Solid Earth, 5, 1189–1203, 2014
www.solid-earth.net/5/1189/2014/
doi:10.5194/se-5-1189-2014

Solid Earth Open Access

Interpretative modelling of a geological cross section from boreholes: sources of uncertainty and their quantification

R. M. Lark, S. Thorpe, H. Kessler, and S. J. Mathers

British Geological Survey, Keyworth, Nottingham, UK

Correspondence to: R. M. Lark (mlark@bgs.ac.uk)

Received: 28 May 2014 – Published in Solid Earth Discuss.: 17 July 2014
Revised: 20 October 2014 – Accepted: 23 October 2014 – Published: 28 November 2014

Abstract. We conducted a designed experiment to quantify sources of uncertainty in geologists' interpretations of a geological cross section. A group of 28 geologists participated in the experiment. Each interpreted borehole record included up to three Palaeogene bedrock units, including the target unit for the experiment: the London Clay. The set of boreholes was divided into batches from which validation boreholes had been withheld; as a result, we obtained 129 point comparisons between the interpreted elevation of the base of the London Clay and its observed elevation in a borehole not used for that particular interpretation. Analysis of the results showed good general agreement between the observed and interpreted elevations, with no evidence of systematic bias. Between-site variation of the interpretation error was spatially correlated, and the variance appeared to be stationary. The between-geologist component of variance was smaller overall, and depended on the distance to the nearest borehole. There was also evidence that the between-geologist variance depends on the degree of experience of the individual. We used the statistical model of interpretation error to compute confidence intervals for any one interpretation of the base of the London Clay on the cross section, and to provide uncertainty measures for decision support in a hypothetical route-planning process. The statistical model could also be used to quantify error propagation in a full 3-D geological model produced from interpreted cross sections.

1 Introduction

Three-dimensional (3-D) models are now the state of the art for presenting geologists' knowledge and interpretation of subsurface structures, and are supplied to varied users of geological information. There is no single methodology for the production of models, and the method will reflect the geological setting and the nature of the information available to the modeller, which may include geophysical imagery, boreholes and surface observations. Models can be produced by geostatistical interpolation (e.g. Lark and Webster, 2006) or by a combination of geostatistical methods with expert intervention to ensure geologically realistic results (e.g. Gunnink et al., 2013). Models may also be based on inversions of geophysical data, constrained by geological knowledge and interpretation (Jessell et al., 2010). The approach of particular interest here is based on expert interpretation of boreholes along interlocking sets of cross sections with subsequent interpolation from the interpreted sections to produce models of volumes in 3-D. This is exemplified by the GSI3D software (Kessler and Mathers, 2004; Kessler et al., 2009). Expert interpretation of a cross section entails the interpretation of boreholes and the sequential construction of the basal contact of each geological unit in the stack. This process depends on the expert interpretation of boreholes in line with rules, explicit or tacit, which control the shapes of surfaces and the circumstances in which faults must be invoked to explain their observed positions. Because these rules encapsulate geological knowledge, they provide a sound basis for modelling, particularly when limited observations are available. However, the interpretation of the cross sections inevitably has an attendant uncertainty, and this is propagated when the interpreted cross sections are combined to model volumes in 3-D by interpolation.

The uncertainty in a 3-D model is of interest to data users who will apply it for decision making. For this reason, there has been considerable interest in the development of quantitative or semi-quantitative operational methods to

characterise the uncertainty in 3-D models and the variation of this uncertainty in space (e.g. Lelliott et al., 2009).

If information in 3-D is produced by geostatistical interpolation, then the uncertainty can be quantified directly on the basis of the geostatistical model (Lark and Webster, 2006). In the study reported by Gunnink et al. (2013), the geostatistical predictions were modified to ensure geological consistency. The original geostatistical measures of uncertainty no longer hold for the modified values, so Gunnink et al. (2013) used a cross-validation method to quantify uncertainty. However, this is feasible only if many borehole observations are available. Bistacchi et al. (2008) present a case study where the uncertainty in the modelled position of planar surfaces in the 3-D space could be computed from information about the uncertainty of the angular observations on which the model was based, and the distance over which these observations were projected. Tacher et al. (2006) used the simple kriging variance as a measure of uncertainty for the position of modelled geological surfaces, the parameters of the variogram being informally elicited to reflect expert judgement about uncertainty and its spatial dependence. In many cases 3-D modelling is supported by interpretation of geophysical data. Bond et al. (2007) and Torvela and Bond (2011) examine the uncertainties in expert interpretation of seismic imagery, and particularly how uncertainties in the conceptual geological model which underly the interpretations contribute to the final uncertainty. Aitken et al. (2013) discuss a measure of "data richness" to quantify the extent to which the geological interpretability of geophysical data, the complexity of these data and their quality determine the uncertainty of resulting models, and the variation of this uncertainty in space.

In this paper, we are particularly interested in the uncertainty of models produced by the cross-section interpretation methodology. Lark et al. (2013) made a direct empirical assessment of the quality of one such model in a designed experiment. They compared the predicted heights of units with observed heights at a set of validation boreholes. This gave a quantitative measure of uncertainty. However, Lark et al. (2013) concluded that it is necessary to understand how error enters into the initial interpretation of cross sections prior to interpolation, since the errors in the cross sections may be predictable from factors such as the distance to boreholes or crop lines, but propagated into the 3-D model by the interpolation step in a complex way. If we can understand and quantify the uncertainty in cross-section interpretation, then it may be possible to develop quantitative models of the uncertainty for different "benchmark" geological settings, and to use this to develop uncertainty measures for application to geological models.

To this end, we undertook, and report here, an experiment to study the error in cross-section interpretation, hypothesizing that the variability of the interpretation error changes along the section in ways that can be described by a statistical model. We considered statistical models in which the variance of the interpretation error at some location depends on two factors. The first factor was the distance from the location to the nearest borehole available to support the interpretation of the cross section. Our hypothesis was that the variance of interpretation error would increase with the distance to the nearest available borehole. The second factor was the experience of the geologist making the interpretation; our hypothesis was that the variance of interpretation error would diminish with increasing geologist experience.

If our hypothesis is verified, then we could compute confidence intervals for the interpreted height of a contact along a cross section, and model how this uncertainty may propagate in the subsequent interpolation from the interpreted cross section into a 3-D geological model. If statistical models of the uncertainty in cross-section interpretation could be estimated for a variety of geological settings, then these could be used to compute uncertainty measures for new geological models, and so to calculate, for example, decision–theoretical measures of the value of the model information (Howard, 1966) or other criteria by which model users can make rational decisions that account for model uncertainty.

2 Methods

2.1 Geological context of the cross section

This study is based on an 8 km cross section in London which roughly follows the A12 road from Hackney northeast across the Lea Valley to Wanstead. The local geology (Fig. 1) consists of Quaternary deposits comprising alluvium along the valleys of the rivers Lea and Roding, with river terrace deposits at several levels beneath and flanking the alluvium and capping the low interfluvial ridge.

The Quaternary deposits are generally less than 5 m in total thickness, except beneath the Lea Valley, where up to 10 m are encountered. They rest everywhere on Palaeogene bedrock units. In order of increasing age and depth, these are the London Clay Formation, the Lambeth Group and the Thanet Formation. The Quaternary deposits rest on the London Clay Formation along part of the section, but cut down beneath the Lea Valley into the underlying Lambeth Group (Fig. 1). The Palaeogene deposits are underlain by the Chalk Group (Upper Cretaceous), which is several hundred metres thick and is the lowest unit considered that is encountered here in approximately 10 % of the 143 available boreholes along the cross section (Fig. 1).

The Palaeogene strata in this region are affected by the Alpine Orogeny, and underwent gentle folding, faulting and tilting in Oligocene–Miocene times (Sumbler, 1996). In this study, our interest lies in the definition of the base of the London Clay Formation. In the London area, the London Clay comprises a grey marine silty clay with thin interbeds of sandy clay, sand and pebble beds (Ellison et al., 2004). The whole sequence locally exceeds 100 m in thickness, although

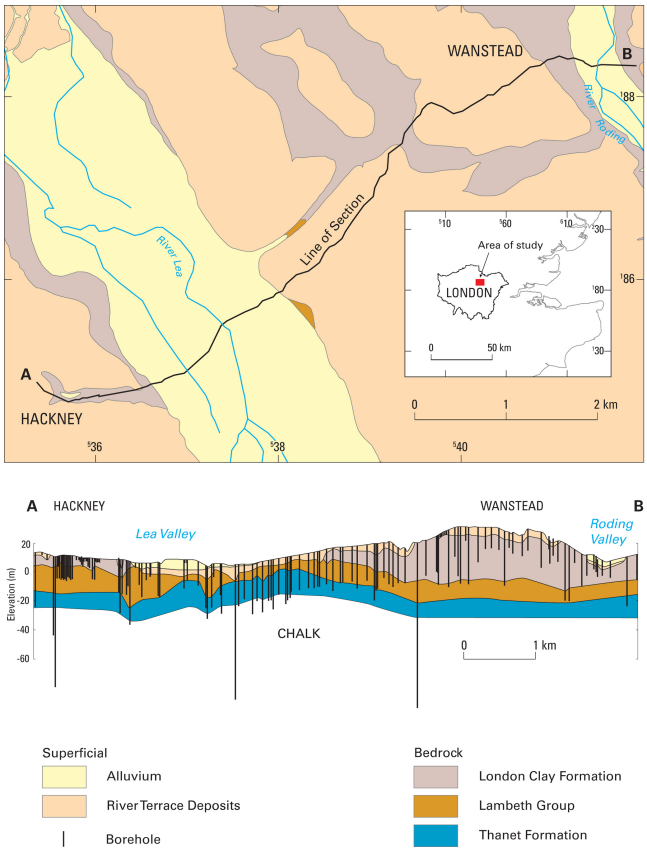

Figure 1. Map of surface geology (superficial and exposed bedrock) in the study area, with the line of the cross section shown. Map coordinates are in km on the British National Grid. One interpretation of the cross section is shown below, with the position and depth of the full borehole set indicated.

due to erosion, considerably less is preserved along the line of the cross section discussed here. It rests conformably on the Lambeth Group, which consists of about 15 m of interbedded colour mottled clays, sands and silts arranged in a complex and variable vertical sequence of facies (Ellison, 1983). The London Clay Formation is present over about 60 % of the length of the cross section, and the base of this unit, the surface of interest, is proven by 51 of the boreholes. Along the section, the elevation of the base of the London Clay Formation observed in the boreholes varies from nearly 10 m to −13 m relative to the Ordnance Datum (Fig. 1).

In the London area, the London Clay Formation is a relatively thick firm clay without significant water flow, and it is therefore regarded as a good medium for tunnels and excavations (Ellison et al., 2004). The Lambeth Group, by contrast, contains thin layers of alternating clay and water bearing soft sands and silts, and the clays are also characterised by a strong propensity to shrink-swell during cycles of wetting and drying. As such, it consists of a very difficult medium for excavation and tunnelling, and is best avoided wherever this is possible (Ellison et al., 2004). As the cross section demonstrates, parts of London are underlain at a few tens of metres in depth by these two units (Fig. 1). Hence, the position of the base of the London Clay Formation is critical, as it separates these two units of radically different engineering behaviour, and the measures of uncertainty derived in this study have considerable potential for application in this context.

2.2 Data subsetting, geologists' self-assessment, and modelling

The key idea of the experiment was that each of a set of participating geologists would make an interpretation of the three Palaeogene bedrock units on the cross section, drawing continuous (if occasionally interrupted) basal contacts of the units as interpretations of the information in a set of boreholes. Any one participant would use a subset of all available boreholes, so that their interpretation could be compared directly with each of a complementary validation subset. The difference between the interpreted and observed elevations of the base of the London Clay, the cross-section error, would then be treated as a variable for statistical analysis to identify important features of its variability. Note that, while we only examined the base of the London Clay, the participants interpreted this in the wider stratigraphical context by also drawing the bases of the other Palaeogene units.

The 51 available boreholes which prove the base of the London Clay were subdivided by independent random sampling without replacement into ten non-overlapping subsets of five validation boreholes. We call each of these subsets a *validation batch*; each is paired with its corresponding *interpretation batch* – the complementary subset of 46 boreholes. In this way, ten different although overlapping interpretation batches, each with 46 boreholes which proved the base of the London Clay, were prepared for use by geologists in the experiment. Any one participant would use just one interpretation batch. His or her interpretation of the cross section could then be compared with the five boreholes in the corresponding validation batch, boreholes not used in the interpretation of the cross section, to generate five observations of cross-section error.

A total of 28 geologists participated in the experiment. Of these, 22 were delegates at the GSI3D workshop which took place at the British Geological Survey (BGS), Keyworth, from 17 to 18 October 2012, and the GSI3D software was used for the experiment. Some of the workshop participants were staff of BGS, others were geologists from a variety of organisations and countries, with varying levels of experience in geological modelling, but all with some interest and experience, if rudimentary, in the use of the GSI3D software, which was used for this experiment by all participants. The remaining six geologists were BGS staff who participated in the experiment after the workshop.

Each participant was asked to complete a questionnaire before undertaking the exercise. Their unique number was recorded on the form. They had the option of

Table 1. Questionnaire on modelling experience and responses received.

Question: "Please indicate with a tick which of the 4 descriptions below best reflects your experience of 3-D modelling."	
Description	Number of participants selecting this description
I have no experience of geological modelling in 3-D	2
I have some experience of geological modelling in 3-D (perhaps through a training course) but little (up to 6 months) or no experience of modelling independently	8
I have moderate experience of geological modelling in 3-D (six months to 2 years of modelling independently)	8
I have substantial experience of geological modelling in 3-D (more than 2 years of modelling independently)	10

recording their name and contact details on the form, or of remaininganonymous. In the questionnaire, each participant was asked to record a self-assessment of their experience of geological modelling in 3-D by identifying the most appropriate of four general descriptions. The descriptions and responses are presented in Table 1. Note that there was some variation in experience among the participants: two were novices in 3-D modelling, and eight had limited experience. This allows us to quantify the effect of increasing experience on the variability of interpretation error.

The key principle of the experiment was explained to all delegates, who were also provided with an explanation of the units in the cross section. Each participant in the experiment, on presenting at the workstations, was given a unique number, and an interpretation batch of boreholes. In addition to the boreholes, a standard interpretation of the superficial material (as a single unit) was provided, so that all participants were working on a common rockhead surface. The intersections of outcrops, as mapped in 2-D, with the cross section were also provided to all participants. A set of guidance notes on the GSI3D software was available, and at all times a staff member experienced with the software was available to help. When the interpretation was complete it was saved with a code which indicated the participant's unique number and the number of the interpretation batch and complementary validation batch of boreholes which had been allocated. As each geologist presented to participate, they were allocated one of the interpretation batches of boreholes, so that a more or less even distribution of participants over batches was achieved.

Once each geologist had completed and saved their interpretation, this was compared with the corresponding batch of validation boreholes, and the observed and interpreted elevation of the base of the London Clay was extracted. One modeller's interpretation was not correctly saved, so this was lost, and in some cases the London Clay was not present in the interpretation at the location of a validation borehole. Over all validation batches, we were able to make a total of 129 comparisons between an interpreted elevation of the base of the London Clay at the location of a borehole in a validation batch observed elevation in that validation borehole (i.e. in a borehole which had not been available to the geologist who made the particular interpretation). As described in Sect. 3.1 below, and formalised in Eq. (1), an observation of interpretation error is the difference between the interpreted and observed elevation of the base of the London Clay for one such comparison. Between 10 and 20 interpretation errors could be calculated for any validation batch.

3 Data analysis

3.1 Overview of models and analyses

This section provides an overview of the analyses undertaken to test our hypothesis, avoiding the statistical detail. The reader will find technical information about the statistical models and their estimation in Sects 3.2–3.3, and these can be ignored by the reader who requires only a summary of the statistical methods. Section 3.4 explains how the selected statistical model for cross-section errors was interrogated to represent the cross-section uncertainty with confidence intervals and an analysis of the implications of this uncertainty for a hypothetical application.

As reported in the previous section, the experimental results consist of a set of 129 comparisons between the interpreted and observed elevations of the base of the London Clay, where each interpretation in the set had been made without access to that particular observation. The variable for statistical analysis is the cross-section error, obtained for each of the 129 comparisons by subtracting the interpreted elevation of the base from the observed elevation. An error of zero therefore means that the observed and interpreted elevations were the same in the particular comparison. A negative error means that the interpreted base was higher than the observed base in that comparison.

The statistical analysis of these values was done with linear mixed models. These treat the cross-section errors as a combination of a fixed effect (here a constant, the mean cross-section error) with random effects. The random effects represent sources of variation in the observed errors, and here account for differences between batches of validation boreholes (are the mean errors for the different batches significantly different?), between the sites of validation boreholes within batches (are the mean errors for different locations within each batch significantly different from each other?), and between the geologists. The means of the random effects are zero; their variances are interesting because they quantify the uncertainty introduced into the interpretation of the cross section by the factors which they represent (differences between modellers, differences between locations). In some of the more complex models, we used the variance of a random effect that was modelled as a function of some covariate. For example, in one case, the variance of the effect of location was modelled as a function of the distance from the location to the nearest borehole available for interpretation (i.e. the nearest borehole in the interpretation batch allocated to the particular geologist). Such models could be used to predict how interpretation uncertainty varies along a cross section.

We considered seven linear mixed models which were fitted in order, so in some cases a statistical inference about one model (i.e. showing that a particular random effect was not significant) determined the form of subsequent models (that effect was dropped).

The random effects which we considered can be defined with respect to two properties. The first is dependency. If a random effect is independent, then the value that it takes for one instance tells us nothing about the value that it takes in other instances. In the first model, 1a, the random effect that models differences between batches was independent, because the batches were formed by independent random sampling. In other models, a random effect may not be independent, but may have a correlation structure. In all models, the random effect that models differences between sites had a spatial correlation structure: one might expect cross-section errors at two nearby sites to be more similar than errors at two sites which are far apart. In models 1a and 1b, the random effect which accounts for variability of geologist interpretations was independent within any site (the effect for one geologist is independent of the effect for another), but the cross-section errors for any one geologist at different sites were modelled as correlated (a geologist who tends to interpret the base too high at one site might make a similar error at other sites).

The second property of random effects is stationarity in the variance (stationarity hereafter). A stationary random effect has a constant variance. However, the variance of a non-stationary random effect may be modelled as a variable which depends on some other factor. For example, in model 2a, the variance of the geologist random effect depends on the level of experience that each geologist recorded in the questionnaire (Table 1).

Table 2 summarises the differences between the models. Mode 1a is a general one in which there are stationary random effects for batch, site and geologist differences. The batch effect is also independent, the site effect is spatially correlated (as in all models) and the geologist effect shows correlation between errors made by the same geologist. Models 1b and 1c were fitted to test, respectively, whether the variance of the batch effect could be assumed to be zero and whether the geologist random effect could be modelled as independent. The final model in group 1, 1d, was meant to see whether the variance of the site effect was non-stationary, depending on the distance to the nearest available borehole.

In all the models in a second group of three, the batch effect was dropped, and the site effect was spatially correlated and stationary. The geologist effect was independent, but we considered non-stationary alternatives in which the variance depended on (2a) the distance to the nearest borehole available for interpretation, (2b) modeller self-identified experience, and (2c) both these factors.

We compared models in two ways (details in Sect. 3.2). In some cases, it was possible to compare models by a log-likelihood ratio statistic L. These are presented in Tables 4 and 5 for comparisons where they can be made. In each case the compared models are indicated and the statistic presents the strength of evidence for the effect of additional terms in the more complex model. The recorded p value is the probability of finding evidence as strong or stronger than the value of L if the simpler model were true. If p is larger than 0.05, we retain the simpler model. Not all models can be compared this way, and where the log-likelihood ratio statistic could not be used, we compared models by Akaike's information criterion (AIC, details in Sect. 3.3). In any comparison, the model for which AIC is smallest was selected. The AIC is not a formal significance test, but by selecting the model with smaller AIC, one minimises the expected information loss through the selection decision (Verbeke and Molenberghs, 2000). A summary of the key comparisons between models, and the inferences made from them, is provided in Table 6.

3.2 Statistical methodology: linear mixed models, the general model (1a), and three variants

The results from this experiment were analysed by the fitting and comparison of linear mixed models (LMM) (Verbeke and Molenberghs, 2000) for the cross-section errors. One observation of cross-section error corresponds to a particular geologist's interpretation at one of the sites in the validation batch corresponding to the interpretation batch to which that geologist had been allocated. The interpretation at that site had therefore been made without access to the information in the borehole there. This gives us a total of $N = 129$ observations of cross-section error. If the interpreted elevation of the base of the London Clay by geologist m at site

Table 2. Summary of statistical models. In all cases the form of the random effects component for between-batch, between-site and between-geologist effects is indicated. For each term the dependency is given (independent or an indicated correlation structure) and it is indicated whether the variance is constant (stationary) or modelled as a variable quantity. A ↓ indicates that a term is dropped from the model.

Model	Batch		Site		Geologist	
	Dependency	Variance	Dependency	Variance	Dependency	Variance
1a	Independent	Stationary	Spatially correlated	Stationary	Correlated[1]	Stationary
1b	↓		Spatially correlated	Stationary	Correlated[1]	Stationary
1c	↓		Spatially correlated	Stationary	Independent	Stationary
1d	↓		Spatially correlated	DNAB[2]	Independent	Stationary
2a	↓		Spatially correlated	Stationary	Independent	DNAB[2]
2b	↓		Spatially correlated	Stationary	Independent	Experience[3]
2c	↓		Spatially correlated	Stationary	Independent	DNAB + Experience

[1] Errors of interpretations by the same geologist are correlated. [2] Variance depends on distance to nearest available borehole for interpretation. [3] Variance depends on geologists self-identified experience of 3-D modelling (Table 1).

k within batch i is $z^{\mathrm{s}}(\mathrm{b}_i, \mathrm{s}_k, \mathrm{g}_m)$, and the corresponding observed elevation in the validation borehole is $z^{\mathrm{o}}(\mathrm{b}_i, \mathrm{s}_k)$, then the correspondingobservation of cross-section error is defined as

$$\varepsilon(\mathrm{b}_i, \mathrm{s}_k, \mathrm{g}_m) = z^{\mathrm{o}}(\mathrm{b}_i, \mathrm{s}_k) - z^{\mathrm{s}}(\mathrm{b}_i, \mathrm{s}_k, \mathrm{g}_m). \quad (1)$$

A negative error therefore means that the geologist's interpretation is higher than the observed elevation of the base of the London Clay.

The fixed effect in all LMM that were considered here was the mean cross-section error. The random effects modelled the contribution of differences between batches, differences between sites and differences between geologists. In an LMM, the random effects are modelled as Gaussian random variables with mean zero and a variance. The variance may be stationary, a parameter of the LMM, or it may be a variable expressed as a parametric function of some covariate with parameters to be estimated (e.g. Nelder and Lee, 1998; Lark, 2009). The random effects and their parameters are of interest because they may be informative about sources of cross-section error, and allow us to predict cross-section error variance in similar settings. Once an appropriate model for the random effects has been selected, one may use generalised least squares to estimate the overall mean model error and test whether it appears to be significantly different from zero.

Model 1a takes the following form for a set of observations of cross-section error in a vector $\boldsymbol{\varepsilon}$ of length N:

$$\varepsilon = \mathbf{M}\alpha + \mathbf{X}_{\mathrm{b}}\beta_{\mathrm{b}} + \mathbf{X}_{\mathrm{s}}\mathbf{Z}_{\mathrm{s}} + \eta_{\mathrm{g}}, \quad (2)$$

where $\mathbf{M}$ is an $N \times p$ design matrix that associates each observation of cross-section error in $\boldsymbol{\varepsilon}$ with a value of a fixed effect variable, contained in the vector $\boldsymbol{\alpha}$ of length p. In all models considered in this paper, the fixed effect is a constant, the mean cross-section error, so $p = 1$, $\boldsymbol{\alpha}$ contains the mean and $\mathbf{M}$ is an $N \times 1$ vector of ones. Other terms in the model are explained in the following paragraphs.

The matrix $\mathbf{X}_{\mathrm{b}}$ is an $N \times N_{\mathrm{b}}$ design matrix for the between-batch random effect where N_{b} is the number of batches. Row n of $\mathbf{X}_{\mathrm{b}}$ corresponds to the nth observation of cross-section error. If the nth observation of cross-section error belongs to the mth batch out of N_{b}, then the element in column m of row n of $\mathbf{X}_{\mathrm{b}}$ is one, and all other elements in the row are zero. The vector $\boldsymbol{\beta}_{\mathrm{b}}$ is an $N_{\mathrm{b}} \times 1$ vector which contains the mean errors for the batches, which are treated as random variables. One may write down an expression for the covariance matrix of the N between-batch components of the observed cross-section errors, $\mathbf{C}_{\mathrm{b}}$.

Because the sites are randomly allocated to batches, it is assumed that the batch effects are independent, and so

$$\begin{aligned}\mathbf{C}_\mathrm{b} &= \sigma_\mathrm{b}^2\mathbf{R}_\mathrm{b}\\ &= \sigma_\mathrm{b}^2\mathbf{X}_\mathrm{b}\mathbf{X}_\mathrm{b}^\mathrm{T},\end{aligned} \tag{3}$$

where σ_b^2 is the variance of the batch effect, and $\mathbf{R}_\mathrm{b}$ denotes the correlation matrix of batch effects which is obtained, given the assumptions of independence, as the product of the batch design matrix with its transpose (denoted by the superscript T).

The term $\mathbf{X}_\mathrm{s}$ is an $N \times N_\mathrm{s}$ design matrix which associates each of the N observations of cross-section error with one of the N_s validation sites. These sites are not assumed to be independent of each other, since they were chosen by purposive sampling, and not by an independent random sampling design. Since the sampling design does not allow us to treat the between-site effect as an independent random variable, we rather invoke a random statistical model of the between-site effect (de Gruijter et al., 2006). The random variable, which is contained in the length-N_s random vector $\mathbf{Z}_\mathrm{s}$ is assumed to be N_s-variate Gaussian with mean zero and $N_\mathrm{s} \times N_\mathrm{s}$ covariance matrix $\mathbf{S}$:

$$\mathbf{Z}_\mathrm{s} \sim \mathcal{N}(\mathbf{0}_{N_\mathrm{s}}, \mathbf{S}), \tag{4}$$

where $\mathbf{0}_{N_\mathrm{s}}$ is a vector length N_s of zeroes. We assume that $\mathbf{Z}_\mathrm{s}$ is a second-order stationary random variable, so that the covariance of the values at any two sites depends only on the interval in space between those sites (Stein, 1999). Here we use a standard covariance function from geostatistics, the Matérn function (Matérn, 1986). Under this model, the covariance between two locations separated by distance d is

$$\begin{aligned}C(d) &= c_0 + c_1, \qquad d = 0\\ &= c_1\left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1}\left(\frac{d}{\phi}\right)^{\kappa} K_\kappa\left(\frac{d}{\phi}\right), \quad d > 0,\end{aligned} \tag{5}$$

where $K_\kappa(\cdot)$ is a modified Bessel function of order κ, κ is a smoothness parameter – see Diggle and Ribeiro (2007) for a discussion – ϕ is a distance parameter, and c_0 and c_1 are, respectively, the spatially uncorrelated and correlated components of variance of the between-site variable. Note that, while in principle, the covariance can be modelled as a function of the direction as well as the length of the separation vector between locations, when our observations of cross-section error are aligned on an almost-straight cross section, we consider distance only.

If the distance between site k in batch i and site l in batch j is denoted by $d_{\{i,k\},\{j,l\}}$, then one may compute a between-site covariance matrix $\mathbf{S}$, which is an $N_\mathrm{s} \times N_\mathrm{s}$ matrix. If the rth out of N_s sites is site k in batch i, and the cth out of N_s sites is site l in batch j, then

$$\mathbf{S}[r,c] \;=\; C(d_{\{i,k\},\{j,l\}}), \tag{6}$$

and the $N \times N$ between-sites effect covariance matrix for the LMM for all N observations of cross-section error is given by

$$\mathbf{C}_\mathrm{s} = \mathbf{X}_\mathrm{s}\mathbf{S}\mathbf{X}_\mathrm{s}^\mathrm{T}, \tag{7}$$

where $\mathbf{X}_\mathrm{s}$ is the $N \times N_\mathrm{s}$ design matrix for sites. Given the site design matrix, and the distances among the observations of cross-section error, this covariance matrix is determined by the four parameters of the Matérn covariance function: c_0, c_1, κ and ϕ.

The geologist effect in model 1a, the term $\boldsymbol{\eta}_\mathrm{g}$ in Eq. (2), is somewhat more complex. At each site within a batch, a cross-section error is observed for each geologist who was allocated the corresponding batch of boreholes. The term $\boldsymbol{\eta}_\mathrm{g}$ is the difference between the cross-section error for a particular geologist at a particular site, and the mean cross-section error at that site. It is therefore the between-geologist within-site effect, but we call it the geologist effect for brevity.

If each geologist had one and only one validation borehole, then the geologist effect would be simply nested within sites as an independent random error (regardless of whether there was one or more observations of cross-section error at each validation site). However, in the current experiment, each of the geologists was allocated all validation boreholes in a particular batch, and so we must choose an appropriate statistical model for the between-geologist effect observed at each of a set of boreholes. In model 1a, we treat the geologist effects as correlated random variables within batches. If we denote by $\bar{\varepsilon}(\mathrm{b}_i, \mathrm{s}_k)$ the mean cross-section error at site k in batch i, the geologist effect for geologist m at the same site is

$$\eta(\mathrm{b}_i, \mathrm{s}_k, \mathrm{g}_m) = \varepsilon(\mathrm{b}_i, \mathrm{s}_k, \mathrm{g}_m) - \bar{\varepsilon}(\mathrm{b}_i, \mathrm{s}_k). \tag{8}$$

In the random effects component of model 1a, we assume that the correlation of the geologist effects is

$$\begin{aligned}\mathrm{Corr}\{\eta(\mathrm{b}_i, \mathrm{s}_k, \mathrm{g}_m), \eta(\mathrm{b}_j, \mathrm{s}_l, \mathrm{g}_n)\} &= 1, \quad i = j, k = l, m = n\\ &= \rho, \quad i = j, k \neq l, m = n\\ &= 0, \quad \text{otherwise.}\end{aligned} \tag{9}$$

In words, the geologist effects for observations of cross-section error at two different sites are uncorrelated if the geologists are different (which includes all between-batch comparisons), and have a correlation of ρ if the geologist is the same. The covariance matrix for the geologist effect in model 1a is therefore

$$\mathbf{C}_\mathrm{g} = \sigma_g^2\mathbf{R}_\mathrm{g}, \tag{10}$$

where $\mathbf{R}_\mathrm{g}$ is an $N \times N$ correlation matrix of geologist effects with values 1 on the main diagonal, ρ on off-diagonal elements which correspond to pairs of cross-section errors corresponding to the same geologist, and zero in all other elements. The variance of the between-geologist effect is σ_g^2.

The random effects of the model in Eq. (2) are characterised by the between-batch variance, σ_b^2, the four parameters of the Matérn covariance model for the between-site variable (c_0, c_1, κ and ϕ), the between-geologist within site variance σ_g^2 and the correlation parameter ρ. We used residual maximum likelihood (REML) to estimate these parameters (Patterson and Thompson, 1971; Smyth and Verbyla, 1996). This proceeds on the assumption that $\boldsymbol{\varepsilon}$ in Eq. (2) is a realisation of a multivariate Gaussian random variable, $\mathbf{E}$:

$$\mathbf{E} \sim \mathcal{N}(\mathbf{M}\alpha, \mathbf{V}) \tag{11}$$

where $\mathbf{V}$ is the covariance matrix given by

$$\mathbf{V} = \mathbf{C}_b + \mathbf{C}_s + \mathbf{C}_g. \tag{12}$$

Under this model the residual log-likelihood, ignoring constants, is given by

$$\ell_R = -\frac{1}{2}\left\{\log|\mathbf{V}| + \log\left|\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M}\right| + \boldsymbol{\varepsilon}^T\mathbf{P}\boldsymbol{\varepsilon}\right\}, \tag{13}$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{M}(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^{-1}. \tag{14}$$

The Gaussian assumption can not be tested strictly, because it is an assumption about a multivariate distribution of which we have a single realisation. However, its plausibility can be tested by examining a histogram and summary statistics of the residuals of an ordinary least squares fit of the fixed effects model, equivalent to the statistics of the data in this case where a uniform mean is the only fixed effect. Where necessary, data may be transformed to a new scale of measurement to make the assumption more plausible.

We used the *optim* procedure in the *R* package (R development core team, 2013) to find REML estimates of the random effects parameters, the values that maximise the likelihood as defined in Eq. (13). The `L-BFGS-B` optimisation method was selected, a quasi-Newton optimiser in which upper and lower bounds are supplied for the parameters to be estimated (Byrd et al., 1995).

In the proposed model, there are $P = 7$ random effects parameters (or variance parameters) to be estimated by REML. One may consider the "null hypothesis" that one of these parameters can be set at a fixed value, to simplify the model. For example, if one assumed that the cross-section errors for the same geologist at two sites within a batch are uncorrelated, then $\rho = 0$. In general a "null" model with $P - g$ parameters is said to be nested within a more complex "full" model with P parameters if the null model can be regarded as a particular case of the full model with the g additional parameters taking fixed values. The maximised residual likelihood for the full model $\ell_{R,F}$ is at least as large as that for the null model, $\ell_{R,N}$. To test whether the improvement of fit from the g additional parameters is large enough to justify their inclusion within the model one may compute the log-likelihood ratio statistic (Verbeke and Molenberghs, 2000)

$$L = 2(\ell_{R,F} - \ell_{R,N}). \tag{15}$$

We call a comparison between two models a "standard case" if the g additional parameters in the more complex model all take definite values in the null model, and these parameter values are not on the boundary of the parameter space in the null model. In a standard case where the null model is the true model, L is distributed as χ^2 with g degrees of freedom (Stram and Lee, 1994). Note that this procedure is valid for residual likelihoods only when the models have the same fixed effects structure.

One may use this procedure to compare the LMM in Eq. (2) with one in which the geologist effects are regarded as uncorrelated between sites within batches. In the full model, $\rho \in [-1, 1]$, so the fixed value, $\rho = 0$, in the null model is not at a boundary. The comparison is therefore a standard case with $L \sim \chi^2(1)$ under the null hypothesis.

However, if we consider a null model in which the between-batch variance is zero this is not a standard case since zero is the lower bound for a variance. A more general criterion for comparing models of differing complexity, although not a formal test, is to compute for each model Akaike's information criterion – AIC (Akaike, 1973):

$$A = -2\ell + 2P, \tag{16}$$

where ℓ is the maximised log likelihood (natural logarithms) and P is the number of parameters. That model is preferred for which A is smallest, so the term $2P$ is, in effect, a penalty for model complexity.

Model 1b is a variant of 1a in which the between-batch variance is dropped. Since the batches were formed at random, one may expect that the mean error does not differ between the batches, except for random sample variation. However, in a comparison between these two models, the null (1b) is formed by fixing the between-batch variance at zero, which is a boundary in parameter space (variances cannot be negative). The models are therefore compared on the AIC.

Model 1c is a variant of 1a in which the correlation $\rho = 0$. As noted above, this comparison can be made by computing the log-likelihood ratio statistic L and testing it against $\chi^2(1)$.

Having selected one model from among 1a–1c, a variant was considered in which the correlated variance of the between-site random variable, c_1 in Eq. (5), depends on the distance from that site to the nearest borehole available for interpretation (i.e. not in the validation set for the batch). We considered the possibility that this variance is a linear function of distance to the nearest borehole. The intercept and slope of this function, $\alpha_{s,0}$ and $\alpha_{s,1}$, respectively, are therefore substituted for c_1 in model 1d. The comparison of between the null model selected from among 1a–1c and the more complex variant 1d can be made using the log-likelihood ratio, assumed to be distributed as $\chi^2(1)$ under

Table 3. Summary statistics of cross-section error.

Mean	0.70
Median	0.38
SD	2.90
Min	−6.67
Max	7.44
Skewness	0.28
Kurtosis	−0.14

the null model since model 1d has one more parameter than the null.

3.3 Statistical methodology: refining the model to explain the geologist variance (models 2a, 2b and 2c)

Here we consider the possibility that the between-geologist variance can be replaced by a parametric function. In principle this is compatible with any variant of the models considered so far. The expression for the between-geologist covariance matrix in Eq. (10) is modified to

$$\mathbf{C}_\mathrm{g} = \Sigma_g \mathbf{R}_\mathrm{g} \Sigma_g, \tag{17}$$

where $\mathbf{R}_\mathrm{g}$ is defined as for Eq. (10), and

$$\Sigma_g = \mathrm{diag}(\boldsymbol{\sigma}_g), \tag{18}$$

where $\boldsymbol{\sigma}_g$ is a vector of length N which contains the standard deviation of the between-geologist effect for each observation of cross-section error, predicted from some parametric function. The operator "diag" denotes that the elements of this vector are put in order on the main diagonal of an $N \times N$ matrix, with off-diagonal elements equal to zero.

Three parametric functions were considered. In the first, the between-geologist variance for the rth observation of cross-section error depends on the distance from the site which corresponds to the rth observed cross-section error and the nearest borehole available for interpretation to the corresponding site. Again, a linear function was considered, so the parameter σ^2_g in the first and second group of models was replaced by the intercept and slope of this predictive relationship, $\alpha_{\mathrm{g},0}$ and $\alpha_{\mathrm{g},1}$, respectively. These parameters, along with the remaining ones, were estimated by REML.

The second parametric model considered used the geologist's self-assessment of experience in 3-D geological modelling. There were four levels of experience to choose from, so the parameter σ^2_g in the first and second group of models was replaced by four parameters, variances for each level of experience: $\sigma^2_{\mathrm{g},1}$, $\sigma^2_{\mathrm{g},2}$, $\sigma^2_{\mathrm{g},3}$, $\sigma^2_{\mathrm{g},4}$.

A final model was considered which combined the last two variants, with separate intercepts and slopes of the linear function for the geologist standard deviation being specified for each level of experience (i.e. eight new parameters replacing σ^2_g in the first and second group of models.

Note that the parametric functions in these three models return variances, which may vary from one observation of cross section to another. The terms in $\boldsymbol{\sigma}_g$ are standard deviations, i.e. the square roots of the corresponding variances.

3.4 Simulating from the selected model to represent cross-section uncertainty

We used the selected model (model 2a as described in the results section below) to simulate realisations of the random component of cross-section error along a part of the cross section (from 4000 m from the start of the section to the end). We considered a situation where all the boreholes along the cross section were available to the geologist. We assumed that the cross-section error is zero at the location of a borehole, and then simulated the components of the error under model 2a conditional on this at regularly-spaced locations along the cross section. The between-site component was simulated as a multivariate normal random variate by Cholesky decomposition of the joint covariance matrix of the regularly spaced sampling locations and the borehole locations. This is described in detail by Goovaerts (1997). We used the CHOL R procedure (R development core team, 2013). To simulate the between-geologist component, we evaluated the variance of this component at each regularly spaced location on the cross section from the parameters of model 2a as a function of the distance to the nearest borehole. A realisation of the between-geologist component of model error at each location was then simulated as a normal random variable, with mean zero and variance set to this computed value. We used the **rnorm** R procedure to do this (R development core team, 2013). The overall cross-section error was then simulated by the sum of these two components. A total of 10 000 independent realisations of cross-section error were simulated this way.

By finding the 2.5th and 97.5th percentiles of the simulated cross-section errors at any location, we approximate the 95 % confidence interval for model error. This can be used to visualise the uncertainty. The simulations can also be used to answer other questions. Consider, for example, an engineer who wishes to dig a tunnel through the London Clay along the length of this part of the cross section. We assume that the engineer wants to put the route of the tunnel as close as possible to the base of the London Clay, but wants to avoid intruding on the underlying Lambeth Group. The conditional simulations can be used to assess the risk of intruding on the Lambeth Group if the tunnel route is k m above the interpreted base of the London Clay everywhere along the route. Assume that the engineer specifies that the tunnel should enter the Lambeth Group over no more than 1 % of its length. What is the smallest value of k consistent with this? One could examine the 10 000 realisations of cross-section error and find, for increasing values of k, the number of realisations for which the engineer's specification is met: n_k.

Table 4. Model 1 and variants, parameter estimates and inferences.

Model	Random effects parameters							ℓ_R	AIC	Contrast*	L	p
	Batch	Site				Geologist						
	σ_b^2	c_0	c_1	κ	ϕ	σ_g^2	ρ					
1a	0.0	0.0	6.84	2.5	4.36	1.45	−0.093	−148.89	311.78			
1b	** ↓	0.0	6.84	2.5	4.36	1.45	−0.093	−148.89	309.78			
1c	↓	0.0	6.86	2.5	4.38	1.45	↓	−149.75	309.50	1c vs. 1b	1.72	0.19
1d			$\alpha_{s,0}$ $\alpha_{s,1}$									
	↓	0.0	6.03 0.01	2.5	4.38	1.45	↓	−149.48	310.96	1c vs. 1d	0.54	0.46

* The first-named model is the null model. ** A ↓ indicates that a term has been dropped from the model.

Table 5. Model 2 and variants, parameter estimates and inferences.

Model	Random effects parameters								ℓ_R	AIC	Contrast	L	p
	Site				Geologist								
	c_0	c_1	κ	ϕ									
2a						$\alpha_{g,0}$		$\alpha_{g,1}$					
	0.0	6.63	2.5	4.73		0.0		0.0217	−117.18	246.36	1c vs. 2a	65.1	$< 10^{-15}$
2b					$\sigma_{g,1}^2$	$\sigma_{g,2}^2$	$\sigma_{g,3}^2$	$\sigma_{g,4}^2$					
	0.0	7.53	2.5	4.59	4.44	2.25	1.32	0.46	−144.16	304.31	1c vs. 2b	11.2	0.01
2c*					$\alpha_{g,1,1}$	$\alpha_{g,1,2}$	$\alpha_{g,1,3}$	$\alpha_{g,1,4}$					
	0.0	6.72	2.5	4.58	0.022	0.021	0.030	0.018	−116.63	257.26	2a vs.2c	1.1	0.98
											2b vs. 2c	55.1	$< 10^{-10}$

* In this model, a separate slope and intercept to compute the between-geologist variance as a function of distance to the nearest borehole was computed for each level of experience. All estimates of the intercept were zero exactly, so only the slopes are reported here.

The probability of meeting the specification given some k can then be estimated as $n_k/10\,000$.

4 Results

4.1 Summary statistics on model error from all validation sites

Figure 2 shows a scatter plot of interpreted and observed heights of the base of the London Clay for all observations of cross-section error. The points are scattered around the bisector (where observed and interpreted heights are equal), and there is no visual evidence of a systematic bias. Table 3 shows the summary statistics of cross-section error, and Fig. 3 shows the histogram of this variable. The symmetrical form of the histogram and the weak skewness and kurtosis values suggest that an assumption of normality is plausible for the analysis of these data. They also suggest that, if there is any systematic tendency for the base of the London Clay to be interpreted too high or too low, then this effect is small.

4.2 Model comparisons

The results for model 1a and its variants are shown in Table 4. Note that the estimated between-batch variance is zero. When a REML estimate of a parameter is at the boundary of parameter space, as here, it is advisable to examine the likelihood profile in the vicinity of the estimate. To compute the likelihood profile for a model parameter, that parameter is fixed at a series of values and, for each, the remaining parameters are estimated by maximum (residual) likelihood. The maximised likelihoods are then plotted against the values of the parameter of interest. The profile likelihood should increase smoothly towards the estimated value. The profile likelihood for the batch variance satisfied this requirement. This is not unreasonable; because the batches were formed at random, we would hope that the between-batch variation is purely explicable in terms of sampling error. The comparison of models 1a and 1b can be done by examining the AIC, which is smaller for the latter model, in which the batch effect is dropped. Model 1b is therefore selected over 1a. The profile likelihood for the uncorrelated between-site variance in these models, c_0, also approached the estimated value, 0.0, smoothly.

Table 6. Summary of model comparisons. In each case, the first-named simpler "null" model is compared with a more complex alternative, either on the log-likelihood ratio L or Akaike's information criterion (AIC). The key conclusion from the comparison is indicated.

Models Null model		Criterion for comparison	Selected model	Conclusion
1b	1a	AIC	1b	Batch effect can be dropped.
1c	1b	L	1c	Correlation of geologist errors can be dropped.
1c	1d	L	1c	No evidence that between-site variance depends on the distance to the nearest borehole.
1c	2a	L	2a	Evidence that between-geologist variance depends on the distance to the nearest available borehole.
1c	2b	L	2b	Evidence that between-geologist variance depends on the geologist's experience.
2b	2c	L	2c	Evidence that the relationship between between-geologist variance and experience depends on the distance to the nearest available borehole.
2a	2c	L	2a	No evidence that adding modeller experience improves the model, with the distance to the nearest available borehole already included.

Table 7. Estimate of the mean cross-section error conditional on model 2a.

Estimated mean	0.52
Standard error	0.42
Wald statistic*	1.56
p value	0.21

* The Wald statistic tests the null hypothesis that the true mean error is zero. The p value is the probability of obtaining a Wald statistic this large or larger under the null hypothesis.

In model 1c, the correlation between within-site effects for particular geologists is dropped (set to zero). The maximum likelihood is slightly smaller than for model 1b, in which this parameter is estimated. However, the log-likelihood ratio statistic, L, for the comparison of (null) model 1c with (the full) model 1b is small, and the probability of obtaining a value of L this large or larger under the null model is large, and so the more complex model is rejected in favour of the null one. This is also consistent with the small estimated value of this correlation, -0.09.

In model 1d, a stationary correlated variance for the between-site effect (as in model 1c) is replaced by two parameters for a linear function which expresses this variance as a function of distance to the nearest borehole available for interpretation. This (full) model can be compared with a (null) model (1c) with a stationary variance by the log-likelihood ratio test. Once again, L is too small to support a choice of the more complex model.

In summary, the consideration of model 1a and its variants in Table 4 leads us to the selection of model 1c (smallest AIC in the table), in which the batch effect and the correlation parameter ρ for geologist effects are dropped, and the between-site variation is modelled as a stationary correlated random variable.

Table 5 shows results for model 2a and its variants. These models are based on 1c, but differ in that, rather than assuming a stationary geologist effect, the between-geologist within-site variance is modelled as a function of covariates. In model 2a, the geologist variance is modelled as a linear function of distance to the nearest borehole available to the geologist for interpretation. The zero value of the intercept, $\alpha_{g,0}$, is plausible, under the assumption implicit in our analysis that the borehole data are correct, and the cross-section error should be zero at the location of a borehole. The positive value of $\alpha_{g,1}$ implies that the geologist variance increases with increasing distance from a borehole, which is also plausible. Model 1c can be regarded as nested within 2a, a null model with $\alpha_{g,0}$ equivalent to σ_g^2 and $\alpha_{g,1} = 0$. The models can be tested by the log-likelihood ratio statistic; Table 5 shows that the null model (1c) can be decisively rejected in favour of the full model 2a.

Model 2b is an alternative to 2a, in which the geologist variance depends on the self-identified experience of the geologist in 3-D modelling. The estimated parameters in Table 5 are plausible in that the variance is largest for geologists who identified themselves as having "no experience of modelling in 3-D" and smallest for those who identified themselves as having "substantial experience of more than 2 years of modelling independently." Once again, this model could be compared with 1c by a log-likelihood ratio test, and the null model (1c) can be rejected, indicating that there is significant evidence for differences in geologist variance, related to geologist experience. However, the evidence for this model is weaker than for 2a.

In model 2c different relationships between geologist variance and distance to nearest borehole were fitted for the four levels of geological experience. In the fitted model the intercepts were all zero, the smallest slope is for the geologists with the highest experience level. However, while the log-likelihood ratio test shows that model 2c is significantly better than model 2b (i.e. adding the information on distance to nearest borehole to a model with geologist experience gives

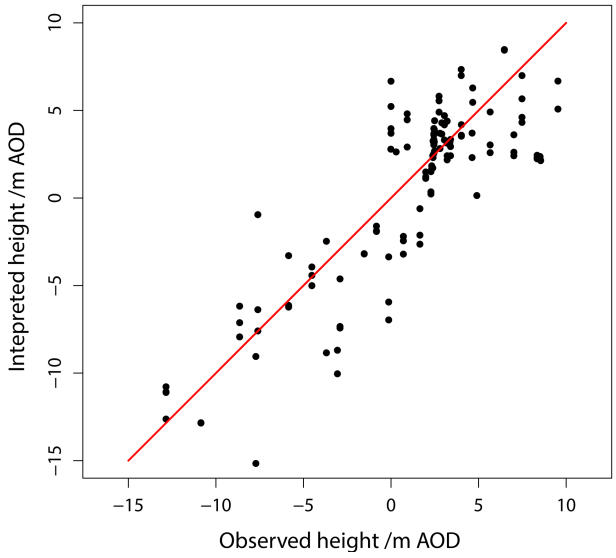

Figure 2. All validation observations of the interpreted and observed height of the base of the London Clay AOD. The red line is the bisector.

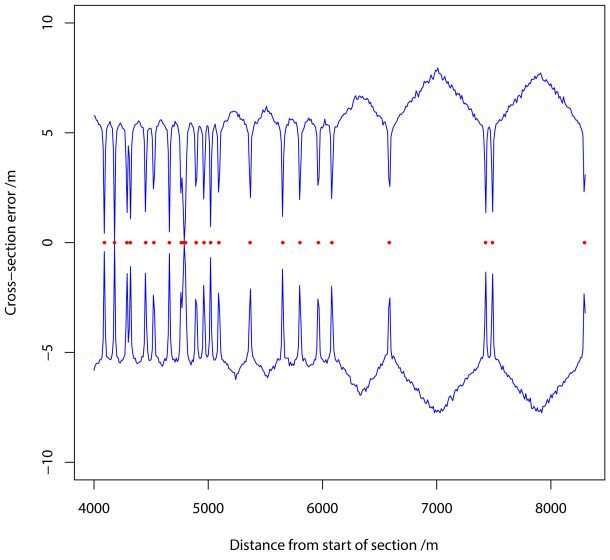

Figure 4. 95 % probability interval for simulated cross-section errors conditional on the location of the nearest borehole (red symbol) and model 2a. Note that these are evaluated at discrete locations.

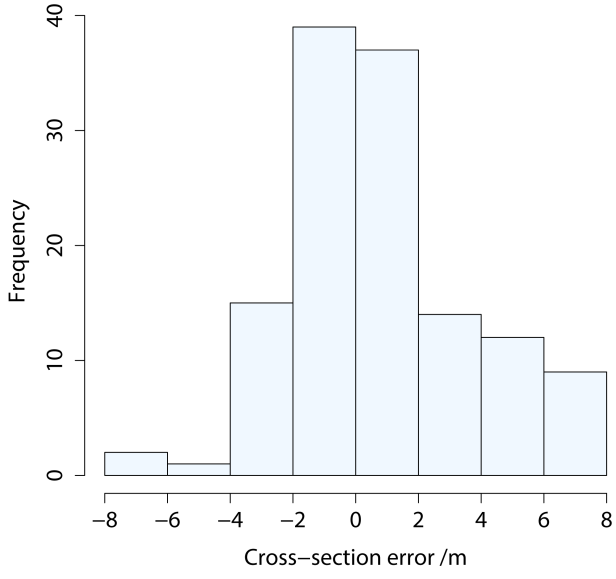

Figure 3. Histogram of cross-section errors.

a significant improvement), the comparison of model 2c with 2a leads to the conclusion that adding geologist experience to a model which already has the distance to nearest borehole incorporated does not give a significant improvement. On the basis of the AIC, model 2a is preferred among all those considered in this study. Table 6 summarises all the key comparisons between models and the inferences which arise from these comparisons.

Table 7 shows the estimated mean cross-section error and its standard error, under model 2a. The Wald statistic (e.g. Dobson, 1990) is a test of the null hypothesis that the mean error is zero, and the large p value shows that this cannot be rejected, so the data provide no evidence for systematic bias in the interpretation. Note that the estimate of mean error is rather less than the average for all observations reported in Table 3. That is because (i) the mean reported in Table 6 is the model mean, the fitted effect in the underlying statistical model for cross-section error, and (ii) the original data were not a random sample in space, and show some local clustering which is likely to bias the arithmetic average as an estimate of the spatial mean of cross-section error.

Fig. 4 shows the 95 % probability interval for cross-section errors along the section, approximated by the 2.5th and 97.5th percentiles of the conditionally simulated errors. The red symbols show the locations of the boreholes. There are two features of the interval. First, there is a rapid narrowing near the boreholes (the interval is zero at the boreholes, but this is only seen if the borehole coincides with a point where the error is sampled). This arises from the spatial correlation of the between-site component of cross-section error. The second feature is a gradual widening of the interval to a local maximum at the midpoint between successive boreholes. This is particularly apparent in the second half of the plot. This arises from the dependence of the between-modeller effect on the distance to the nearest borehole, showing how the constraint of the borehole on model error decays with distance. In Fig. 5, the confidence intervals are added to the interpretation of the base of the London Clay by one of the modellers.

Fig. 6 shows a plot of the estimated probability that a tunnel built k m above the interpreted base of the London Clay will intrude on the underlying Lambeth Group over no more than 1 % of its length for different values of k. This shows

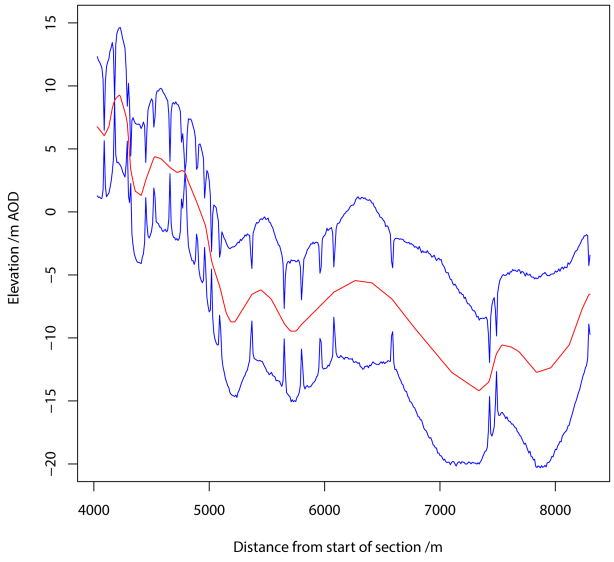

Figure 5. One geologist's interpretation of the base of the London Clay (red) with 95 % confidence intervals (blue).

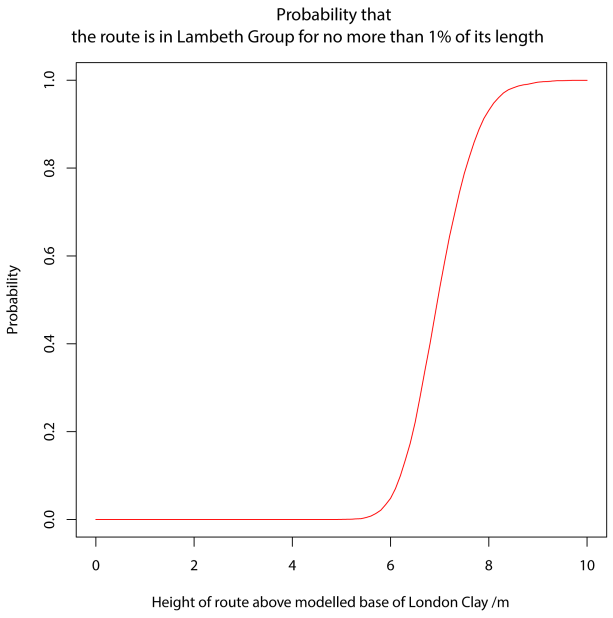

Figure 6. How close to the modelled base of the London Clay could you build a tunnel (over the last 4 km of the cross section) and have a specified probability (ordinate) that the tunnel will stray into the underlying Lambeth Group for no more than 1 % of its length?

that the engineer can be 90 % confident that this specification will be met if the route is a little less than 8 m above the interpreted base.

5 Conclusions

Both the summary statistics and the scatter plot (Fig. 1), and the estimate of the mean cross-section error from the selected model 2a (Table 7), show that the data obtained in this study provide no evidence that there is any bias in the interpretation of the base of the London Clay by the geologists in this study; i.e. the mean error is not significantly different from zero.

We established this experiment to test the hypothesis that the variability of the error of interpretations of cross sections varies spatially. This hypothesis has been supported. First, we found that there is spatial dependence in the variability of the between-site component of cross-section error. This is to say that the cross-section error at one location is likely to be more strongly correlated with the error at a nearby location than at one farther away. This is reasonable, since if, for example, a surface tends to be interpreted as being too high above the Ordnance Datum at a site, perhaps because of faulting, then it is likely that a similar error will occur at nearby sites. There was no evidence, however, that the between-site variance depends on the distance to the nearest borehole.

The between-geologist variance is rather smaller than the between-site variance (compare c_0 with σ_g^2 in model 1c). However, there was evidence that the variance of this error depends on geologist experience and also on the distance to the nearest borehole available for interpretation. The results for these two models are consistent with our hypothesis, and also make intuitive sense in that the variance of cross-section error declines with the geologist's experience, and increases with increasing distance from the borehole. However, the preferred model for the data, given a penalty on model complexity, considers only the distance to nearest borehole. It is interesting to note that our results on how model uncertainty increases with distance to constraining interpretation boreholes, and the effect of modeller experience, are consistent with the opinions on sources of uncertainty that have been elicited in published studies (e.g. Lelliott et al., 2009). This study provides empirical evidence for these opinions, and a direct quantification of the effects.

The fitted model can be used to simulate cross-section errors, conditional on a distribution of boreholes. One may use this procedure to compute confidence intervals around the interpreted cross section which quantifies uncertainty in this interpretation and shows how this changes in space. One could also use this simulation method to study the propagation of cross-section error in further processing to interpolate the surface into 2-D, and so produce 3-D volumes.

The methodology presented in this paper could be deployed in a wider range of geological settings in order to generate statistical models of cross-section error for those settings. These could then be used to compute confidence intervals for new models or measures of uncertainty specific to the requirements of particular data users, such as the example for the London Clay illustrated in Fig. 5.

The experimental design used in this study allowed us to make best use of somewhat sparse boreholes by examining multiple geologist interpretations at each validation site.

However, if there had been a significant correlation between within-site effects for the same geologist, then subsequent modelling of the geologist variance would have been complicated. Alternatively, one might use an experimental design in which validation sites are nested within modellers (so each modeller has a unique batch of validation sites). This requires there to be many boreholes available, however, since each validation borehole is compared with just one interpretation. It also reduces the information that we obtain on between-modeller differences.

One way to get around the problem of insufficient validation observations is to generate synthetic cross sections, perhaps conditioned on geophysical data such as interpretations from seismic lines. These synthetic cross sections can then be notionally sampled at as many locations as we want to provide synthetic borehole data for interpretation and validation. In such an experiment, the syntheticvalidationboreholes should be sampled according to an optimised design (e.g. Lark, 2002) to ensure good estimation of the spatial variance parameters and to give good coverage of possible covariates, e.g. spanning a range of distances to the nearest borehole available for interpretation.

Acknowledgements. We are grateful to our colleague Luz Ramos Cabrera for her contribution to setting up the trial, to those delegates to the 2012 GSI3D workshop, and to other colleagues at the British Geological Survey who participated. This paper is published with the permission of the executive director of the British Geological Survey (NERC).

Edited by: K. Zeigler

References

Aitken, A. R. A., Holden, E-J., and Dentith, M. C.: Semiautomated quantification of the influence of data richness on confidence in the geologic interpretation of aeromagnetic maps, Geophysics, 78, 1–13, 2013.

Akaike, H.: Information theory and an extension of the maximum likelihood principle, in: Second International Symposium on Information Theory, edited by: Petov, B. N. and Csaki, F., Akademia Kiado, Budapest, 267–281, 1973.

Bistacchi, A., Massironi, M., Dal Piaz, V. G., Monopoli, B., Schiavo, A., and Toffolon, G.: 3-D fold and fault reconstruction with an uncertainty model: An example from an Alpine tunnel case study, Comput. Geosci., 34, 351–372, 2008.

Bond, C. E., Gibbs, A. D., Shipton, Z. K., and Jones, S.: What do you think this is? 'Conceptual uncertainty' in geoscience interpretation, GSA Today, 17, 4–10, 2007.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C.: A limited memory algorithm for bound constrained optimization, SIAM J. Sci. Comput., 16, 1190–1208, 1995.

de Gruijter, J. J., Brus, D. J., Biekens, M. F. P., and Knotters, M.: Sampling for Natural Resource Monitoring, Springer, Berlin, 2006.

Diggle, P. J. and Ribeiro, P. J.: Model-Based Geostatistics, Springer, New York, 2007.

Dobson, A. J.: An Introduction to Generalized Linear Models, Chapman & Hall, London, 1990.

Ellison, R. A.: Facies distribution in the Woolwich and Reading Beds of the London basin, England, P. Geologist. Assoc., 94, 311–319, 1983.

Ellison, R. A., Woods, M. A., Allen, D. J., Forster, A., Pharaoh, T. C., and King, C.: Geology of London. Memoir of the British Geological Survey, Sheets 256 (North London), 257 (Romford), 270 (South London) and 271 (Dartford) (England and Wales), British Geological Survey, Keyworth, 2004.

Gunnink, J. L., Maljers, D., van Gessel, S. F., Menkovic, A., and Hummelman, H. J.: Digital Geological Model (DGM): a 3-D raster model of the subsurface of the Netherlands, Netherland Journal of Geosciences – Geologie en Mijnbouw, 93, 33–46, 2013.

Goovaerts, P.: Geostatistics for Natural Resources Evaluation, OUP, New York, 1997.

Howard, R. A.: Information value theory, IEEE T. Syst. Sci. Cyb., 2, 22–26, 1966.

Jessell, M., Ailleres, L., and de Kemp, E.: Towards an integrated inversion of geoscientific data: what price geology? Tectonophysics., 490, 294–306, 2010.

Kessler, H. and Mathers, S. J.: Maps to models – finally capturing the geologists' vision, Geoscientist, 14, 4–6, 2004

Kessler, H., Mathers, S. J., and Sobisch, H.-G.: The capture and dissemination of integrated 3-D geospatial knowledge at the British Geological Survey using GSI3D software and methodology, Comput. Geosci., 35,1311–1321, 2009.

Lark, R. M.: Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood, Geoderma, 105, 49–80, 2002.

Lark, R. M.: Kriging a soil variable with a simple non-stationary variance model, J. Agr. Biol. Envir. St., 14, 301–321, 2009.

Lark, R. M. and Webster, R.: Geostatistical mapping of geomorphic variables in the presence of trend, Earth Surf. Proc. Land., 31, 862–874, 2006.

Lark, R. M., Mathers, S. J., Thorpe, S., Arkley, S. L. B., Morgan, D. J., and Lawrence, D. J. D.: A statistical assessment of the uncertainty in a 3-D geological framework model, P. Geologist. Assoc., 124, 946–958, 2013.

Lelliott, M., Cave, M., and Wealthall, G.: A structured approach to the measurement of uncertainty in 3D geological models, Quart. J. Engin. Geol. Hydrogeol., 42, 95–105, 2009.

Matérn, B.: Spatial Variation, Lecture Notes in Statistics, No. 36, Springer, New York, 1986.

Nelder, J. A. and Lee, Y.: Joint modelling of mean and dispersion, Technometrics, 40, 168–171, 1998.

Patterson, H. D. and Thompson, R.: Recovery of inter-block information when block sizes are unequal, Biometrika, 58, 545–554, 1971.

R Development Core Team: R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org/, 2013.

Smyth, G. K. and Verbyla, A. P.: A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models, J. Roy. Stat. Soc. B, 58, 565–572, 1996.

Stein, M. L.: Interpolation of Spatial Data: Some Theory for Kriging, Springer, New York, 1999.

Stram, D. O. and Lee, J. W.: Variance components testing in the longitudinal mixed effects setting, Biometrics, 50, 1171–1177, 1994.

Sumbler, M. G.: British Regional Geology: London and the Thames Valley, 4th edition, HMSO London, 1996.

Tacher, L., Pomian-Srednicki, I., and Parriaux, A.: Geological uncertainties associated with 3-D subsurface models, Comput. Geosci., 32, 212–221, 2006.

Torvela, T. and Bond, C. E.: Do experts use idealised structural models? Insights from a deepwater fold–thrust belt, J. Struct. Geol., 33, 51–58, 2011.

Verbeke, G. and Molenberghs, G.: Linear Mixed Models for Longitudinal Data, Springer-Verlag, New York, 2000.