, we wanted to give context of how geostatistical mapping has been used as a tool to provide information about soil and crop micronutrient properties in relation to mineral micronutrient deficiencies in sub-Saharan Africa. Studies conducted in this region provided evidence that concentration of micronutrients vary spatial and spatial information is important to design efficient interventions. We have provided this background information and the rationale why we grouped soil scientists, agronomists, and public health experts, in Section 3 from L98. We also revised the Introduction to reflect the views of the referee, and we already have presented the proposed revisions for the introductions (from L22 in the revised manuscript) in the first response to this referee comments (RC1).

conditional on a logistical model (i.e., a function of sampling effort and statistical information about the estimates of the cost of errors) can be modelled as the loss function that makes a particular decision on sampling effort rational (Lark and Knights, 2015)' is impossible to parse. The paper's emphasis on the intuitiveness and simplicity of the offset correlation method is presented as an advantage. However, this approach seems to oversimplify the complexities involved and may even have biased the stakeholders towards this method. The way the study was constructed raises concerns that the stakeholders might have been subtly steered towards favouring the offset correlation method. As already indicated, I suggest that a more engaging introduction is written underlining the theoretical foundations of the study and providing compelling justification for study's approach.	
2 Theory: This section, as it currently stands, does not appear to add significant value to the overall paper. Although some readers might find it informative, its current content might be better suited for an appendix. I recommend relocating this material and replacing it with a comprehensive literature review that outlines the current state of knowledge regarding decision-making in soil and plant surveys. This could include case study examples highlighting the tangible costs incurred by stakeholders who failed to adequately plan and make informed decisions prior to their survey efforts. Further, it would be enlightening to specify the types of stakeholders you have in mind for this research. Detailing their background and roles will help readers better understand how their specific attributes might influence their decision-making process. By making these adjustments, you can create a section that not only maintains the reader's interest but also lays a more robust foundation for the arguments and findings presented later in the paper.	We acknowledge the referee's comment, and we will move this text to the Appendix A for the benefit of readers for whom the mathematical content is of limited interest.

3 Materials and methods **3.1 Basic approach** (i) L176: “We used the four methods, described above, to assess uncertainty in relation to sampling density, considering the problem of measuring a soil property relevant to crop management: soil pH, and a property of the crop: Se_{grain} concentration.” àI am not sure what is meant here with assessing uncertainty in relation to sampling density. Also, what “problem” is there when measuring a soil property relevant to crop management? And why specifically soil pH? And Se? Providing this context will not only enhance the reader's understanding but also reinforce the motivation behind the study, making it easier to follow the progression of the research and its significance within the broader scientific landscape. L177-180: “We used variograms from a national survey in Malawi for each variable (Gashu et al., 2021) to obtain sampling densities for further notional sampling for an administrative district in Malawi, Rumphi District, with an area of 4769 km2. The outputs were presented to participants”. à While the paper draws on the dataset from Gashu et al. (2021), further details on how this dataset was collected, along with the rationale behind its selection, would strengthen the connection between the data and the study's objectives. Specifically, it is essential to explain the methodology used in collecting the dataset, including how the parameters of the variograms were selected to derive the sampling densities. This information will provide readers with a clear understanding of the data's reliability and relevance to the study. The paper should address potential biases that could arise from using variograms of national level data to derive	(i) We thank the review for these comments to improve our paper and we wish to edit our manuscript to reflects the points raised by the review. We have revise the section on the Materials and Method and added information about why we used the data from the GeoNutrition project and how it was collected and whether this was adequate to support predictions at regional level (see L98 to L125 revised manuscript). We described in detail how the variograms were modelled in Section 2.2.1 – where we described statistical modelling and spatial predictions of grain Se and pH.

	regional sampling densities, especially considering the comparison of four different methods. Are there similar machine learning approaches? This section must articulate the steps taken to minimize biases, ensuring that all four methods were optimal for the input dataset. Providing a context for how the study's scenario would apply to stakeholders needing to understand uncertainty without national-level data will help readers gauge the broader applicability of the findings. Further, clarity on how the output was presented to the participants, whether through PowerPoint, poster format, or other means, and the order of presentation is crucial. These factors could significantly influence participants' understanding and choices and acknowledging them in the paper will enhance the transparency of the process. By addressing these points, the paper can offer a more comprehensive and clear understanding of the data, methods, and process, enhancing both its scientific rigor and accessibility to a broader audience.		
(ii)	L180: “The participants considered each method in turn and were asked to select a sampling grid density based on the method. After doing this they were asked, for each method: Has the method helped you assess the implication of uncertainty in spatial prediction in as far as it is controlled by sampling? They were then asked: Which of these methods was easiest to interpret? Finally, the participants were asked to rank the method in terms of ease of use. Evaluation of the test methods were done using an online questionnaire on Microsoft Forms” à How! Which aspect of the methods were considered? Was the quality of the method with regards to the output or which specific aspect? On the question of “easier to interpret”, how do authors define “easier”? This question is	(ii)	The list of specific questions used to elicit stakeholder are listed in Table 1. These questions were sufficient for the participants to understand for example: “We show you here some pairs of examples map of soil pH/Se grain, each pair being based on a different grid spacing, and so, with a different offset correlation. We also show scatter plots which illustrate the strength of the correlation. What do you think is the smallest correlation that would be acceptable if one of the maps were to be used to make decisions?” We think this is clear enough question which prompts critical thinking about the smallest correlation they deemed acceptable to decide. However, we have edited the text so that it becomes clearer to the reader see L149 to L158 in the revised manuscript.

	loaded with so much subjectivity that without a clear unbiased scale of what “easy” means, it is impossible to derive any meaning from their answers.	
(iii)	L187-195: “The invited participants self-identified as (i) agronomist or soil scientist or (ii) public health or nutrition specialists. The participants also self-assessed their statistical/mathematical background and their frequency of use of statistics in their job role (perpetual, regular, occasional use)”. à Given that this information is one of the pillars of your findings in this work, I wonder why there wasn’t the attempt to standardize the backgrounds of the participants. For instance, what qualifies one as any of the professions (agronomist, soil scientist, nutritionist, and public health specialist). Is it based on education level, years of practice, specific training, or other criteria? Was there a reason why such experts were chosen? Do these experts typically have training in interpreting uncertainty in maps? Elaborate on why the distinctions among these professionals were used as the basis for the response. Address whether the 26 participants were intended to represent a broader population or if they were selected for specific reasons. Justify the choice of only 26 participants for this study. Explain why this number was deemed sufficient, considering the scope and objectives of the research. If the sample size is indeed small, acknowledging its limitation and potential biases will improve the rigor of the study.	(iii) We will add the following text from L392, revised manuscript, to address the issue raised by the referee: All the stakeholders recruited in this study were employed in public sector institutes in roles (e.g., universities, civil organisations, research, and extension) and had experience in their respective fields in an SSA setting. In terms of sample size, we have no prior basis to select a sample size because, as this was the first study of this topic, it was not clear how to select an appropriate effect size. As a result, our major consideration was recruiting individuals willing to participate and with experience in their respective institutions. We therefore attempted to recruit the entire set of suitable respondents in each country. In future work initial power analysis might be considered.
(iv)	L195-200: “In the exercise, an introductory talk was given to explain the study’s objectives. During the talk, we explained the four test methods (offset correlation, prediction intervals, conditional probabilities and implicit loss function) and how they can be used to assess the implications of uncertainty in	(iv) Our participants were stakeholders in that they had an interest in being better able to contribute to the planning, execution, and interpretation of surveys to address MND. They were volunteers, recruited from national-level institutions with responsibility for interventions and policy, they were familiar with the GeoNutrition

	spatial predictions to determine appropriate sampling grid space for a geostatistical survey. We explained the structure of the questionnaire to the participants. We emphasized to the participants that we were not testing their mathematical/statistical skills and understanding but rather were testing the accessibility of the methods using their responses"à What drove the participants to engage in the exercise? Understanding their motivations can shed light on the relevance of their input and the validity of their responses. Were they incentivized in any way? Did they have personal or professional interests in the outcome? The term "stakeholder" typically implies an individual or group with a vested interest in the outcome of a particular process or decision. In this context, it remains unclear if the participants indeed stood to gain or lose anything from the exercise. If they did not have a direct stake in the findings or implications of the research, using the term "stakeholder" might be misleading. An explanation or justification for this terminology would enhance the clarity and precision of the paper.		project and so were aware of the importance of being able to engage with the process. They gave informed consent to participate in the elicitation. No remuneration was offered, but all participants in African countries who were not able to participate from institutional offices were provided with a one-day data bundle to allow them to join online. We have provided this information in the revised manuscript please see L97 to L118.
(v)	L205-210: "The offset correlation was the first method presented to the participants. This was followed by prediction intervals and conditional probabilities. The implicit loss function was the final method presented to the participants. We started with a measure we thought all our stakeholders would most easily understand and then moved on to the more complex methods." à The presentation of the offset correlation method within the research design appears to have been conducted in a manner that may have inadvertently favored this approach. Was there any randomization in how the different methods were presented to the participants? If the offset correlation was consistently presented first, or in a way that highlighted it	(v)	It is easy to "blind" in an experiment when you are giving a subject one of two indistinguishable pills, but hardly relevant to this case. We revised the Materials and Methods section to address the issues raised by the referee from L97 to L163: **3. Materials and Methods**

	more prominently, this could influence participants' perceptions and evaluations. Were all the methods described with equal clarity and neutrality? Any differences in language, emphasis, or complexity might have created an uneven playing field, leading participants to gravitate toward the offset correlation method. Was there any attempt to control or assess the potential for bias in how the methods were presented and evaluated? Implementing and reporting on measures such as blinding or counterbalancing could strengthen the credibility of the results. Were participants' preconceived notions or preferences regarding these methods assessed or controlled for? Their prior knowledge or beliefs could also contribute to a bias in their evaluations. Addressing these questions would help to ascertain whether the apparent favoring of the offset correlation method is a genuine reflection of its merits or a product of the research design. A robust examination of these concerns would enhance the rigor and validity of the findings, ensuring that the conclusions drawn are founded on an unbiased assessment of the methods in question.	This study was conducted with information-users who have been involved with the GeoNutrition project (http://www.geonutrition.com/), which examined strategies to alleviate micronutrient deficiencies (MNDs) in Ethiopia and Malawi and included surveys to provide baseline information on MN concentrations in staple crops and soils, and soil properties (such as pH) which influence soil to plant transfers of MN. The GeoNutrition project had teams from multiple disciplines (agriculture, soil science, human nutrition, and public health). It has been shown that concentration of micronutrients in staple crops and in soils vary spatially, as do biomarkers for MN status and so interventions to address the deficiencies should be based on spatial information on all these variables (Gashu et al., 2021; Botoman et al., 2022). The spatial information therefore must be interpreted by information users from this broad set of disciplines, and all of them might also contribute to decisions on the amount of effort to be expended on field survey. It is plausible that experts with training in different disciplines might find different quantitative methods to express uncertainty in information useful for decision-making, and so we recruited a panel for elicitation which spanned these disciplines. We recruited the panel from institutions which were partners of the GeoNutrition project research team and the allied Translating GeoNutrition project in Zimbabwe (ZimGRTA) and the University of Zambia. The participants were volunteers, recruited from national-level institutions with responsibility for interventions and policy working on agricultural research and extension services, public health bodies and nutritional research. Soil scientists from the UK were also included. Panel members were invited by email from the local GeoNutrition/ZimGRTA lead. Most of the participants were familiar with the GeoNutrition project. They had an interest in contributing to the planning, execution, and interpretation of surveys to address micronutrient deficiencies, and so were aware of the importance of being able to engage with the process. Ethical approval to conduct this study was granted by the University of Nottingham, School of Biosciences Research Ethics Committees (SBREC202122022FEO) and participants gave informed

	consent to their participation and subsequent use of their responses. No remuneration was offered, but all participants in African countries who were not able to participate from institutional offices were provided with a one-day data bundle to allow them to join online. We sought to explore how the information from the sampling conducted in the GeoNutrition project could be used to support decisions about sampling density for other similar projects. We used data from the GeoNutrition project, crop and soil properties were measured at national scale in Malawi. In this survey, field sampling was undertaken to support the spatial prediction of micronutrient concentration in crops and soil across Malawi. The sampling design was selected to achieve spatial coverage and used 'main-site' and 'close-pair' sampling to support the estimation of variance parameters of the linear mixed model (Lark and Marchant, 2018). The location of sample points were the centroids of the Delauny polygons, resulting from the stratification function in the spcosa library for the R platform (Walvoort et al. 2010). The sample support (0.1 ha circular plot) for the data consisted of bulk soil and grain samples from aliquots within a single field (Gashu et al., 2020). Therefore, the uncertainty quantification of the predictions relates to the mean values of the target variable across such as support within a field at a specified location, and this is appropriate for deriving regional sampling densities. Details about field data collection in Malawi are presented by Gashu et al. (2021), Botoman et al. (2022) and Kumssa et al. (2022). Grain and soil samples were prepared and analysed using methods described in Gashu et al., 2021. We used variograms for soil pH and Se_{grain} to obtain sampling densities for further notional sampling for an administrative district in Malawi, Rumphi District, with an area of 4769 km^2. The outputs were presented to

participants in poster format through PowerPoint, and examples of the posters are shown in Figs. S5 – S10 in the Supplement.

2.1 Format of the exercise

We wanted to elicit from stakeholder the usefulness of proposed methods (offset correlation, prediction intervals, conditional probabilities) in helping them assess the implications of uncertainty in spatial prediction in as far as this is controlled by sampling, considering the problem of measuring a soil property and a micronutrient from a crop. Soil pH and concentration of Se in grain were used as examples for this case study. We invited professionals working in agriculture, nutrition and health at civic organisations, universities, government departments from Ethiopia, Malawi and wider GeoNutrition sites (United Kingdom, Zambia, and Zimbabwe). In total we had 26 participants (18 were agronomists or soil scientists and 8 public health or nutrition specialists).

The elicitation was conducted online using \cite*{Zoom} in two sessions, 26$^{\rm{th}}$ and 28$^{\rm{th}}$ April 2022. There were two sessions to accommodate participants from different time zones, and to manage the participants in smaller groups to allow for questions and feedback. The invited participants self-identified as (i) agronomist or soil scientist or (ii) public health or nutrition specialists. The participants also self-assessed their statistical/mathematical background and their frequency of use of statistics in their job role (perpetual, regular, occasional use).

In the exercise, an introductory talk was given to explain the study's objectives. During the talk, we explained the four test methods (offset correlation, prediction intervals, conditional probabilities, and implicit loss function) and how they can be used to assess the implications of uncertainty in spatial predictions to determine appropriate sampling grid space for a geostatistical survey. We explained the structure of the

	questionnaire to the participants. We emphasized to the participants that we were not testing their mathematical/statistical skills and understanding but rather were testing the accessibility of the methods using their response. The participants considered each method in turn and were asked to select a sampling grid density based on the method. Evaluation of the test methods was done through a questionnaire, as shown on Table 1. Using the first four questions, Q1 to Q4, we wanted to find out if the method helped to identify a sampling grid spacing. On Q5, we wanted the participants to assess the test methods in terms of their effectiveness in finding an appropriate grid spacing. We asked the participants to rank these methods in an order of their effectiveness, in their experience, and in terms of finding a level of uncertainty that they were able to tolerate when deciding about a sampling grid spacing. We asked them to put rank 1 as the most effective method and rank 4 the least. The participants recorded their responses using and online questionnaire on Microsoft Forms. The offset correlation was the first method presented to the participants. This was followed by prediction intervals and conditional probabilities. The implicit loss function was the final method presented to the participants. We started with a measure we thought all our stakeholders would most easily understand and then moved on to the more complex methods.
3.2 Test methods (i) Most of the information in 3.2.1 largely repeats the information in 3.1. Thus, I suggest to fuse the information here with that of the section 3.1. I think some of the questions I raised in 3.1 is answered here so I suppose it makes it easy to fuse them. While it is common to cite previous studies for established	(i) Section 2.1 is an overview of the key experimental work, the engagement with the participants. Section 2.2 describes our analysis of the data sets and production of the outputs for the participants to use. We think it important to keep these quite distinct, and do not think that there is more than a superficial

	methods or data, in this case, where the dataset is central to the analysis, it may be beneficial to provide specific details rather than merely referring to other works. For instance, it is unexplained how soil pH and Segrain is measured. This will give readers a more comprehensive understanding of the methods and rationale behind the chosen measurements.		overlap. Also, we do not believe that the analytical methods are of special relevance to this paper so prefer not to include them.
(ii)	L218-220: I wonder how the mean of the soil pH was calculated. This is because it will be incorrect to just calculate the arithmetic mean of a phenomenon (like pH) that is on a log scale.	(ii)	We are reporting here are summary statistics to describe the distribution of the variable soil pH and to evaluate whether it appears normally distributed.
(iii)	L228-230: Any specific reasons for these minimum and maximum grid spacings?	(iii)	The grid spacings were considered because the span the axis from finer grid to a coarser grid to fully illustrate the different prediction intervals that can be achieved by sampling effort.
(iv)	L231: “We considered different prediction for each variable, but the prediction interval was fixed, depending only on grid spacing. The three predictions of soil pH were 4.8, 5.5 and 6.0 and those of Segrain were 20, 55 and 90 µg kg−1.”à I can understand the need to keep the same prediction intervals, but considering that the soil pH as a soil property and Segrain as a plant property will be subjected to different dynamics of spatial change, was there a way to account for this in the predictions?	(iv)	From our previous study Gashu et al. 2020 we showed that it is possible to examine spatial variation of soil and grain properties, sampled on an appropriate joint sampling design, by using model-based statistical analysis. The empirical best linear unbiased prediction (E-BLUP) has the allowance to add collocated and non-collected data and has an associated prediction error distribution that allows account for the differences in the predictions. We have provided this information in the revised manuscript see L122 to L132.

(v) L233: What kind of chart? Is it Figure 1? If so, then please state it.	(v) This chart refers to Figure 2 (revised manuscript). To make this clear, we edited the text on L191to L192: The predictions of soil pH and ${\rm Se_{grain}} $ concentration were presented to the participants in a chart as shown in Figure 2.
(vi) L234: “From the chart, we asked the participants to select the grid spacing that gives the widest prediction interval that would be acceptable if the mapped predictions were to be used to make decisions about soil management or interventions to address human Se deficiency.” à I find difficulty in embracing the premise upon which this question is constructed. Initially, my understanding was that the inquiries were primarily concerned with the planning of a geostatistical survey. Therefore, it confounds me as to why participants are questioned about employing the maps as a foundation for decision-making. In a theoretically optimal scenario, what would constitute the best choice of a prediction interval for such a decision?	(vi) The whole point of a survey is that it produces predictions, and the basic premise of our study, which we hope will be clear in the revised paper. Also, that prediction quality responds to survey effort. The geostatistical methods can capture the measures of the quality of spatial information explicitly. We have proposed changes in the introduction which makes it clear that the width of the prediction interval depends on the conditional prediction distribution and so on grid spacing.
(vii) L240-245: If a conditional probability of 1 indicates that the prediction is a equivalent to the overall mean of the dataset, does it suggest that the conditional mean of <1 is an indication of underestimation or over estimation of the true value at the given location? Also, I have the same issue with the question posed here as that posed in L234 above. L245-263: My concerns mirror those I previously expressed in section 234.	(vii) The probability goes to zero or to one because the prediction goes to the mean which either indicates the intervention or not, depending on the threshold and the mean value.

(viii) Participants are queried about interventions, but their responses are then utilized as a foundation for planning a geospatial survey. This connection appears incongruent, and it might be worth clarifying how the answers to these questions directly inform the planning process.	(viii) As noted above, it is fundamental for these approaches to survey design that the information is used for a purpose. We have made this clear, in the revised materials and method section.
(ix) Section 3.2.5: I'm grappling with a particular aspect of the offset correlation method, namely its use as a measure of similarity between two grid spaces. For this measure to function meaningfully in decision-making, one grid space must be taken as a reference, representing the closest approximation to reality. Then, higher correlation with this given reference space would indicate an optimal choice among the others. However, in the method's current presentation to participants, an issue arises. Specifically, there's a risk of bias propagation; grid spaces that are closer together are likely to show higher correlation compared to those farther apart. Similarly, coarser grid spaces might exhibit greater correlation across the board. These biases can distort the method's effectiveness. How did the authors address this potential source of error?	(ix) We hope that our revisions in the Introduction (see L83 to L90) and Appendix A4 (see L492 to L494), ensure the offset correlation concept is well understood by a broader audience. Making the grid spacing coarser always increases the offset correlation. The point is that we posit two maps based on the same grid spacing but offset by the maximum possible distance in each axis (half the grid spacing). Neither map is expected to be closer to reality than the other, the question is how consistent they are. The best analogy would be when a lab does triplicate analyses on some soil samples. We do not say “one of those three analyses must represent reality and we compare the other two with them”. Rather, we say, if our method is good enough then the three measurements should be consistent with each other. We added the following at L183 to L184: Neither of the posited pair of maps, based on offset grids, is to be regarded as closer to reality than the other, the question is how consistent they are.
(x) Figure 4: Please check, the caption mentions Segrain, but the figure indicates soil pH. Also, it would be meaningful for the reader to know the grid space of map1 and map2 that is giving the correlation value of 4. As I have indicated in my comment	(x) The caption has been edited to reflect the referee’s suggestions, see L188 revised manuscript.

above, it will be useful to know which of these two is closer to reality.	Figure 1. The pair of hypothetical maps of pH value and corresponding scatterplot for offset correlation 0.4. As noted above in (ix), there is no reason to believe that one map is closer to reality than the other. The question is how consistent are they?
3.3 Data analysis Section 3.3.1: “The expected number of responses under the null hypothesis, ei,j in a cell [i, j], is a product of row (ni) and column (nj) totals dived by the total number of responses (N), and this the null hypothesis of the contingency table which is equivalent to an additive log-linear model of the table” à What is intended by this sentence? Please consider revising it to be more comprehensible for readers who may not be statisticians. For example, instead of stating 'Contingency tables allowed us to test the null hypothesis of random association of responses with the different factors in the columns,' it would be more helpful to specify what the null hypothesis was in relation to the different responses. This clarification would illuminate the process and make the statement more approachable for a broader audience.	Thank you for this suggestion; we propose to edit the statement on L225 to L227. We analysed the contingency table is analysed on the basis of a null hypothesis that the distribution of observations between responses (e.g. selected grid spacing) is independent of the factor in the column (e.g. professional group). If evidence is provided to reject the null hypothesis, then this would indicate that how a respondent interprets the information presented to select a grid spacing depends on their professional group.
Table 2 appears to neither enhance the flow of the paper nor contribute to its content. Consider relocating it to the appendix, where it can be accessed if needed without interrupting the main narrative of the paper.	Table 2 was moved to the Appendix B (page 27 revised manuscript).
Section 3.3.2: “However, in our analysis we reversed the order by assigning a score of 4 for the most preferred method and 1 for the least.” à Why was it necessary to do this? Why wasn’t it possible to also offer to the participants the same way you analysed the data? Perhaps, you could have also tested if the sequence of the choices offered would have had an effect on the decision.	We did this to assign a score of 4 to the most preferred method, and 1 for the least to the most ranked method. The mean rank is computed from the product of the rank and score dived the number of participants. However, it makes no difference reversing the order or not. The respondents ranked the methods in order of their preference. To make it clearer, we edited the statement on L266 to L267 (revised manuscript):

	However, to calculate the mean rank, r, for each method for all the respondents, we assigned a score of 4 for the most preferred method and 1 for the least.
4 Results Section 4.1 test methods can be removed as there is no text under this section	Sub-sections 3.1.1 to 3.1.4 all are under the section 3.1 which presents the results from test methods. Then 3.2 presents results from assessment of the methods. Therefore, this heading is necessary to make this distinction.
Section 4.1.1 presents a discrepancy in the order of the methods, with the 'offset correlation' appearing last in the methods section but first in the results. To enhance clarity and consistency, I recommend aligning the order of appearance in both sections.	The offset correlation, see Section 2.2.2, now appears first in the methods section and this aligns to presentation of results.
L338-340: From what I understand so far about offset correlation, the correlation value is combination pair of two grid spaces, so what does it mean here that the grid spacing for soil pH is 25 km and that for Segrain is 12.5 km? What is the other pair in this correlation combination? Also, from figure 5, it can be seen that while most people indicated 0.7 correlation value, there were still a substantial number of people that selected the full range of the correlation values. Given the low number of participants (n) it will be useful to not only report on the most but also critically consider the other correlations. I think this is one of the major flaws of this study.	The summary given here does not reflect how the offset correlation is defined. The offset correlation is dependent on the variogram model of the property- we had two variograms one for soil pH and the other for $Se_{grain.}$ so, there is no reason to expect that the offset correlation will be the same for two different variables at a given grid spacing. We will revise the paper to improve our explanation (see comments above). In the revised paper we will comment on the range of values for the offset correlation.
Section 4.1.2: L345, do you mean there were no differences considering the p-value you reported? While I agree that the reported p-values suggest that the null hypothesis of uniformity in response cannot be rejected, it can be see from Figure 6 that the percentage of people that selected the grid spacing of 100 km (< 5 %) were substantially lower compared to the rest of the population, so what accounts for it?	While there is a fluctuation at 100 km the analysis tells us that it is potentially misleading to look for an explanation as the overall result is quite compatible with a random distribution.
5 Discussion L383-385: “In this study, we presented to groups of stakeholders, four methods (offset correlation, prediction intervals, conditional probabilities and implicit loss functions) that can be used to support decisions on	We will change stakeholders to “information user” and this change will be made on the revised manuscript.

sampling grid spacing for a survey of soil pH and Segrain.“à I don’t think you can regard your participants as stakeholders in this case. It still remains to be answered what is at stake for them.	
L385-390: “Offset correlation was ranked first as the method the stakeholders found easy to interpret (see Figure 9), and over 70% of the stakeholders specified a correlation of 0.7 or more as a criteria for adequate sampling intensity” à I am unsure where this 70 % is coming from because from Figure 5, it is only 30% that chose that 0.7 correlation. Since, the 0.7 value was chosen as part of an ordered categorical set of variables (from 0.4 to 0.9), it is inconsistent to draw a conclusion like “0.7 or more”. As I have already indicated in an earlier comment in the results, it is equally important to know why people chose 0.4 or 0.9 as their best choice of correlation coefficient for intervention.	We do not agree that this is an inconsistent interpretation. We clearly state that “**correlation of 0.7 or more.**” This means >30% for 0.7 plus 28% for 0.8 and ~15% for 0.9 which is over 70%. Furthermore, a respondent who thinks that an offset correlation of 0.7 is necessary would regard a design for which the OC was 0.8 as acceptable with respect to quality, but one for which it was 0.6 as not, so there is an asymmetry, and we can state that 70% of respondents thought that an offset correlation of 0.7 or more was acceptable. We have edited the sentence L333 to L336: Offset correlation was ranked first as the method the stakeholders found easy to interpret (see Figure 9), and most respondents (30%) selected an offset correlation of 0.7, and slightly fewer selected 0.8 so over half of respondents are accommodated within this range of values.
“During the feedback session, stakeholders highlighted that they were more familiar with the concept of correlation, with a closed interval of [0,1]. This explains why there more consistent responses under this method.” à This here is another major flaw in the whole study. Was it the stakeholders that selected 0.7 who made this declaration or was is it also the same for those who chose 0.4 and 0.9, because if the concept of correlation is familiar, then you would strive for a stronger correlation of 0.9 and not 0.7. Also, it is wrong to indicate that correlation has a close interval of 0 to 1, because the interval of correlation is -1 to 1.	We propose to make the following change on L337 to L338: This is likely to explain the consistency of the results for this criterion, with over half the respondents selection 0.7 or 0.8 as a minimum acceptable correlation. The question to the respondents was not “what correlation indicates the best design” but rather, what is the minimum acceptable correlation. That is very different. Also, to mention the offset correlation 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. We added the following text on L492 to L495:

	The offset correlation is bounded [0,1], which makes it intuitively easy to interpret as an uncertainty measure. It 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.
L390-393: "Our results are consistent with findings of Hsee (1998), that relative measures of some uncertain quantity (Hsee gives an example of the size of a food serving relative to its container) are more readily evaluated than absolute measures (the size of serving). An easy-to-evaluate attribute, such as the bounded correlation of [0,1], has a greater impact on a person's judgement of utility. Hsee (1998) describe this as the "relation-to-reference" attribute. It is therefore, not surprising that the offset correlation is highly-ranked." à As I have already explained above, correlation is not bounded between 0 and 1, and the fact that authors' failed to grasp this clearly indicates that it is not a simple "easy-to-evaluate" attribute. It will be helpful for readers if the greater impact of an easy-to-evaluate attribute on judgement of utility is explained, given that this seem to be one of the main conclusions from this study. It will also be useful if Hsee(1998) "relation-to-refence" attribute can be explained as to how it relates to this study.	The offset correlation is bounded between 0 and 1, *pace* the reviewer. This is not difficult, many correlation measures used in statistics are non-negative such as intra-class correlations or heritabilities. We have revise our explanation of the offset correlation L83 to L90 and L492 to L495 to make this explicit.
L394-395: "The offset correlation will be more useful for stakeholders who are not able to express their quality requirement for information in terms of quantities such as kriging variance." à Was this statement also derived from the feedback session? In which way will it be more useful?	We edited this to make this clear, see L343 to 344 (revised manuscript): The offset correlation seems to be a criterion which respondents are more likely to find comprehensible, and so a basis for selecting the sample density for a geostatistical survey, than alternatives such as kriging variance.
L395-399: "Furthermore, it is an intuitively meaningful measure of uncertainty, it recognises that spatial variation means that maps interpolated from offset grids will differ but that the more robust the sampling strategy the more consistent they will be. There is a paradox	We made the following change from L344 to 351 to reflect this. Furthermore, it appears to be a measure of uncertainty which participants in the study found comprehensible, and so were able to use to select a grid

here, however, in that the previous study Chagumaira et al. (2021) showed that interpretation of survey outputs in terms of uncertainty was easiest for stakeholders with measures related directly to a decision made with the information. The offset correlation is a general measure, and the absolute magnitude of uncertainty." à I wonder how offset correlation is an intuitive measure of uncertainty, can you please explain. And can you also explain the paradox you mention?	sample spacing. It recognises that spatial variation means that maps interpolated from offset grids will differ but that the more robust the sampling strategy the more consistent they will be. However, Chagumaira et al. (2021) found that measures of uncertainty related to a specific management threshold of the mapped variable were preferred by participants for the interpretation of uncertain spatial information to general quality measures without a specific management or policy implication. In this case, in contrast, the preferred criterion, the offset correlation, is a general measure of map quality, which is not directly linked to specific interpretation.
L403-411: Interesting explanation. I wonder if author's don't find it strange that the same people (stakeholders/participants) that could understand the bounded attribute [0,1] of the offset correlation cannot seem to understand a similar attribute of the conditional probabilities simply because it is "probabilities"?	We do not find it strange. Offset correlation [0,1] tells us that the maps made from offset grids are either completely independent of each other (0) or identical (1). In contrast the probability is (i) conditional on data and (ii) is a joint probability so it measures the overall probability of making a particular error in interpretation of the information: failure to recommend an intervention, resulting from (a) the uncertainty of the prediction and (b) the overall probability of the corresponding intervention being required.
L411-418: As I have already indicated in the results section the response on the grid spacing of 100 km is markedly lower than the rest, so I expected some explanation as to why this is so. The explanation given here is too superficial and inadequate to explain such a complex decision-making process.	Our analysis shows that there was no evidence to reject the null hypothesis that the responses are uniformly distributed see Table 3. The result is quite compatible with a random distribution. It would be potentially misleading to give an explanation to as why there are fewer responses on 100km, yet the evidence suggests these responses are uniformly distributed.
General comment: Based on the issues I've highlighted throughout my review, it's apparent that the remainder of the discussion and conclusion sections also warrant similar concerns. I strongly recommend a comprehensive revision of the manuscript to explicitly delineate its unique contributions to this most important field of science communication and particularly on communicating uncertainty.	We thank the referee for thorough review of our work, and we believe we have addressed all the concerns raised.

Referee Comments 2 (RC2):

Referee Comment	Author Comment
The work is nicely thought out and well written, and I think this will be useful to add to the literature about how easy users find it to understand uncertainty when presented in different ways. It also nicely demonstrates that different sampling densities would result if the different methods of communicating the uncertainties would be used. I have only very minor comments to add, plus some typos.	We would like to thank the referees for their thorough review of our manuscript. We wish to revise the manuscript based on their suggestions and comments.
I think the introduction should mention other mapping methods, which can make use of appropriate environmental covariates to model part of the variation, and then justify the focus of the work on kriging and gridded sample designs. If other mapping methods were used, optimal sampling schemes would then not be a grid, could the methods be applied to help in this case?	We have proposed to make changes in the introduction as we were responding to RC1 comments, see L22 to L96.
14: correlation is bounded on [-1,1]?	Correlations are bounded [-1,1] however, offset correlation 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. Therefore, we wish to make the following change in the manuscript from L492 to L495 to make it clearer for the reader: The offset correlation is bounded [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.
35: Clarify that different grid spacings is for the data (not grid of predictions).	We have revised the introduction to make this clear, please see L36 to 51.

42: Should this specifically say “ordinary kriging predictions” (ie not simple/regression/universal kriging)	We have revised the introduction to make this clear, please see L36 to 51.
Eq 2: I think z(x0) on right-hand side should be z(xi), also in the text below the equation.	Suggested edit has been made on the manuscript L421to L422 and on Equation A1: $$\tilde{Z}(\mathrm{x}_0) = \sum_{\mathrm{I}=1}^{\mathrm{N}} \lambda z(\mathrm{x}_\mathrm{i}),$$ where $z(\mathrm{x}_\mathrm{i})$ is the data and λ are the kriging weights (Webster and Oliver, 2007).
Eq 3: This is a repeat of Eq 1. Probably better here (ie after presenting formula for prediction), so suggest deleting Eq 1.	Suggested deletion was made.
103: definition of epsilon here should align with what is given on line 112 (probably line 112 version is better).	Suggested edit was made on L424 to L425: Cross-validation predictions of the statistical model need to be examined by exploratory analysis of the error of the kriging prediction, $\varepsilon(x_0)$, defined as $\{\tilde{Z}(x_0) - z(x_0)\}$, to check the if the assumption of normality holds.
134-136: I’m confused by this sentence. If z is 0 in Eq 6, then the true data does not appear in the equation, and the error doesn’t seem to matter, only the value of the prediction?	In the equation 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 (Z* - z). If the true value is 0 then the error is equal to Z*, that is not the same as saying that the data value has vanished from the error. We edited the sentences from L445 to L446 to make this clear.
162: I think it would be clearer to put “made from data collected on a square grid”	Suggested edit was made on L483. The expected correlation between the kriging predictions. $\tilde{Z}_1(x_0)$, made from data collected on a square grid, of interval ζ, and predictions, $\tilde{Z}_2(x_0)$, made from a second grid, a translation of the first grid by ζ/2 in both directions is known as the offset correlation.

230: Was the size of blocks the same for all grids, or was it set to be the same size as the grid? The second option (block size = grid size) doesn't really make sense to me in this context, as the meaning of the predictions (and their variances) would be different for the different sample designs.	We used the same size of blocks for all the grids, and it was 0.01 km^2 square block, and we made the following change in the manuscript on L189. We then computed the cell-centred block kriging variance the spacings we were considering by block kriging (Webster and Oliver, 2007). For all the grid spacings, we computed cell-centred block kriging variance on 0.01 km^2 square blocks.
Fig 4: Can brief details of how these maps be added? Were data for a pair of offsetted sampling grids jointly simulated, then those simulated data (for each of the two offsetted grids in turn) used to predict on the fine-scale mapping grid (as shown in the figure)?	The maps were generated to allow the participant to visualize what two spatial variables correlated by some specified amount would look like. They were generated as realizations of two coregionalized variables with the same mean and variance and the target correlation specified. We have added the following sentences to make this clear, please see: L178 to L179: We presented the participants with correlated pairs of hypothetical maps of soil pH and Segrain, with differing correlations, so that the extent to which maps might differ as a result of the grid offset could be visualized. L183 to L184: Neither of the posited pair of maps, based on offset grids, is to be regarded as closer to reality than the other, the question is how consistent they are. Caption of Figure 1 (L188): Figure 1. The pair of hypothetical maps of pH value and corresponding scatter-plot for offset correlation 0.4.

450: Needs a sentence added here to summarise what you did (before talking about responses in the next sentence).	We have added the following sentence from L404 as suggested by the referee: In this study we evaluated four methods of communicating uncertainty associated with kriging predictions made from data from a geostatistical survey, to determine an appropriate sampling density to meet stakeholders expectations.
Tables A2 and A3: these could easily be combined into one table	Tables B2 (revised manuscript in the Appendix, from page 23) shows the subtable for responses when pooled within the variable used (soil pH and Se_{grain}) when partitioning the full table as illustrated on Table B1. Then table B3 shows the pooled counts of response for offset correlation after all the partitioning for Q1, and this was used to examine if the responses were uniformly distributed.
Line number: suggested new text: 7: “from data on sample grids…”	Suggested was made on the manuscript L11: …. the correlation is a measure of the consistency of kriging predictions made from data on sample grids with the same spacing but different origins
Line number: suggested new text: 20: “concentrations”	Suggested was made on the manuscript L103 to L104: It has been shown that concentrations of micronutrients in staple crops and in soils vary spatially, as do biomarkers for micronutrient status and so interventions to address the deficiencies should be based on spatial information on all these variables (Gashu et al., 2021; Botoman et al., 2022).
Line number: suggested new text: 26: “survey efforts”	We revised this text to a new sentence from L31 to L33: These surveys have costs, for example, travel and logistics, staff costs, time for community engagement, management costs and analytical costs for processing material collected in the field.
Line number: suggested new text:	We have revised the introduction to make this clear, please see L37 to 51.

33: “is quantified”	
Line number: suggested new text: 53: “tied to particular decisions”	We have revised the introduction to make this clear, please see L64 to 71.
Line number: suggested new text: 61: delete comma after x0	Suggested was made on the manuscript L61: For a conservative measure of uncertainty, x_0 may be at a general location where uncertainty is largest e.g., at the centre of a square grid cell.
Line number: suggested new text: 67 “to an acceptable”	We have revised the introduction and replaced this statement, see L67 to L69: The second measure was based on the joint probability that a location requires some intervention (because the surveyed variable exceeds or falls below a threshold) and that the spatial prediction at that location indicates the contrary.
Line number: suggested new text: 79: “or from a comparable region”	We have revised the introduction and replaced this text with the paragraph from L72 to 82.
Line number: suggested new text: 138: “reduces the error”	Suggested was made on the manuscript L459: Increasing sample size reduces the minimum expected loss in so far as it reduces the error variance.
Line number: suggested new text: 139: “an additional sample point”	Suggested was made on the manuscript L460: Therefore, the cost of obtaining *n* samples can be measured at which the marginal cost of an additional sample point.

Line number: suggested new text: 145: “the sampling exercise”	Suggested made on the manuscript L464: The implicit loss function is conditional on a logistic model, that expresses the marginal costs of the sampling exercise
Line number: suggested new text: 149: “number of samples”	Suggested was made on the manuscript L470: where $\bar{n}$ is the specified number of samples, C(n) is the function that returns the cost of n samples and ϕ is a vector of variogram
Line number: suggested new text: 155: “denotes”	Suggested was made on the manuscript L476: Where, F^{-1} denotes the quantile of the prediction distribution for a probability P_0 obtained from
Line number: suggested new text: 158: “loss functions”	Suggested was made on the manuscript L478: Lark and Knights (2015) suggested that a stakeholder group might consider an implicit loss function for different $\bar{n}$ 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.
Line number: suggested new text: 231: “three different predictions”	Suggested was made on the manuscript L190: We considered three different predictions for each variable, but the prediction interval was fixed, depending only on grid spacing.
Line number: suggested new text: 244: “asked the participant what grid”	Suggested was made on the manuscript L203: We then asked the participant what grid spacing they thought corresponded to the largest acceptable value of this probability.
Line number: suggested new text: 267: “pair of maps”	Suggested was made on the manuscript L180: Figure 1, shows an example of pair of maps of Se_{grain} concentration and the corresponding scatterplot (see Figure S5 and S6).
Line number: suggested new text:	Suggested was made on the manuscript L501:

289: “partitioned into components corresponding to a pooled table”. 291: I think “Figure 3” should be “Table 2”	The full table in Table B1, was partitioned into components corresponding to subtables for soil pH (subtable 1 in Table B1), and Segrain concentration (Subtable 2 in Table B1).
Line number: suggested new text: 302: “differences in responses”	Suggested was made on the manuscript L251: We first tested for differences in responses recorded for each test method, by the variable used (soil pH or Se_{grain} concentration) using contingency tables.
Line number: suggested new text: 311: “no difference”	Suggested was made on the manuscript L256: For some questions, we noted differences in the responses when pooled within variable used (soil pH or Segrain concentration) and there was no difference in responses in professional groups and frequency of use of statistics for all
Line number: suggested new text: 312: “were uniformly”	Suggested was made on the manuscript L261: We further analysed the pooled tables or separate subtables to examine if the responses were uniformly distributed and the null hypothesis is a random distribution.
329: “responses of the?” Missing something here.	We have made the following edit on L274: There were no differences in the responses when the columns were pooled by the variable used, soil pH vs. Se_{grain} concentration, p = 0.656 (Table 2).
Line number: suggested new text: 372: I don’t think this should be “by all respondents” (ie not every single respondent ranked it first?), maybe should be “Amongst all respondents, the offset…most effective”	Suggested was made on the manuscript L320 to L321: Amongst all the respondents, the offset correlation was ranked as the most effective (Figure 9a) and implicit loss function as the least effective.
Line number: suggested new text: 380: “statistics”	Suggested was made on the manuscript L326: Those who always use statistics, ranked conditional probabilities second.

Line number: suggested new text: 388: “explains why there were”	We revised this statement to be more clear on L334 to L335: This is likely to explain the consistency of the results for this criterion, with over half the respondents selection 0.7 or 0.8 as a minimum acceptable correlation.
Line number: suggested new text: 404: “This suggests”	Suggested was made on the manuscript L350: This suggests that the stakeholders may not have fully understood the method.
Line number: suggested new text: 417: “by the respondents” and “would be of greatest value”	Suggested was made on the manuscript L366 to L367: Similar reasons were given by the respondents. We expected that prediction intervals would be of greatest value for specific interpretation of particular sites but would be of limited value for survey planning.
Line number: suggested new text: 432: “beginning” and “explanation of”	Suggested was made on the manuscript L381 to L383: At the beginning of the online workshop, we explained each method with the aid of illustrations. After an explanation of each method, there was a feedback session to allow the participants opportunities to seek clarity on ambiguous and unfamiliar concepts from the presenters.
Line number: suggested new text: 441: “different variables”	Suggested was made on the manuscript L390: All the methods may give different results for different variables, because they depend on the variogram of the variable in question.