Estimating pile length uncertainty with Kriging

Martti Hallipelto, M.Sc., Ramboll Finland Oy
PL 25, 02601 Espoo
Ari Hartikainen, M.Sc., Ramboll Finland Oy
PL 25, 02601 Espoo
Taavi Dettenborn, M.Sc., Ramboll Finland Oy
PL 25, 02601 Espoo
martti.hallipelto@ramboll.fi

Abstract. Piles are used to transfer loads from the structure into the bearing soil layer. The accurate estimation of the pile lengths is a complex statistical problem that includes several spatial and measurement-based uncertainties. Bearing subsurface is modeled with linear triangular networks, and geometry related uncertainty is considered with overall safety factor. Currently, it is common practice is to solve the unknowns in this problem with simplistic borehole analysis and engineering judgment. Our paper presents a statistical pile length model based on the borehole uncertainty analysis. Modern geotechnical designing contains multiple steps including ground investigations, computer modeling, and expert decision making. All the included steps add uncertainty to the design process. The problem with these traditional methods is that the sources and the magnitudes of the uncertainty are not visible to the designer, which can lead to non-optimal decisions. With the help of the state-of-the-art probabilistic models, a design pipeline can be constructed that propagates the uncertainty from process to process and is transparent about its sources and magnitudes. In our paper, we show how the probabilistic pipeline approach enables us to make more informed decisions. This is due to uncertainty propagation from ground investigations to the 3D volumetric model. Given the uncertainty, we can estimate the mean and variance of the future costs and the impact that different design decisions have. The paper presents a practical example where probabilistic models are utilized to improve decision making.

1. Introduction
Commonly piles are used to transfer loads from the structure into the bearing soil layer. The accurate estimation of the pile lengths is a complex statistical problem that includes several spatial and measurement-based uncertainties. Bearing subsurface can be modeled with simple geometrical objects such as linear triangular networks. The uncertainty related to the geometry is usually considered in other phases in the analysis pipeline, such as overall safety factor. Currently, it is common practice to solve the unknowns in this problem with simplistic borehole analysis and engineering judgment.

Modern geotechnical designing pipeline contains multiple steps including ground investigations, computer modeling, and expert analysis and decision making. All the included steps add uncertainty to the design process pipeline. The problem with these traditional methods is that the sources and the magnitudes of the uncertainties are not visible to the designer, which can lead to sub-optimal decisions.
The aim of this paper is to present the workflow to handle uncertainty sources correctly with statistical uncertainty modeling. Statistical modeling is flexible and enables the modeler to select the complexity based on the needed application. Each situation can be modeled with the bespoke statistical model to include optimal amount of information.

2. Theory

2.1. Gaussian process

Interpolation of spatially distributed data is usually modeled with non-linear regression, where covariates $x$ are used to model response $y$ for $n$ observed covariate-response pairs.

$$y_i = f(x_i) + \epsilon_i$$

where covariate $x_i$ is a vector of length $p$ and the error $\epsilon_i$ and $y_i$ are scalars. The Gaussian Process Regression, GPR, is a common non-parametric model. A subset of GPR models are commonly known as Kriging models, but in this paper, we are using the wider definition. We define the basics for the GPR, but a more in-depth introduction to the subject can be read in [1, 2]. In the standard GPR, the model residuals are assumed to be independent and identically distributed with a Gaussian distribution of mean 0 and constant variance $\sigma^2$ [3].

The Gaussian Process model is defined with mean and covariance functions. Usually, the mean function is kept simple and all the information is incorporated with the covariance function. The standard GPR is shown in equations 2 and 3. The measurement variable vector $y$ is modeled as a multivariate normal distribution with a simple mean vector $\mu$ and covariance matrix $\Sigma$.

$$y \sim N(\mu, \Sigma)$$

(2)

Common simple model information such as bias and linear regression models can be incorporated with mean function, but usually mean vector $\mu$ is defined as 0 vectors More complex models are also possible, but usually, it is preferred to let covariance function to handle these interactions.

Covariance functions $\Sigma$ are usually defined with hyperparameters which are estimated from the data. The covariance for $y$ can be defined as

$$\Sigma_{i,j} = Cov(y_i, y_j) = K(x_i, x_j) + \delta_{ij}\sigma^2$$

(3)

where it is a sum of a covariance matrix for $x$ and $\sigma^2$ against the delta function $\delta$, which is 1 if $i=j$ else 0. Heteroskedastic noise model for input data can be included if the noise vector $\sigma^2$ is used instead of a scalar value $\sigma^2$.

The covariance matrix is positive definite and can be created by combining multiple covariance matrices. Our paper uses the common exponentiated quadratic covariance function with hyperparameters for length scale $\rho$ and marginal standard deviation $\alpha$, see equation 4.

$$K(x_i, x_j) = \alpha^2 \exp \left(-\frac{(x_i-x_j)^2}{\rho^2}\right)$$

(4)

The length scale $\rho$ can be seen to describe the smoothness of the solution and marginal standard deviation $\alpha$ describes the variation in the solution amplitude. After the hyperparameters are solved, the solved results can be used to predict new values. To predict location $x_k$ given the solved hyperparameters and data, one can calculate the conditional mean and covariance seen in equations 5 and 6.

$$\bar{\mu} = \mu(x_k) - \Sigma_{x_k, x} \Sigma_x^{-1} (y - \mu(x))$$

(5)
where $\Sigma_{x_k,x}$ is $K(x_k, x_i)$ for $i$ in $1 \ldots n$.

$$\tilde{\Sigma} = \Sigma_k - \Sigma_{x_k,x} \Sigma_x^{-1} \Sigma_{x_k}$$ \hspace{1cm} (6)

Conditional mean and covariance can be used to sample from the conditional distribution.

3. Case

The case study area is in Southern Finland. The construction plans show a total of 749 piles. The area also contains 10 test piles which are used to estimate the virtual bearing soil layer depth.

The topmost soil layer consists of a clay layer that has an average thickness of 15 meters based on the borehole measurements. A sand or moraine surface is found on top of the bedrock surface layer with an average thickness of 2 meters. It is probable that the virtual bearing soil layer is close to the layer edge between clay and sandy layers.

Based on the spatial locations for the test piles as the covariates $x$ and their relative distance, length-scale $\rho$ was given relatively high normal prior helping to resolve the smooth result. Marginal standard deviation $\alpha$ was fixed to 4.0 which is a reasonable estimation based on the prior predictive checks. The prior for sigma was also fixed and calculated as a function of the relative depth from the surface following [4].

The GPR hyperparameter for length-scale $\rho$ was solved with Stan probabilistic language using the default sampler dynamical Hamiltonian Monte Carlo, HMC, method with no U-turn, NUTS [5, 6]. The solved mean value for $\rho$ was 12.235 which is reasonable estimation. For the purpose of simplifying the prediction, only the mean value from posterior density was used for the prediction. For this application the simplification was suitable, but it is recommended to use the whole posterior distribution.

Predictions for pile lengths were done for each planned location and a total of 10000 samples were draw from the conditional distribution.

4. Results

![Map showing the planned and test piles and section location. The contour values show the mean virtual bearing soil layer depth estimate and the contour values describe the absolute height in meters from the sea-surface. The ground surface is near +112 meters.](image-url)
The resulting conditional mean surface is shown as a contoured surface in Figure 1 with the location for the cross-section A-A. The map also shows the locations for the test and the planned piles. The cross-section A-A is shown in Figure 2 and Figure 3 where the first shows the predicted mean surface with 5% - 95% uncertainty envelope and the latter shows the prediction samples which describe the possible virtual bearing soil layer surfaces that are described by our model. The elevation in the figures are shown as the absolute elevation from the sea-surface.

**Cross section A – A**

![Cross section A – A](image)

Figure 2. Cross-section A-A with the results of test piles shown on top of the estimated pile depth.

**Cross section A – A**

![Cross section A – A](image)

Figure 3. Cross-section A-A with the five sampled predictions from conditional distribution.

To estimate the total material flow, 10000 predictions were calculated for each of the 749 piles and for each prediction the total length of piles was calculated. The distribution of a total sum of piles is shown in Figure 4, which shows that the 5%-95% uncertainty envelope estimates the total sum to lie between 8.55 to 10.07 kilometers with the mean value of 9.31 kilometers.
Figure 4. Total pile length distribution in kilometers with 5% and 95% uncertainty envelope.

5. Conclusions and discussion
We have shown in the paper a successful use of a statistical model to estimate pile lengths by incorporating uncertainty. The obtained results can be further processed and new estimates for different metrics can be calculated based on the prediction dataset. This methodology enables decision-makers to do a probabilistic evaluation with the specified risk level.

Further work with the probabilistic models enables advanced decision making for planning and construction purposes. These include optimization of the material workflow to minimize the material costs, optimization of the number and locations for new soundings.

The presented statistical model was kept simple, but with suitable prior information considering the spatial location, the model can be made more complicated with updates on the mean and covariance functions. These updates might include modeling the discontinuity locations, the use of adding a heteroskedastic likelihood model and more informed mean function.
References

[1] C. E. Rasmussen and K. I. Williams, Gaussian Processes for Machine Learning, Massachusetts: the MIT Press, 2006.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, Bayesian Data Analysis 3rd edition, Boca Raton: CRC Press, 2014.

[3] W. Chunyi and N. M. Radford, "Gaussian Process Regression with Heteroscedastic or Non-Gaussian Residuals," arXiv:1212.6246 [stat.ML], 2012.

[4] M. Hallipelto, Pohjasuhteiden epävarmuuden mallintaminen Kriging-menetelmällä, Espoo: http://urn.fi/URN:NBN:fi:aalto-201809034884, 2018.

[5] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li and A. Riddell, "Stan: A probabilistic programming language," Journal of Statistical Software, vol. 76, no. 1, 2017.

[6] Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, Version 2.22, 2020.