Prediction of interface of geological formations using generalized additive model

Xiaohui Qi[a], Zhiyong Yang[b*], Jian Chu[c]

[a]Lecturer, Department of Mechanical and Construction Engineering, Northumbria University, Newcastle upon Tyne, UK. Email: xiaohui.qi@northumbria.ac.uk.
[b]Assistant Professor, School of Civil Engineering, Sun Yat-Sen University, Guangzhou, China. Email: yangzhy85@mail.sysu.edu.cn.
[c]Professor, School of Civil and Environmental Engineering, Nanyang Technological University, Singapore. Email: cjchu@ntu.edu.sg.
*Corresponding author

Abstract: Geological information such as geological interfaces is important for the design of underground excavation and supporting measures. This in turn requires a method to predict accurately the locations of geological interfaces for the gap areas between boreholes. This study presents a generalized additive model (GAM) to predict the location of the geological interfaces. The performance of the GAM method is evaluated using both simulated data and borehole data for the determination of rockhead in two different geological formations in Singapore. The results show that the GAM method can provide a reasonable confidence interval (CI) of the mean trend and the prediction interval (PI) in the sense that the 95% CI covers about 95% of the actual mean curve while the 95% PI covers around 95% of testing data. Furthermore, the geological complexity can be well reflected as the prediction uncertainty in the geologically complex area is larger than that in the geologically regular area. More importantly, the users can impose prior information or personal judgment regarding the shape of the geological profile on the model. This is an important feature to enable further improvement in the accuracy of the prediction.

Keywords: spatial prediction, geological interface, generalized additive model, cubic spline

1. Introduction

Geological model including the interfaces of different geological formations is indispensable information for underground constructions as it may affect the construction method and supporting measures. It is necessary to predict the location of geological interfaces in the gap areas between boreholes. This task is difficult as the site exploration data are always sparse and limited. What makes this task more challenging is the large variability in the geological interface, especially the rockhead in rock formations. The reason is that the weathering of the rock is affected by many factors, including climate, topography, hydrological conditions, biological systems, rock mass discontinuities, rock composition and permeability (Zhao et al. 1994). It is vital to find an effective method to accurately predict the location of geological interfaces. The prediction should be able to provide a predicted value of the location of the geological interface as well as its uncertainty.

A variety of methods have been used for interpolation problems in geotechnical or geological engineering. These methods can be divided into two categories, deterministic methods and statistical methods. Deterministic methods such as the inverse distance weighting method, spline interpolation or the triangle-based tessellation method (e.g., Aswar and Ullagaddi 2017; Burke et al. 2017) cannot automatically quantify the uncertainty of the prediction or provide any confidence interval of the predicted property. This problem can be to some extent addressed using cross-validations, as shown in Lark et al. (2013), but the results highly depend on the employed testing data and the quantified uncertainty may not be reliable when the testing data are limited. The statistical interpolation methods include the coupled Markov chain method (Qi et al. 2016; Li et al. 2019; Liu et al. 2020), Markov random field method (Wang et al. 2017, 2018), Bayesian compressive sampling method (Wang and Zhao 2016, 2017; Zhao, Hu, and Wang 2018), random field method (Gong et al. 2020; Zhao et al. 2021), geostatistical methods such as kriging and conditional random field method (Qi et al. 2019, 2021a). The coupled Markov chain method can characterize the geological uncertainty using limited borehole data, but it can only be used when the transition of geological types has a Markovian property. The Markov random field method can model complex geological structures, but some of its parameters lack clear physical meaning (Mariethoz and Caers 2014). The recently developed Bayesian compressive sampling method can quantify the interpolation uncertainty using limited data and has a high interpolation accuracy (Wang, Akeju, and Zhao 2017). Moreover, it is quite versatile in that it can model both stationary and non-stationary random fields (Wang, Zhao, and Phoon 2018; Wang et al. 2019). One potential problem of the method is that its robustness degrades when the number of measurements is smaller than the length of the discrete signal or when the measurement noise is relatively large (Huang et al. 2014). Geostatistical methods such as kriging or conditional random field have gained wide popularity (e.g., Qi et al. 2019, 2021a). One problem with geostatistical methods is that they are purely mathematic based and may not lead to realistic soil or geological profiles. For example, the conditional random field or the kriging method normally produces a soil or geological profile with extreme values only at known data points. Furthermore, some artificial intelligence methods such as neural networks (Zhou and Wu 1994) and the support vector machine (Smirnoff, Boisvert, and Paradis 2008) were also applied to spatial prediction problems of geological conditions. The drawback of these methods is that they lack interpretability in the sense that they behave like black boxes and the effect of individual explanatory variables is difficult to examine.

Recently, Qi et al. (2020, 2021b) applied a spline regression method to spatial predictions of the location of geological interfaces. It has been shown that the method can provide a clear spatial trend of the geological interface, which well reflects the geological complexity. One problem of these studies is that the uncertainty in the mean trend is not well quantified or distinguished from the uncertainty in the random error. Herein the uncertainty in the mean trend represents the bias of the fitted curve in a regression. To be specific, if two different sets of data for the same explanatory and response variables are employed to perform regression, the two fittings generally produce different mean trends. The variability in the fitted curves is called uncertainty in the mean trend. The uncertainty in the random error denotes the deviation of the data points from the fitted curve. To address this issue, this study uses a generalized linear model (GAM) to perform the spatial prediction of the geological interface. For the GAM, the response variable is expressed as the weighted average of some basis functions (Wood 2017). It can be viewed to be a non-parametric or semi-parametric method in the sense that the structure of the model is not fixed. The uncertainty in the mean trend and random error can be explicitly considered by the GAM. An additional advantage of the GAM is its interpretability, which means the contribution of each independent parameter to the prediction is explicitly modeled and can be readily examined.

In this study, the GAM is firstly briefly introduced. Secondly, the performance of the GAM model is investigated based on cross-validations using simulated data. Finally, some borehole data from Singapore, which reveal the rockhead elevation of two rock formations, i.e., Bukit Timah Formation and Jurong Formation, are used to predict the rockhead. Herein the rockhead is the interface of soil and rock layers in a rock formation, which is mainly determined according to the weathering degree of the geological layers (Qi et al. 2020, 2021b). Prediction errors and the capability of GAM in characterizing the two types of uncertainty are evaluated using a cross-validation procedure. The reasonableness of the prediction uncertainty, which was generally ignored in existing studies, is investigated in this paper. The two types of uncertainties, i.e., uncertainty in the mean trend and random error are well differentiated and quantified in the investigation. The role of engineering judgment in spatial predictions is also discussed in the example.

2. Generalized additive model

The section introduces briefly the generalized additive model to be used for the prediction of the rockhead elevation. The GAM is originally developed by Hastie and Tibshirani (1986, 1990). It can be viewed to be a generalization of the linear regression model. The main advantage of the GAM is that it can flexibly identify the nonlinear relation between explanatory variables (also called predictors or covariates) and a response variable (Hastie and Tibshirani 1986; Wood 2017). To be specific, the users do not need to specify a particular parametric function to represent the nonlinear pattern. Instead, non-parametric or semi-parametric smooth functions are used to relate the predictors and responses. Another advantage of the GAM model is the interpretability, which means that the effect of predictors can be examined separately and explicitly (Hastie and Tibshirani 1986). The GAM has been widely applied to various disciplines since its advent, such as environmental engineering (e.g., Gong et al. 2017; Ma et al. 2020), soil science (e.g., de Brogniez 2015), ecology (e.g., Yee and Mitchell 1991; Simpson 2018), transportation engineering (e.g., Khoda Bakhshi and Ahmed 2021). In geotechnical engineering, the model is applied for the determination of landslide susceptibility, as shown in Goetz et al. (2015) and Bordoni et al. (2020). The spline regression methods investigated in Qi et al. (2020, 2021b) are also GAM. This study extends the work in Qi et al. (2020, 2021b) by taking the uncertainty in the fitted mean trend of the response variable into consideration. The basic idea and the fitting of the GAM are introduced as follows.

2.1 Representation of smooth functions

This study intends to investigate a one-dimensional problem, namely the prediction of the rockhead elevation along a line. The response variable is the rockhead elevation while the explanatory variable is the distance to the leftmost point on the line. Besides, since the rockhead elevation does not have any capped value, the Gaussian distribution is taken to be the probability distribution of the response. In this case, the GAM can be simplified into

$$y = f(x) + \varepsilon \quad (1)$$

where ε is a normally distributed random variable with a mean of 0 and variance of σ_ε^2. The smooth function is usually represented by the weighted sum of several basis functions, i.e.,

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i \quad (2)$$

where $b_i(x)$ is a basis function of which the expression is already known; β_i is an unknown coefficient and q is the total number of basis functions, also called the dimension of basis. This study adopts the commonly used cubic spline basis functions because the cubic spline interpolant always provides a solution that is smooth and closely approximates to the true function whatever the true function is (Wood 2017). Polynomial bases are not chosen because a high-order polynomial function usually causes an oscillation problem, as shown in De Boor (2001). One example of curve fitting using cubic spline basis functions is plotted in Fig. 1. In

Fig. 1, the 19 circles denote blow count value data from standard penetration tests taken from Baecher and Christian (2008). The vertical coordinate denotes the elevation while the horizontal coordinate denotes the blow count value. The 19 data points are fitted with three, six and twelve cubic spline basis functions using a least squared method in Figs. 1a, 1b and 1c, respectively. The basis functions are denoted as the blue dashed lines. Since the values of the basis function are too small (< 1), the basis functions are magnified by 10 times to make them discernible in the figure. Also, the locations of the knots are denoted as the red dotted lines. These knots are the connection points of two neighbouring pieces or sections of the fitted curve, each of which can be expressed by a cubic spline function. For illustration purposes, the knots are set to be evenly spaced in the elevation direction. As shown in Fig. 1, the resulting smooth function is a piecewise cubic polynomial function. The fitted curve becomes wigglier when more basis functions are used.

2.2 Degree of smoothing

After the structure of the smooth function is known, one natural question is how to determine the number and location of the knot given some data. The number and location of knots control the degree of smoothing of the resulting function. Too many knots result in a wiggly curve running across all the data points. This curve normally suffers from overfitting and performs poorly when it is used for prediction. In practice, the number and location of the knots can be determined by model selection methods or cross-validation, as shown in Qi et al. (2020, 2021b). An alternative method to control the smoothness is fixing the basis dimension at a relatively large size and adding a wiggliness penalty term in the least-squares objective (Wood 2017), i.e.,

$$||\sqrt{\mathbf{W}}(y\text{-}\mathbf{X}\boldsymbol{\beta})||^2+\lambda \int [f''(x)]^2 \, \mathrm{d}x \tag{3}$$

where $||\cdot||^2$ is the squared Euclidian length of a vector; $\mathbf{X}\boldsymbol{\beta}$ denotes the fitted values of the smooth function in which $\mathbf{X}$ is the model matrix denoting the values of the basis functions at locations of observation data while $\boldsymbol{\beta}$ is the coefficient vector; $\mathbf{W}$ is a diagonal matrix denoting the weights of data points. Assigning a weight value of w_i for a data point is equivalent to put w_i identical data points at the same location. Normally the weights for all the data points are set to be 1. A value larger than 1 can be used when a data point reveals an important geological feature (such as an abruptly low rockhead caused by faults) and controls the shape of the geological profile. $f''(x)$ is the second derivative of the smooth function $f(x)$; λ is the smoothing parameter that controls the tradeoff between the model fit and the model smoothness. $\lambda = 0$ results in an un-penalized spline regression and would produce a very wiggly curve. An infinitely large of λ would lead to a linear estimate of the true function (Wood 2017). The first term in Eq. 3 represents the fitting errors while the second term denotes the penalty against wiggliness of the fitted function. An optimal solution can be sought out by minimizing the objective expression in Eq. 3. For this alternative method, the number and location of knots do not significantly affect the fitted curve once the number of knots or basis function is large enough (Wood 2017).

It is worth noting that the users of the GAM can impose their prior information, personal knowledge or judgment regarding the geological profile on the model. This can be accomplished by assigning suitable weights to specific data points which reveal a geological feature and dominate the shape of the geological profile. The prior information can also be fed into the model by setting a proper value of the smoothing parameter. For example, if an area is subject to intensive tectonic activities in history, the geological profile is expected to be wavy.

In this case, the smoothing parameter can be set to be relatively small.

2.3 Selection of smoothing parameter

In addition to manually setting a value for the smoothness parameter, its value can also be determined using other ways. A variety of methods have been proposed to determine the value of the smoothing parameter. One simple way is to minimize certain index which denotes the prediction errors, such as the Akaike information criterion (AIC), ordinary cross-validation score, generalized cross-validation (GCV) score (e.g., Wood 2017; Simpson 2018). The second way is to treat the smoothing parameter as a random variable and estimate its value using the maximum likelihood or the restricted likelihood method. In this study, the GCV score is used to determine the value of λ.

For a known value of λ, the coefficient parameters $\boldsymbol{\beta}$ can be estimated using the penalized least square estimation method. The variance of the random error can be estimated as the residual sum of squares divided by the residual degree of freedom. Details of the estimation of these parameters can be found in Wood (2017).

2.4 Mean trend uncertainty, confidence interval and prediction interval

The fitted smooth function can hardly be the actual function of the response variable because of various uncertainties. A reasonable practice is to provide a mean trend as well as a band that denotes the uncertainty. It is worth noting that there are two different kinds of uncertainty, namely the uncertainty in the mean trend and the uncertainty in a prediction (Ruppert, Wand, and Carroll 2003). The former means that if two different sets of data for given explanatory and response variables are used to perform the regression, the fitted curves are expected to be different. This variability in the fitted curve is called the uncertainty in the mean trend. The latter means that if the fitted smooth function is used for predicting a response, the predicted value will be different from the actual value. This uncertainty is caused by both the error in the mean trend and the random error which denotes the deviation of data points from the mean trend (see Fig. 1). The random error can be attributed to measurement errors in the data or other sources of errors. For example, for the rockhead elevation, the random error can be caused by the subjective judgments of engineers in determining the weathering degree of the geological layer. The standard deviation of the mean trend, SD_{μ} indicates the epistemic uncertainty in the spatial prediction, which can be reduced if more observations are available. SD_{μ} reflects the geological complexity as well as the data quantity of the investigated area. The value of SD_{μ} would be quite small if many data exist or the geological profile is very simple such as a flat curve. The standard deviation of the random errors, $\mathrm{SD}_{\varepsilon}$, suggests the magnitude of the aleatoric uncertainty in the spatial prediction, which cannot be decreased even if sufficient data are available. $\mathrm{SD}_{\varepsilon}$ represents the minimum error or maximum accuracy that can be achieved in spatial predictions.

The first uncertainty is usually expressed by confidence interval (CI) while the latter by the prediction interval (PI). For example, the 95% CI of the mean trend is bounded by the mean trend minus and plus twice the estimated standard deviation of the mean trend, $2\mathrm{SD}_{\mu}$. SD_{μ} is derived from the standard deviation of the coefficient parameter, which can be estimated using either the frequentist or Bayesian approach. For the Bayesian method, the posterior distribution of the standard deviations of the coefficient parameters has an analytical solution when the prior distribution, $f_{\beta}(\beta)$, is given by (Wood 2017)

$$f_{\beta}(\beta) \propto \mathrm{e}^{-\frac{1}{2}\beta^{\mathrm{T}} \sum S_i/\tau_i \beta} \tag{4}$$

where the τ_i are parameters controlling the dispersion of the prior distribution; S_i is an element from the penalty matrix, which is a matrix of known coefficients and is derived from $f''(x)$. The prior distribution in Eq. 4 gives equal probability density to all models of equal smoothness, but larger probability densities to smooth models than wiggly models as normally it is believed that smooth models are more likely than wiggly models. More details regarding the estimation of the uncertainty in the coefficient parameter can be found in Wood (2017). The 95% PI of a prediction is bounded by the mean trend minus and plus twice the standard deviation of a prediction, given by $2\mathrm{SD}_\mathrm{p} = 2\sqrt{\mathrm{SD}_\mu^2 + \mathrm{SD}_\varepsilon^2}$, where SD_ε denotes the standard deviation of the random error. In this study, the fitting of a GAM model is performed using a well-known R package, *mgcv*.

3. Spatial prediction using simulated data

This section evaluates the performance of the GAM model using cross-validation based on simulated data. Firstly, dense data are simulated based on a smooth trend function and a random error. Secondly, the simulated data are divided into two groups, the training group and the testing group. Thirdly, the training data are used to estimate the unknown parameters of the GAM model and generate the 95% CI and 95% PI. Finally, the coverage percentage of the 95% CI is evaluated by counting the percentage of the input trend function covered by the 95% CI (e.g., take 100 evenly spaced data points from the input trend and check how many of them are covered by the 95% CI) while that of the 95% PI is assessed by computing the percentage of the testing data covered by the 95% PI. These steps are repeated by 500 times and the average values of the coverage percentages are computed.

An example taken from Wood (2017) is used to analyze the performance of the GAM, including the CI and PI, the latter of which is not investigated in Wood (2017). The input trend function is given by

$$f(x)=x^{11}(10[1-x])^{6}+10(10x)^{3}(1-x)^{10} \quad (5)$$

The range of the explanatory variable, x is set to be [0, 1] while the range of the input function is scaled to the interval of [0, 1] by dividing $f(x)$ by the maximum value of $f(x)$. The curve of the input smooth function is plotted as the dotted line in Fig. 2(a). 500 samples of x are randomly drawn from the uniform distribution with a range of [0, 1]. Simulated data of y were generated by adding a random error to the input mean trend for the 500 samples of x. The random errors are normally distributed with a mean of 0 and standard deviation of 0.2 and are mutually independent at various locations. 50 data points are randomly drawn from the 500 data points and set as training data while the remainder as testing data. The training data are used to fit the GAM model and create the 95% CI and 95% PI. The dimension of basis was set to be 20 as a larger basis dimension produces quite similar results. Besides, 500 experiments were carried out as such quantities of experiments are sufficient to yield a converged estimation of the coverage percentages.

A typical experiment of the cross-validation is plotted in Fig. 2(a-c). Fig. 2(a) plots the training and testing data, respectively. Fig. 2(b) plots the 95% PI and the 95% CI, which are denoted by the region between the dashed lines and shaded region, respectively. Fig. 2(c) plots the 95% CI of the mean trend as well as 20 samples of the mean trend. The mean trend samples are simulated by first generating samples of the coefficient parameters based on their posterior distribution and then multiplying the model matrix for the testing data by the coefficient vector (i.e., $\mathbf{X}\boldsymbol{\beta}$ in Eq. 3). As shown, the 95% CI can cover most sections of the actual mean trend while the 95% PI can cover most of the testing points, indicating the reasonableness of these intervals. Also, the simulated mean curves generally capture the trend of the response variable. The good performance of the 95% CI and 95% PI can also be seen from Table 1, which summarizes the mean values of the 500 coverage percentages for 500 experiments. For comparison, the performance for the prediction interval ignoring the uncertainty in the mean trend is also studied. This prediction interval refers to the interval derived purely from the random error, namely the interval bounded by the mean trend minus and plus twice the standard deviation of the random errors. As shown in Table 1, both the 95% CI and 95% PI have a reasonable coverage percentage, which is close to the confidence level. The 95% PI ignoring the uncertainty in the mean has a coverage percentage slightly smaller than the confidence level. Furthermore, the coverage percentages of the 95% CI and 95% PI for 30 training points were also evaluated, which is summarized in the last row of Table 1. As shown, the coverage percentage of the 95% CI and 95% PI are still close to the confidence level, 95%. However, the 95% PI ignoring the uncertainty in the mean trend just has an average coverage percentage of 86%, which is a little far from the theoretical value, 95%. The reason is that when the data are limited, the uncertainty in the mean trend is relatively large. Fig. 2(d) plots one typical experiment of cross-validation using 30 training data. As shown, the 95% CI in Fig. 2(d) is considerably wider than that in Fig. 2(b). The reason is that when a relatively small number of data points are used, the coefficient parameters β_i have large uncertainty. In other words, relatively large values of the standard deviations of β_i are estimated in this case. The large uncertainty in β_i is propagated to the mean trend, inducing a wider 95% CI of the mean trend. These phenomena highlight the importance of considering the uncertainty in the mean trend when the data are limited.

One issue that may arouse argument is that herein the residual of the simulated data is assumed to be independently distributed rather than autocorrelated. Spatial autocorrelation is well known to be a property of geotechnical or geological properties. To reflect this nature, it may be more reasonable to model the residual of the response variable as autocorrelated random variables. However, we argue that spatial autocorrelation can to some extent be viewed as an artifact generated by the modelers. This is similar to what was stated by Baecher and Christian (2005), i.e., the division of the spatial variation into a mean trend and a residual around the mean is an artifact. As shown by Baecher and Christian (2005), the variance of the residual and the associated autocorrelation function highly depend on the form of the trend function. Hence, it is possible to obtain independent residuals when a suitable trend function is selected. This statement is supported by Qi et al. (2020), which showed that the residual of the rockhead around a mean trend described by a spline function is independent of each other.

4. Spatial prediction using actual borehole data

This section studied the performance of the GAM in dealing with real borehole data. Some borehole data revealing the rockhead of two formations, Bukit Timah Formation and Jurong Formation in Singapore are analyzed. The Bukit Timah Formation data were extracted from 100 boreholes while the Jurong Formation data were from 60 boreholes. The borehole data are extracted from site investigation reports for the construction of two metro lines in Singapore. The former data are located at the Bukit Timah Road while the latter nearby the Buona Vista metro station. For both sets of data, the data are projected to a line that is approximately parallel to the metro line. These borehole data have been elaborated in Qi et al. (2020) and are not repeated herein. Similar to the last section, cross-validation is used to evaluate the accuracy of the prediction and the reasonableness of the CI and PI. Firstly, spatial prediction is performed using all the data. Secondly, the data are divided into two groups and cross-validation is performed.

Note that there are two major differences between this study and the analyses performed by Qi et al. (2020). Firstly, the two studies use different methods to determine the smoothness of the fitted curve, i.e., Qi et al. (2020) using the knot number and location selection while this study using the wiggliness penalty. Secondly, the uncertainty in the mean trend of the rockhead elevation is ignored by Qi et al. (2020) but considered herein.

4.1 Bukit Timah Formation

4.1.1 Analysis using all the data

The rockhead elevation of the Bukit Timah Formation is first predicted using all the borehole data, i.e., rockhead elevation from 100 boreholes. When performing the model fitting, one input parameter is the dimension of the basis used to represent the smooth term, i.e., q in Eq. 2. The dimension of the basis should be large enough to approximate the true, but unknown function of the response parameter (Wood 2017). As a rule of thumb, the dimension of the basis is considered to be sufficient when the smoothness selection criterion converges as the basis dimension increases. Table 2(a) summarizes the GCV scores and estimated standard deviations of the random error for various basis dimensions. As shown, when the basis dimension reaches 60 ~ 80, both the GCV score and the standard deviation of the random error converge. Hence, the basis dimension is set to 80 and the associated fitted GAM model is plotted in Fig. 3(a). In Fig. 3(a), the 95% CI of the mean trend is represented by the shaded area while the 95% PI of the prediction is bounded by the two dashed lines. As shown, the fitted mean trend is generally consistent with that reported in Qi et al. (2020), indicating the effectiveness of the used method. Besides, the 95% CI of the mean trend has different widths at various locations. For instance, the area within the distance range of [700 m, 2000 m] has a narrower confidence interval than its left-hand and right-hand sides. The reason is that the geological conditions on the left-hand and the right-hand sections are more complex than the middle section. This phenomenon shows that the GAM model can provide an uncertainty that well reflects the geological complexity. Based on the SD_{μ} for various points, it is evaluated that the average value of SD_{μ} is 3.9 m. Hence, the average standard deviation of predictions is $\mathrm{SD}_{\mathrm{p}} = \sqrt{\mathrm{SD}_{\mu}^2+\mathrm{SD}_{\varepsilon}^2} = \sqrt{6.0^2+3.9^2} = 7.2$ m.

Fig. 3(b) plots the variation of the $\mathrm{SD}_{\mu}/\mathrm{SD}_{\mathrm{p}}$ with distance. Since SD_{μ} reflects the magnitude of the geological uncertainty and the data quantity, Fig. 3(b) provides quantitative information on the area with large uncertainty in the trend, which indicates relatively large construction risk and requires additional site investigation. Also, it is worth noting that SD_{μ} denotes the epistemic uncertainty in the spatial prediction that can be reduced to 0 when sufficient data are available while SD_{p} denotes the sum of the aleatoric and epistemic uncertainties in the spatial prediction. The ratio of the two indexes indicates the potential for improvement in the spatial prediction accuracy if additional data are available. For example, a relatively large value of $\mathrm{SD}_{\mu}/\mathrm{SD}_{\mathrm{p}}$ suggests that the prediction accuracy can be further improved by using additional data. By contrast, if the value of the ratio approaches 0, there is no need to carry out additional site investigations. As shown in Fig. 3(b), the area within the distance range of [700 m, 2000 m] has a relatively small value of $\mathrm{SD}_{\mu}/\mathrm{SD}_{\mathrm{p}}$, indicating that this area has less complex geological conditions than the remaining areas. Also, some areas have a $\mathrm{SD}_{\mu}/\mathrm{SD}_{\mathrm{p}}$ value larger than 0.6, suggesting that there is a high potential to improve the prediction accuracy.

Furthermore, Fig. 3(c) plots the autocorrelation functions of the residuals of the rockhead elevation, namely the measured rockhead elevation minus the mean trend. The horizontal coordinate, lag, in Fig. 3(c) denotes the difference in serial numbers for a pair of data points. For example, the lag for any pair of neighboring data points is 1 while the lag of the data pairs, (1^{st} data point, 3^{rd} data point), (2^{nd} data point, 4^{th} data point), ⋯, is 2. In other words, the data pairs with the same difference value of the serial number are used to evaluate an autocorrelation coefficient. This is a rough but simple way to check the autocorrelation of the residual. As shown, the autocorrelation coefficient for a lag of 1 is already negative, indicating the residual of the rockhead elevation is independent. This means that there is no need to use a more complex model, such as a mixed model with correlated residuals, to analyze the rockhead data.

Fig 3(c) plots 20 simulated samples of the mean trend of rockhead. Each sample is generated by first simulating samples of the coefficient parameters based on their posterior distribution and then multiplying the model matrix by the simulated coefficient vector, namely $\mathbf{X}\boldsymbol{\beta}$ in Eq. 3. As shown, the generated mean trend samples generally reveal the spatial trend of the rockhead elevation. Besides, a larger variability in the mean trend can be observed in the left-hand and right-hand sides than that in the middle section. These simulated trend curves can be readily used in numerical analyses of geological structures, such as stability analysis of slopes or tunneling. This is one major benefit of using the GAM model. Note that the method in Qi et al. (2020) can also produce samples of the mean trend, but these samples are generated mainly by the bootstrap method, namely performing regression using randomly drawn subsets of borehole data. These mean trend samples are not as accurate as those created by the GAM method as the former uses fewer data.

4.1.2 Cross-validation

This section evaluates the performance of the GAM model using cross-validation. Similar to Qi et al. (2020), 50 data points with even serial numbers are set as training data while the remaining 50 points with odd serial numbers as testing data. As mentioned in section 2, one feature of the GAM is that the users can impose a wiggliness constraint on the fitted curve or assign different weights to the data points. Hence, three prediction schemes are considered herein. Scheme 1 imposes no prior information on the smoothing parameter and weight. Scheme 2 sets the smoothing parameter to be 1.5, which corresponds to a relatively wiggly rockhead profile. Scheme 3 assigns larger weights to several data points at the geologically complex area, namely weights for the 6^{th} to 9^{th}, 43^{rd}, 45^{th}, 47^{th} training points = 2, and weights for the remaining training points = 1. Assigning a weight value of m to a certain data point is equivalent to placing m identical observations at the same location (Wood 2020). In all three schemes, the dimension of the basis is set to be 40, which is obtained using the same procedure as that in section 4.1.1.

The prediction results for various prediction schemes are plotted in Fig. 4. As shown in Fig. 4(a), the predicted mean trend of the rockhead elevation cannot capture the wavy rockhead profile at the two ends of the section when no prior information is imposed on the model. The reason is that the data at the geologically complex area are so limited and the abnormalities in these data are treated as a random error rather than counted into the mean trend. However, the wavy trends can be well captured by schemes 2 and 3. The reason is that the constraint of the smoothing parameter = 1.5 in scheme 2 makes the mean trend wigglier. Moreover, the larger weights of the data points at the geologically complex area amount to manually adding some data to the critical locations which control the shape of the rockhead profile. This phenomenon shows the capability of the GAM to incorporate prior information of the engineers, such as personal judgment or knowledge of the geological information in the investigated area. On the other hand, even if the users do not have any prior information, they can perform some sensitivity analyses and acquire multiple solutions by altering the values of smoothing parameters or weights. These solutions can be subsequently submitted to experienced engineers and an optimal solution can be decided based on their judgment. This feature is quite useful as the solution is a joint product of the GAM method and engineer judgment. Also, by providing the clients multiple scenarios of the possible rockhead profile, the risks in the construction can be better appreciated and managed.

The prediction accuracies for the three considered schemes as well as the results in Qi et al. (2020) are summarized in Table 2(b), including the mean and standard deviation (SD) of the prediction errors, the estimated SD of the random error, SD_ε, the mean width of the 95% CI and 95% PI. The mean and SD of the prediction error are evaluated from the 50 errors for the 50 testing points. The width of the 95% CI is four times the SD of the mean trend, $4\mathrm{SD}_\mu$ while the width of 95% PI is four times the SD of a prediction, i.e., $4\mathrm{SD}_\mathrm{p} = 4\sqrt{\mathrm{SD}_\mu^2 + \mathrm{SD}_\varepsilon^2}$, as illustrated in section 2.4. Since Qi et al. (2020) did not consider the uncertainty in the mean trend, the width of the 95% PI is set to be $4\mathrm{SD}_\varepsilon$ in the last row. As shown in Table 2(b), all the three considered schemes have slightly smaller prediction errors than that in Qi et al. (2020).

The two schemes implementing prior information have higher accuracy than the one without prior information. This result shows the usefulness of imposing human judgment on the GAM model. Such incorporation of human judgment is lacking in the method in Qi et al. (2020). It also well demonstrates the idea that the data-driven method should incorporate engineering judgment rather than replace engineering judgment, as discussed by Phoon, Ching, and Shuku (2021). Besides, the 95% PIs produced by the GAM model are wider than that reported in Qi et al. (2020) because the uncertainty in the mean trend is considered by the GAM. The former is more rational than the latter, which can be shown by the coverage percentage of the 95% PI. In Qi et al. (2020), only 42 out of 50 (i.e., 84%) testing data were covered by the 95% PI. But for the GAM, 46, 46 and 45 out of the 50 (i.e., 92%, 92%, 90%) data points are covered by the 95% PIs for the three schemes, respectively. This observation justifies the consideration of the uncertainty in the mean trend. The result indicates that the method of Qi et al. (2020) underestimates the prediction uncertainty and induces unsafe designs of geotechnical structures or an underestimation of the underground construction risk. By contrast, the GAM method in this study can reasonably quantify the prediction uncertainty, which leads to a reasonable design of geotechnical structures or underground construction scheme.

4.2 Jurong Formation

To further illustrate the performance of the GAM, this section performs the spatial prediction using the rockhead data of the Jurong Formation. For this case, only 60 data points are distributed along a line with a length of around 3400 m. Such limited data make spatial prediction more challenging. Fig. 5(a) plots the fitted mean trend for the scheme imposing no prior information (referred to as scheme 1) while Fig. 5(b) plots the result for the scheme with the smoothing parameter set to be 5 (referred to as scheme 2). Both schemes use all the data points and a basis dimension of 50, which is selected based on the same method as that in section 4.1.1. Details for the selection of the basis dimension are summarized in Table 3(a). As shown in Fig. 5(a, b), scheme 2 captures more local fluctuation in the mean trend of the rockhead than scheme 1.

Cross-validations of the spatial prediction of Jurong Formation rockhead are performed using 40 training points and 20 testing points. The cross-validation also adopts two schemes, scheme 1 imposing no prior information and scheme 2 setting the smoothing parameter to be 5. The two schemes use the same training data, testing data and basis dimension, i.e., 30. The predicted rockhead profile and associated 95% CI, 95% PI are plotted in Fig. 5(c, d). For comparison, the mean trend obtained from all the data points and scheme 2 is plotted in both Fig. 5(c) and (d). The associated prediction errors are summarized in Table 3(b). Similar phenomena as those in section 4.1 can be observed in Fig. 5 and Table 3(b). These include (i) imposing some prior information involving the wiggliness of the rockhead profile produces a more accurate estimation of the mean trend when the data quantity is limited, (ii) the GAM model can quantify the uncertainty in the mean trend of rockhead, making the 95% PI wider than that reported in Qi et al. (2020).

5. Conclusions

This study uses the generalized additive model (GAM) for the spatial prediction of the interfaces of geological formations. The performance of the GAM is evaluated using both simulated data and actual borehole data for two geological formations in Singapore. The prediction accuracy, the rationality of the 95% confidence intervals for the mean trend and the

95% prediction interval of the prediction are assessed. The benefits of the GAM are summarized as follows.

(1) The GAM can produce a reasonable 95% confidence interval and 95% prediction interval as the analyses using the simulated data show that on average the 95% confidence interval covers 94% or 91% of the actual mean trend while the 95% prediction interval covers 94% or 92% of testing data. Ignoring the uncertainty in the mean produces a 95% prediction interval with an unreasonably low coverage percentage. Furthermore, samples of the rockhead profile such as the grey curves in Fig. 3(d), can be generated by the GAM. These samples can be viewed to be possible rockhead profiles and be used in future numerical analyses of geotechnical structures.

(2) Both the epistemic uncertainty and aleatoric uncertainty in the spatial predictions can be quantified by the GAM method. The former refers to the uncertainty in the mean trend and can be reduced if additional data are available. This uncertainty is affected by the geological complexity and the data quantity. The latter refers to random errors caused by various factors such as engineers' subjective judgments of weathering degrees of the geological layers. The aleatoric uncertainty cannot be reduced and represents the minimum error that can be achieved in spatial predictions. The relative magnitudes of the two uncertainties can be quantified by the ratio of the standard deviation of the mean trend and the standard deviation of predictions. From the ratio values, it is easy to determine the areas with complex geological conditions, which need additional site investigations. It is also easy to check the potential for improvement in the prediction accuracy when additional data are available because in theory the minimum value of the ratio is 0.

(3) The users can apply expert judgment or knowledge on the geological profile to the model by setting suitable values of the smoothing parameter or assigning suitable weights to the data points at critical locations. The analyses using the actual data in the two cases show that the use of expert knowledge improves the prediction accuracy and makes the resulting geological profile more consistent with that obtained from more data.

(4) Due to the large variability of the rockhead locations, the prediction error of the rockhead elevation is still relatively large. In the future, it is of interest to use additional data such as geophysical data to further increase the prediction accuracy and reduce the prediction uncertainties.

Acknowledgment

The second author would like to thank the financial support from the National Natural Science Foundation of China (No. 52109144) and the Open Innovation Fund of Changjiang Institute of Survey, Planning, Design and Research (No. CX2020K07).

Reference

Aswar, D. S. , and P. B. Ullagaddi. 2017. "An overview of 3-D geological modelling part II. Summary of major 3-d geological modelling methodologies." International Journal of Latest Engineering and Management Research 2 (11):15-27.

Baecher, G. B., and J. T. Christian. 2005. Reliability and statistics in geotechnical engineering: John Wiley & Sons.

Baecher, G. B., and J. T. Christian. 2008. "Spatial variability and geotechnical reliability." In Reliability-based design in geotechnical engineering, 88-145. CRC Press.

Bordoni, M., Y. Galanti, C. Bartelletti, M. G. Persichillo, M. Barsanti, R. Giannecchini, G. D.

Avanzi, et al. 2020. "The influence of the inventory on the determination of the rainfall-induced shallow landslides susceptibility using generalized additive models." CATENA 193:104630. doi: https://doi.org/10.1016/j.catena.2020.104630.

Burke, H.F., J.R. Ford, L. Hughes, S. Thorpe, and J.R. Lee. 2017. "A 3D geological model of the superficial deposits in the Selby area." In Groundwater programme commissioned report CR/17/112. Nottingham British Geological Survey.

de Brogniez, D., C. Ballabio, A. Stevens, R. J. A. Jones, L. Montanarella, and B. van Wesemael. 2015. "A map of the topsoil organic carbon content of Europe generated by a generalized additive model." European Journal of Soil Science 66 (1):121-34. doi: https://doi.org/10.1111/ejss.12193.

De Boor, C. 2001. A practical guide to splines. Revised Edition. New York: Springer-Verlag

Goetz, J. N., A. Brenning, H. Petschko, and P. Leopold. 2015. "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling." Computers & Geosciences 81:1-11. doi: https://doi.org/10.1016/j.cageo.2015.04.007.

Gong, X., A. Kaulfus, U. Nair, and D. A. Jaffe. 2017. "Quantifying O3 impacts in urban areas due to wildfires using a generalized additive model." Environmental Science & Technology 51 (22):13216-23. doi: 10.1021/acs.est.7b03130.

Gong, W. P., C. Zhao, C. H. Juang, H.M. Tang, H. Wang, and X.L. Hu. 2020. "Stratigraphic uncertainty modelling with random field approach." Computers and Geotechnics 125:103681.

Hastie, T. J., and R. J. Tibshirani. 1986. "Generalized Additive Models." Statistical Science 1 (3):297-310, 14.

Hastie, T. J., and R. J. Tibshirani. 1990. Generalized additive models. CRC press.

Huang, Y., J. L. Beck, S. Wu, and H. Li. 2014. "Robust Bayesian compressive sensing for signals in structural health monitoring." Computer-Aided Civil and Infrastructure Engineering 29 (3):160-79.

Khoda Bakhshi, A., and M. M. Ahmed. 2021. "Real-time crash prediction for a long low-traffic volume corridor using corrected-impurity importance and semi-parametric generalized additive model." Journal of Transportation Safety & Security:1-35. doi: 10.1080/19439962.2021.1898069.

Lark, R.M., S.J. Mathers, S. Thorpe, S.L.B. Arkley, D.J. Morgan, and D.J.D. Lawrence. 2013. "A statistical assessment of the uncertainty in a 3-D geological framework model." Proceedings of the Geologists' Association 124 (6):946-58.

Li, J.H., Y.M. Cai, X.Y. Li, and L.M. Zhang. 2019. "Simulating realistic geological stratigraphy using direction-dependent coupled Markov chain model." Computers and Geotechnics 115:103147.

Liu, L. L., Y. M. Cheng, Q. J. Pan, and D. Dias. 2020. "Incorporating stratigraphic boundary uncertainty into reliability analysis of slopes in spatially variable soils using one-dimensional conditional Markov chain model." Computers and Geotechnics 118:103321.

Ma, Y., B. Ma, H. Jiao, Y. Zhang, J. Xin, and Z. Yu. 2020. "An analysis of the effects of weather and air pollution on tropospheric ozone using a generalized additive model in Western China: Lanzhou, Gansu." Atmospheric Environment 224:117342. doi: https://doi.org/10.1016/j.atmosenv.2020.117342.

Mariethoz, G., and J. Caers. 2014. Multiple-point geostatistics: stochastic modeling with training images: John Wiley & Sons.

Phoon, K.K., J.Y. Ching, and T. Shuku. 2021. "Challenges in data-driven site characterization." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards:1-13. doi: 10.1080/17499518.2021.1896005.

Qi, X.H., D.Q. Li, K.K. Phoon, Z.J. Cao, and X.S. Tang. 2016. "Simulation of geologic uncertainty using coupled Markov chain." Engineering Geology 207:129-40. doi: 10.1016/j.enggeo.2016.04.017.

Qi, X.H., and H.X. Liu. 2019. "An improved global zonation method for geotechnical parameters." Engineering Geology 248:185-96. doi: 10.1016/j.enggeo.2018.11.013.

Qi, X.H., X.H. Pan, K. Chiam, Y. S. Lim, and S. G. Lau. 2020. "Comparative spatial predictions of the locations of soil-rock interface." Engineering Geology 272:105651. doi: https://doi.org/10.1016/j.enggeo.2020.105651.

Qi, X., H. Wang, J. Chu, and K. Chiam. 2021a. “Effect of autocorrelation function model on spatial prediction of geological interfaces.” Canadian Geotechnical Journal, accepted, 10.1139/cgj-2020-0644.

Qi, X.H., H. Wang, X.H. Pan, J. Chu, and K. Chiam. 2021b. "Prediction of interfaces of geological formations using the multivariate adaptive regression spline method." Underground Space 6(3): 252-66.

Ruppert, D., M. P. Wand, and R. J. Carroll. 2003. Semiparametric regression: Cambridge university press.

Simpson, G. L. 2018. "Modelling palaeoecological time series using generalised additive models." Frontiers in Ecology and Evolution 6:149.

Smirnoff, A., E. Boisvert, and S. J. Paradis. 2008. "Support vector machine for 3D modelling from sparse geological information of various origins." Computers & Geosciences 34 (2):127-43.

Wang, H., J. F. Wellmann, Z. Li, X.R. Wang, and R. Y. Liang. 2017. "A segmentation approach for stochastic geological modeling using hidden Markov random fields." Mathematical Geosciences 49 (2):145-77. doi: 10.1007/s11004-016-9663-9.

Wang, X.R., H. Wang, R. Y. Liang, H.H. Zhu, and H.G. Di. 2018. "A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data." Structural Safety 70:128-38. doi: https://doi.org/10.1016/j.strusafe.2017.10.011.

Wang, Y, and T.Y. Zhao. 2016. "Interpretation of soil property profile from limited measurement data: a compressive sampling perspective." Canadian Geotechnical Journal 53 (9):1547-59. doi: 10.1139/cgj-2015-0545.

Wang, Y. , and T. Zhao. 2017. "Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling." Géotechnique 67 (6):523-36. doi: 10.1680/jgeot.16.P.143.

Wang, Y., O. V. Akeju, and T.Y. Zhao. 2017. "Interpolation of spatially varying but sparsely measured geo-data: A comparative study." Engineering Geology 231:200-17. doi: https://doi.org/10.1016/j.enggeo.2017.10.019.

Wang, Y., T.Y. Zhao, and K.K. Phoon. 2018. "Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation." Canadian Geotechnical Journal 55 (6):862-80.

Wang, Yu, T.Y. Zhao, Y. Hu, and K.K. Phoon. 2019. "Simulation of random fields with trend from sparse measurements without detrending." Journal of Engineering Mechanics 145 (2):04018130.

Wood, S. N. 2017. Generalized additive models: an introduction with R: CRC press.

Wood, S. N. 2020. Package 'mgcv', Mixed GAM computation vehicle with automatic smoothness estimation. R package version.

Yee, T. W., and N. D. Mitchell. 1991. "Generalized additive models in plant ecology." Journal of Vegetation Science 2 (5):587-602. doi: https://doi.org/10.2307/3236170.

Zhao, J., B. B. Broms, Y. Zhou, and V. Choa. 1994. "A study of the weathering of the Bukit Timah Granite Part a: Review, field observations and geophysical survey." Bulletin of the International Association of Engineering Geology 49 (1):97-106. doi: 10.1007/BF02595006.

Zhao, T.Y., Y. Hu, and Y. Wang. 2018. "Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling." Engineering Geology 246:162-75. doi: https://doi.org/10.1016/j.enggeo.2018.09.022.

Zhao, C., W.P. Gong, T.Z. Li, C. H. Juang, H.M. Tang, and H. Wang. 2021. "Probabilistic characterization of subsurface stratigraphic configuration with modified random field approach." Engineering Geology 288:106138.

Zhou, Y.X., and X.P. Wu. 1994. "Use of neural networks in the analysis and interpretation of site investigation data." Computers and Geotechnics 16 (2):105-22.

Table 1 Mean values of the coverage percentages for the simulated data

Number of training point	Mean of coverage percentage for the 95% CI	Mean of coverage percentage for the 95% PI	Mean of coverage percentage for the 95% PI ignoring the uncertain in the mean trend
50	0.94	0.94	0.91
30	0.91	0.92	0.86

Table 2 Spatial prediction of the rockhead elevation for the Bukit Timah Formation

(a) Selection of the dimension of basis for all the data

Basis dimension	20	40	60	80
GCV	74.9	60.9	58.8	58.5
Estimated standard deviation of the random error (m)	8.0	6.5	6.1	6.0

(b) Prediction results of the cross-validation

Fitting scheme	Mean of prediction error (m)	[a]SD of prediction error (m)	Estimated [a]SD of random error (m)	Mean width of 95% CI (m)	Mean width of 95% PI (m)
[b]Scheme 1	1.8	8.5	8.2	14.7	36.0
[c]Scheme 2	1.6	7.5	6.8	21.8	35.0
[d]Scheme 3	1.5	7.6	7.5	19.6	35.9
Qi et al. (2020)	2.0	9.0	6.1	–	24.4

Note: a: SD = standard deviation;

b: scheme 1 imposes no prior information on the smoothness parameter or weights;

d: scheme 2 sets the smoothness parameter to be 1.5;

d: scheme3 gives more weights to the data points at the geologically complex area, namely weight for the 6th to 9th, 43rd, 45th, 47th training points = 2, and weight for the remaining training data = 1.

Table 3 Spatial prediction of the rockhead elevation for the Jurong Formation

(a) Selection of the dimension of basis using all the data

Basis dimension	20	30	40	50
GCV	83.4	86.4	83.4	83.2
Estimated standard deviation of the random error (m)	8.2	7.0	8.2	8.2

(b) Prediction results of the cross-validation

Fitting scheme	Mean of prediction error (m)	[a]SD of prediction error (m)	Estimated SD of random error (m)	Mean width of 95% CI (m)	Mean width of 95% PI (m)
[b]Scheme 1	-2.0	10.6	11.4	22.1	50.9
[c]Scheme 2	-2.3	10.5	10.5	30.6	52.2
Qi et al. (2020)	-3.1	12.7	8.1	–	32.4

Note: a: SD = standard deviation;

b: scheme 1 imposes no prior information on the smoothness parameter or weights;

c: scheme 2 sets the smoothness parameter to be 5.

Figure captions

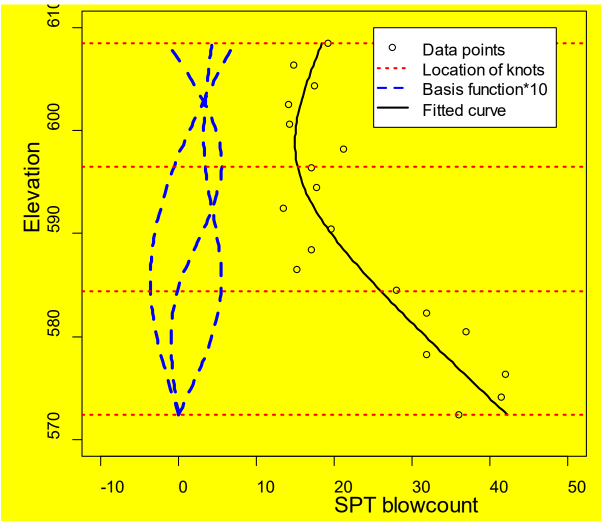
Fig. 1 Example of cubic spline basis and the fitted smooth curve

Fig. 2 Prediction using simulated data

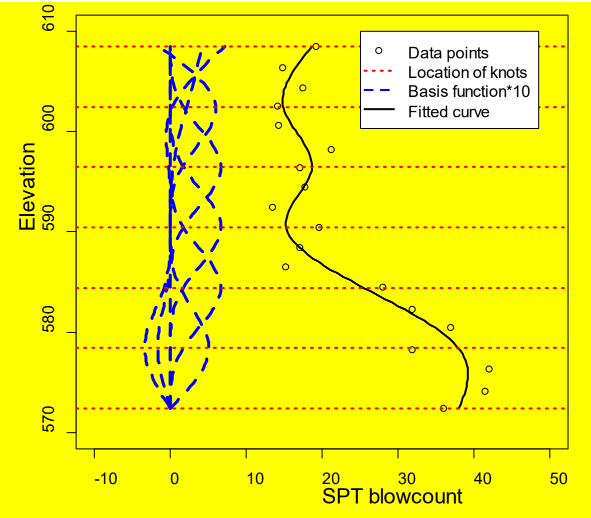
Fig. 3 Spatial prediction of rockhead elevation for the Bukit Timah Formation using all the 100 borehole data

Fig. 4 Cross-validation for the spatial prediction of the rockhead elevation for the Bukit Timah Formation

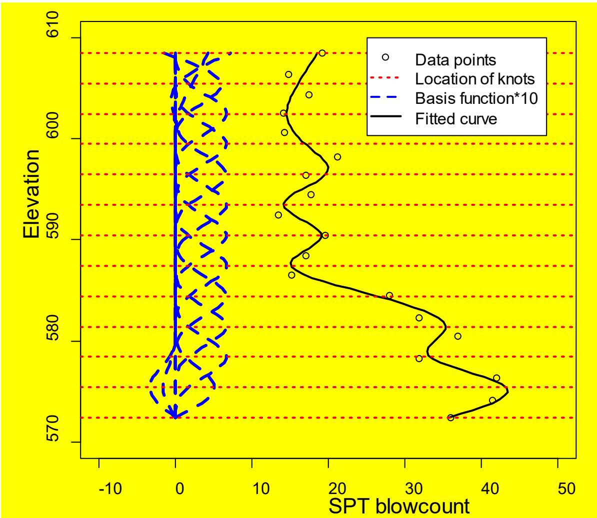
Fig. 5 Spatial prediction of the rockhead elevation for the Jurong Formation



(a) Three cubic basis functions

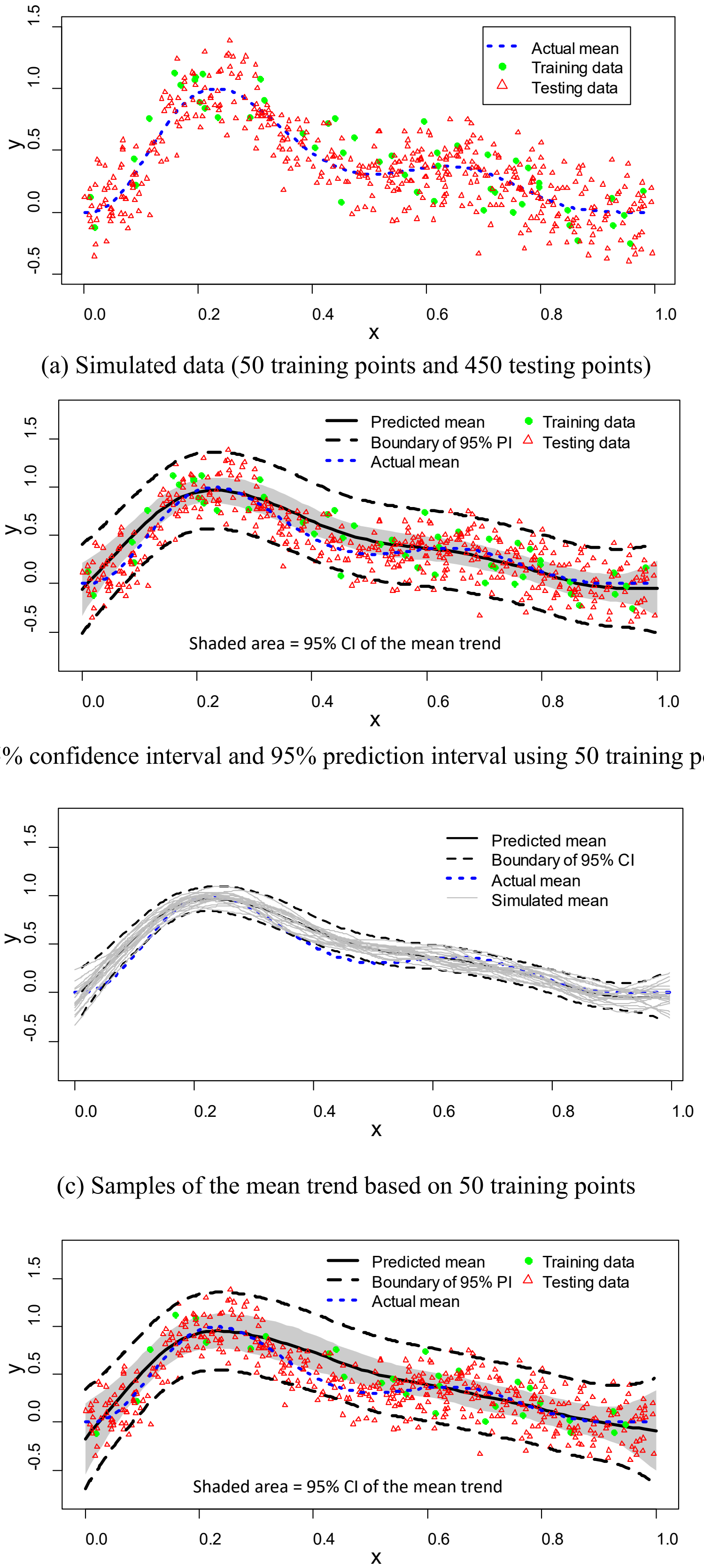


(b) Six cubic basis functions



(c) Twelve basis functions

Fig. 1 Example of cubic spline basis and the fitted smooth curve



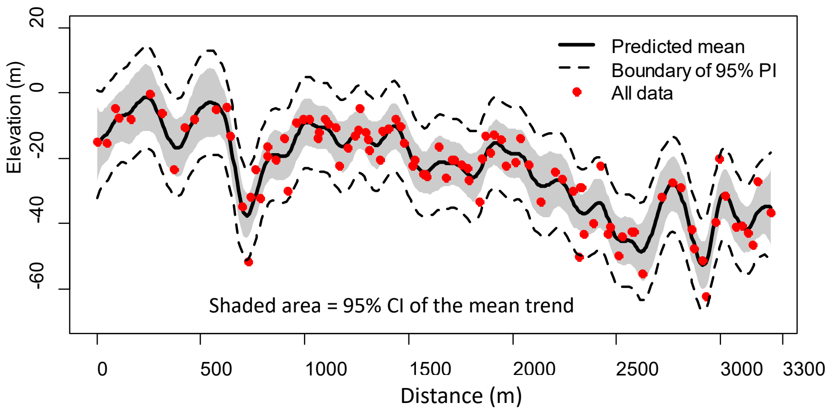
(a) Simulated data (50 training points and 450 testing points)

(b) 95% confidence interval and 95% prediction interval using 50 training points

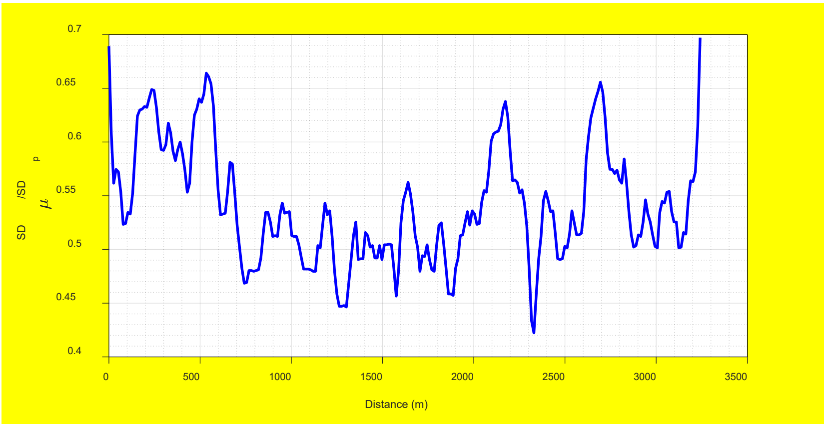
(c) Samples of the mean trend based on 50 training points

(d) 95% confidence interval and 95% prediction interval using 30 training points

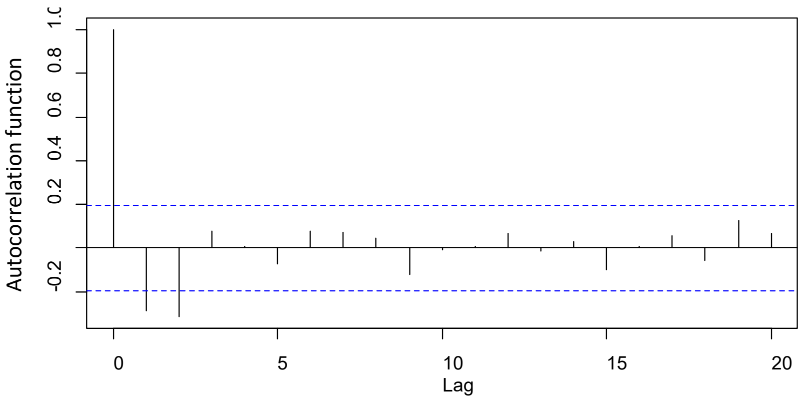
Fig. 2 Prediction using simulated data



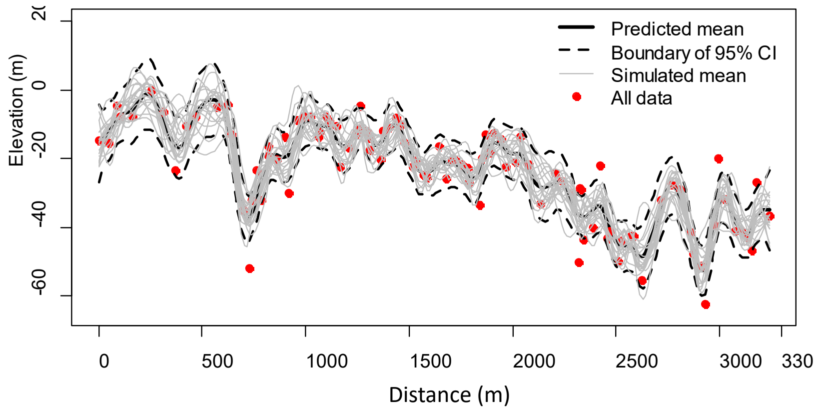
(a) Predicted mean trend, 95% confidence interval, and 95% prediction interval



(b) Ratio of SD_{μ} to SD_{p}

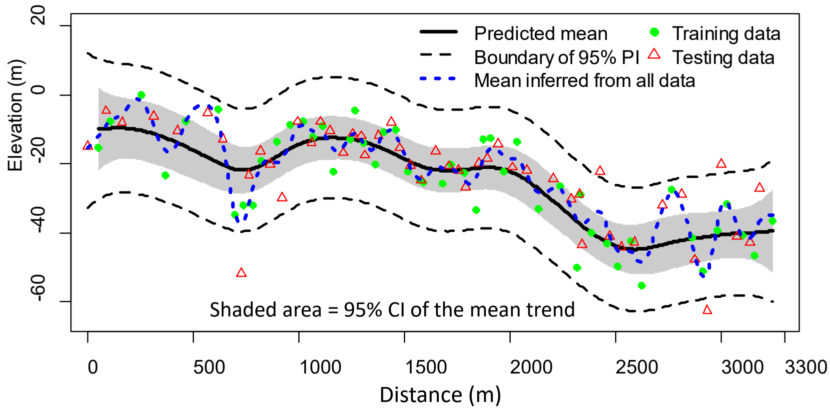


(c) Autocorrelation of the residual of rockhead elevation

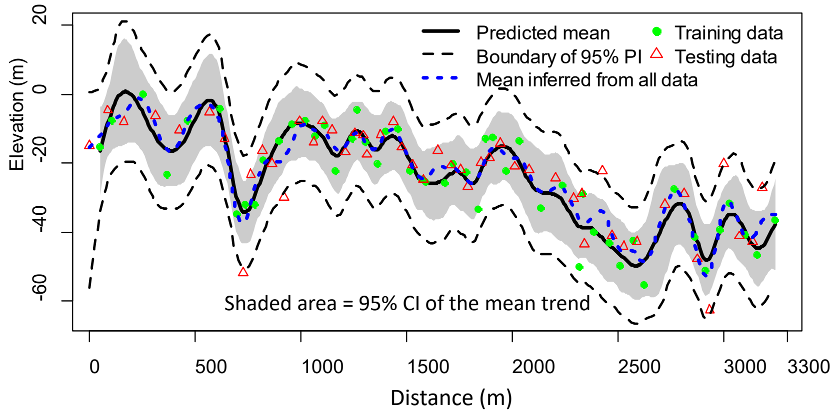


(d) Sample of the mean trend

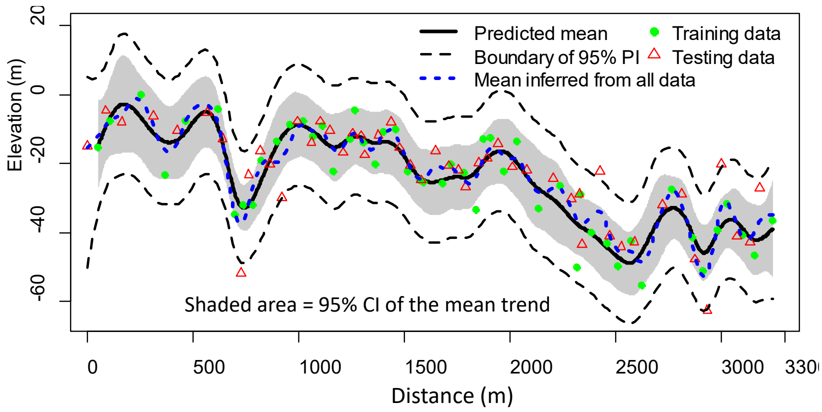
Fig. 3 Spatial prediction of rockhead elevation for the Bukit Timah Formation using all the 100 borehole data



(a) Scheme 1 (no prior information in the smoothness parameter or weights)

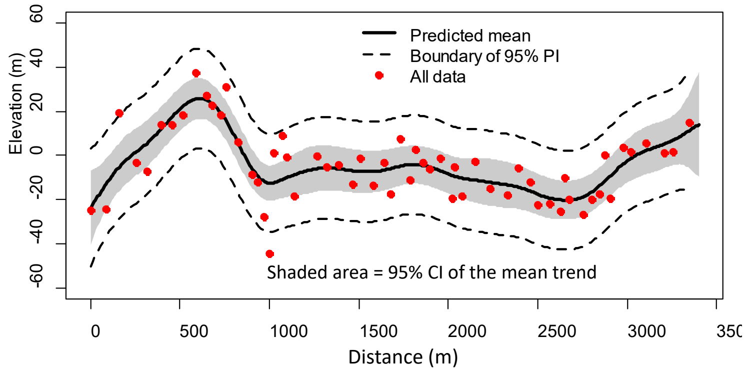


(b) Scheme 2 (the smoothness parameter fixed at 1.5)

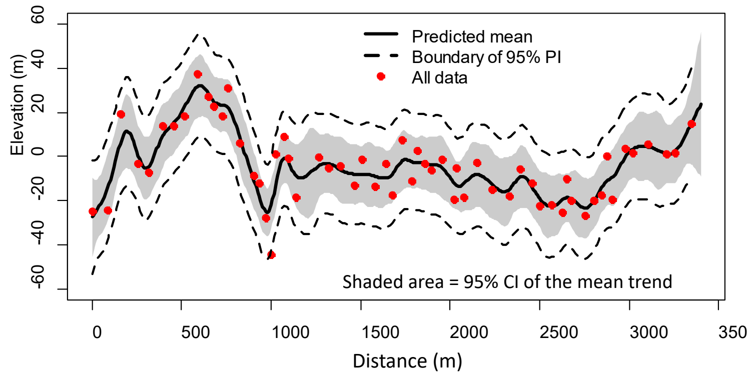


(c) Scheme 3 (Weight of 6th ~ 9th, 43rd, 45th, 47th training points = 2, and weight of remaining training points = 1)

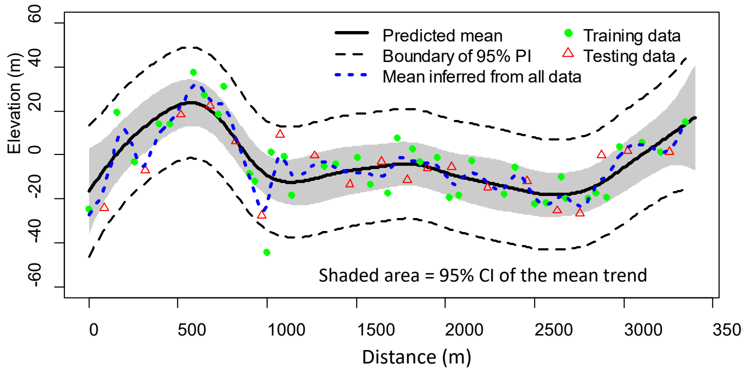
Fig. 4 Cross-validation for the spatial prediction of the rockhead elevation for the Bukit Timah Formation



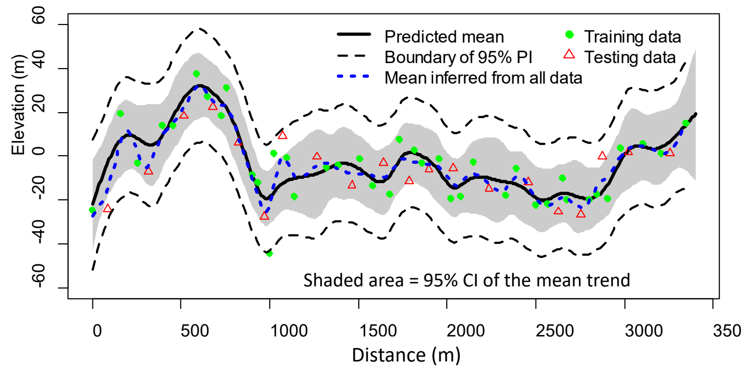
(a) Prediction using all the data (scheme 1: no prior information in the smoothness parameter)



(b) Prediction using all the data (scheme 2: smoothing parameter = 5)



(c) Cross-validation (scheme 1: no prior information in the smoothness parameter)



(d) Cross-validation (scheme 2: smoothing parameter = 5)

Fig. 5 Spatial prediction of the rockhead elevation for the Jurong Formation