A Machine Learning Approach for Characterizing Soil Contamination in the Presence of Physical Site Discontinuities and Aggregated Samples

Alyssa Ngu-Oanh Quach[a], Lucie Tabor[a], Dany Dumont[b], Benoit Courcelles[a], James-A. Goulet[a,*]

[a] *Polytechnique Montreal, CANADA*
[b] *WSP — Parsons Brinckerhoff Engineering Services, CANADA*

Abstract

Rehabilitation of contaminated soils in urban areas is in high demand because of the appreciation of land value associated with the increased urbanization. Moreover, there are financial incentives to minimize soil characterization uncertainties. Minimizing uncertainty is achieved by providing models that are better representation of the true site characteristics. In this paper, we propose two new probabilistic formulations compatible with Gaussian Process Regression (GPR) and enabling (1) to model the experimental conditions where contaminant concentration is quantified from aggregated soil samples and (2) to model the effect of physical site discontinuities. The performance of approaches proposed in this paper are compared using a Leave One Out Cross-Validation procedure (LOO-CV). Results indicate that the two new probabilistic formulations proposed outperform the standard Gaussian Process Regression.

1. Introduction

Rehabilitation of contaminated soils in urban areas is in high demand because of the appreciation of land value associated with the increased urbanization. A common technique to rehabilitate a contaminated site is to remove contaminated soil and either treat or burry it in designated sites. Because there are important costs associated with this activity, it is essential to characterize spatial contaminant concentration in order to classify soil as

*Corresponding author
Email address: james.a.goulet@gmail.com (James-A. Goulet)

either *contaminated* or *non-contaminated* based on the applicable legislation. Any cubic meter unnecessarily removed (i.e. *false*+) or any cubic meter wrongly left in place (i.e. *false*−) will increase the overall rehabilitation costs. Thus, there are financial incentives to minimize soil characterization uncertainties.

In the field of geostatistics, several researchers such as Boudreault et al. [1] and Goovaerts [2, 3] have employed the *Kriging* theory to characterize the spatial distribution of contaminant concentration. Historically, *Kriging* was proposed by Krige and later formalized by Matheron [4]. More recently, the research community has turned toward Machine Learning methods [5]. Most researchers in this field have employed Artificial Neural Networks (ANN) [6, 7, 8, 9]. ANN is a powerful tool, however it requires lots of data (up to millions of data points) to perform well [10]. This condition is seldom met in practice. In the field of *Machine learning*, other techniques analogous to Kriging have recently been the object of numerous publications under the name of *Gaussian Process Regression*, (GPR) [11]. Authors such as MacKay [12] and Rasmussen & Williams [13] have presented modern techniques to calibrate parameters efficiently, process small and large datasets, and provide enhanced formulations that increase the robustness toward numerical instabilities. These latest developments are implemented in several open-source packages such as *GPML* (Gaussian Process Machine Learning) [14] and *GPStuff* [15], both running on the Matlab/Octave language. The motivation for this paper is that current ANN and GPR formulations cannot handle two particular situations that are common during site characterization: (1) experimental conditions where contaminant concentration is quantified from aggregated soil samples and (2) the effect of physical site discontinuities. Note that even if geostatistics methods can handle aggregated soil samples using *Block Kriging* [16], it cannot handle the effect of physical site discontinuities.

This paper proposes a new unified formulation based on the GPR method to address the two limitations identified above. The paper is organized as follows: Section 2 introduces the standard mathematical formulation of Gaussian Process Regression along with specificities associated with soil characterization applications. Section 3 presents the two extensions to the standard GPR formulation that are proposed in this paper. The first extension account aggregated soil samples by creating virtual points that are employed to model the average contaminant concentration. The second extension proposes a new covariance function that can employ discrete attributes corresponding to physical site discontinuities. The justification for these two new probabilistic formulations comes from a case study where both features are present. In section 4, an empirical analysis compares the

performance of these new extensions with the baseline GPR model.

2. Gaussian Process Regression for Contamination Concentration Characterization

This Section summarizes the theory behind Gaussian Process Regression [11, 13]. Subsection 2.1 presents aspects related to the model definition, Subsection 2.2 presents the formulation for estimating the conditional probability of a Gaussian process given observations, and Subsection 2.3 presents the procedure for calibrating hyper-parameters. All subsections are presented in the context of soil contamination concentration characterization.

2.1. Model definition

The characterization of contaminants concentration is based on the following fundamental equation

$$\underbrace{Y_i}_{\text{observation}} = \overbrace{\mathsf{c}(\mathbf{l}_i^{\{Y\}})}^{\text{true contaminant []}} + \underbrace{V_i}_{\text{measurement error}}, \; V_i \sim \mathcal{N}(0, \sigma_V^2) \tag{1}$$

where Y_i is a noise-contaminated observation of the contaminant concentration $\mathsf{c}(\mathbf{l}_i^{\{Y\}})$, and where V_i is a zero-mean Gaussian measurement error such that $V_i \perp\!\!\!\perp V_j, \forall i \neq j$. $\mathsf{c}(\mathbf{l}_i^{\{Y\}})$ describes an unknown, yet deterministic function corresponding to the concentration of contaminants across the tridimensional space. For a location i, $\mathbf{l}_i^{\{Y\}} = [x, y, z]_i^\intercal$ describes spatial coordinates. The model formulation in Equation 1 is defined for any real number; in practice, it is inconsistent with reality, because contaminant concentrations are strictly positive numbers. Therefore, it is common to transform the observations in the logarithmic space [17, 18],

$$\log \underbrace{Y_i}_{\text{observation}} = \overbrace{\mathsf{c}(\mathbf{l}_i^{\{Y\}})}^{\text{true contaminant [] in log space}} + \underbrace{V_i}_{\text{measurement error in log space}}, \; V_i \sim \mathcal{N}(0, \sigma_V^2) \tag{2}$$

This paper only employs the model formulation in the logarithmic space as described in Equation 2. The true contaminant concentration $\mathsf{c}(\mathbf{l}_i^{\{Y\}})$, is hidden so only realizations of the random variable Y_i can be observed. The set of observation $\mathcal{D} = \{(\mathbf{l}_i^{\{Y\}}, \mathbf{y}_i), \; i = 1 : M\}$ corresponds to M pairs of concentration observations and their associated covariate $\mathbf{l}_i^{\{Y\}}$ for which the superscript $^{\{Y\}}$ refers to observation locations.

2.2. Model Estimation

Although the true contaminant concentration $\mathsf{c}(\mathbf{l}_i^{\{Y\}})$ is a deterministic function, our knowledge of it is incomplete and it is thus described by a stochastic process quantifying the probability of contaminant concentration across space, $p(c|\mathbf{l}^{\{C\}})$. The probabilistic estimation of contaminants concentration $\mathbf{C}$ conditional on data $\mathcal{D}$ and estimation location $\mathbf{l}^{\{C\}}$ is denoted $p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D})$. This conditional probability is modelled using a Gaussian Process $p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D}) = \mathcal{N}(\mathbf{M}_{\mathbf{C}|\mathcal{D}}, \mathbf{\Sigma}_{\mathbf{C}|\mathcal{D}})$, where $\mathbf{l}^{\{C\}} = [\mathbf{l}_1^{\{C\}}, \mathbf{l}_2^{\{C\}}, \cdots, \mathbf{l}_N^{\{C\}}]^\intercal$ is a vector containing the coordinates for N tridimensional locations where the concentration needs to be estimated, and where the superscript $^{\{C\}}$ refers to estimation locations. The dependence on the vector of locations $\mathbf{l}^{\{C\}}$ of the posterior mean vector $\mathbf{M}_{\mathbf{C}|\mathcal{D}}$ and the posterior covariance matrix $\mathbf{\Sigma}_{\mathbf{C}|\mathcal{D}}$ are assumed implicitly to simplify the notation. The analytical formulation for computing $\mathbf{M}_{\mathbf{C}|\mathcal{D}}$ and $\mathbf{\Sigma}_{\mathbf{C}|\mathcal{D}}$ is obtained from the Gaussian conditional distribution

$$\underbrace{\begin{aligned}\mathbf{M}_{\mathbf{C}|\mathcal{D}} &= \mathbf{M}_{\mathbf{C}} + \mathbf{\Sigma}_{\mathbf{YC}}^\intercal \mathbf{\Sigma}_{\mathbf{YY}}^{-1}(\mathbf{y} - \mathbf{M}_{\mathbf{Y}}) \\ \mathbf{\Sigma}_{\mathbf{C}|\mathcal{D}} &= \mathbf{\Sigma}_{\mathbf{CC}} - \mathbf{\Sigma}_{\mathbf{YC}}^\intercal \mathbf{\Sigma}_{\mathbf{YY}}^{-1} \mathbf{\Sigma}_{\mathbf{YC}}\end{aligned}}_{\text{Posterior knowledge}} \tag{3}$$

In Equations 3, the subscript $_C$ and $_Y$ respectively refers to estimation and observation locations and matrices on the right-hand side correspond to the prior knowledge

$$\underbrace{\mathbf{M} = \left\{ \begin{array}{c} \mathbf{M_Y} \\ \mathbf{M_C} \end{array} \right\}, \; \mathbf{\Sigma} = \left[\begin{array}{cc} \mathbf{\Sigma_{YY}} & \mathbf{\Sigma_{YC}} \\ \mathbf{\Sigma_{YC}^\intercal} & \mathbf{\Sigma_{CC}} \end{array} \right]}_{\text{Prior knowledge}} \tag{4}$$

The prior knowledge for the mean vector is typically defined following the hypothesis that the prior mean is zero, i.e. $\mathbf{M} = \mathbf{0}$. If additional knowledge is available to describe the prior mean, more complex functions can be employed instead of $\mathbf{M} = \mathbf{0}$. The prior knowledge for each sub-component of the covariance matrix $\mathbf{\Sigma}$ is defined

$$\begin{aligned}[\mathbf{\Sigma_{YY}}]_{ij} &= \rho(\mathbf{l}_i^{\{Y\}}, \mathbf{l}_j^{\{Y\}})\sigma_C^2 + {\sigma_V}^2\delta_{ij}, \; \delta_{ij} = 1 \text{ if } i = j, \text{ else } \delta_{ij} = 0 \\ [\mathbf{\Sigma_{CC}}]_{kl} &= \rho(\mathbf{l}_k^{\{C\}}, \mathbf{l}_l^{\{C\}})\sigma_C^2 \\ [\mathbf{\Sigma_{YC}}]_{ik} &= \rho(\mathbf{l}_i^{\{Y\}}, \mathbf{l}_k^{\{C\}})\sigma_C^2.\end{aligned} \tag{5}$$

In Equation 5, subscripts $i, j = 1, 2, \cdots, M$ and $k, l = 1, 2, \cdots, N$, where M is the number of observations and N is the number of estimation locations. In

this definition of the covariance matrices, σ_C is the prior standard deviation of the concentration C and this one is considered to be constant for all locations $\mathbf{l}_i$. $\rho(\mathbf{l}_i, \mathbf{l}_j)$ is a *correlation function* which describes the correlation between the contaminant concentration $C(\mathbf{l}_i)$ and $C(\mathbf{l}_j)$ at two locations $\mathbf{l}_i$ and $\mathbf{l}_j$. One possible choice for the correlation function is the square exponential basis function defined by

$$\rho(\mathbf{l}_i, \mathbf{l}_j) = \exp\left(-\frac{1}{2}(\mathbf{l}_i - \mathbf{l}_j)^\intercal \mathrm{diag}(\boldsymbol{\ell}^2)^{-1}(\mathbf{l}_i - \mathbf{l}_j)\right) \tag{6}$$

where $\boldsymbol{\ell} = [\ell_x, \ell_y, \ell_z]^\intercal$ is a vector containing the length scale parameter for each spatial dimension. Each length scale parameter defines how correlation decays according to an increase in distance with respect to its corresponding direction. Figure 1 presents examples of unidimensional square-exponential covariance functions for different length-scale parameters where the correlation $\rho(x_i, x_j)$ is expressed as a function of the spatial distance $x_i - x_j$. Although many other correlation functions are available [13], only the square

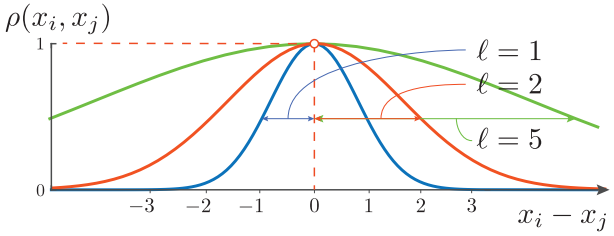


Figure 1: Examples of unidimensional square-exponential covariance functions for different length-scale parameters.

exponential is employed in this paper.

Note that although the formulation in Equation 3 is analytically accurate, it is known to be computationally demanding and to suffer from numerical instability issues. An equivalent formulation that is faster and numerically more stable is obtained by taking advantage of the Cholesky decomposition of $\mathbf{\Sigma_{YY}}$. This formulation is described in detail by Rasmussen and Williams [13] and is implemented in the packages *GPML* [14] and *GPStuff* [15].

2.3. Hyper-Parameter Estimation

The prior distribution of the Gaussian Process presented in §2.1 has five unknown hyper-parameters $\mathcal{P} = \{\ell_x, \ell_y, \ell_z, \sigma_C, \sigma_V\}$. The term *Hyper-parameters* is employed to describe parameters of the prior knowledge. The

posterior distribution of hyper-parameters conditioned on observations can be obtained using Bayes's theorem

$$\overbrace{p(\mathcal{P}|\underbrace{\mathbf{l}^{\{Y\}},\mathbf{y}}_{\text{Observations}})}^{\text{Posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{l}^{\{Y\}},\mathcal{P})}^{\text{Likelihood}}\overbrace{p(\mathcal{P})}^{\text{Prior}}}{\underbrace{p(\mathbf{y})}_{\text{Normalization constant}}}.$$

In practical situations this posterior distribution is difficult to estimate, in part because of the computational demand related to the normalization constant $p(\mathbf{y})$. Instead, a common approximation consists in employing a *Maximum Likelihood Estimate*, (MLE) where $\mathcal{P}^*$ designates the set of hyper-parameters that maximizes the likelihood function

$$\mathcal{P}^* = \arg\max_{\mathcal{P}} \overbrace{p(\mathbf{y}|\mathbf{l}^{\{Y\}},\mathcal{P})}^{\text{likelihood}} \equiv \arg\max_{\mathcal{P}} \overbrace{\log p(\mathbf{y}|\mathbf{l}^{\{Y\}},\mathcal{P})}^{\text{log-likelihood}}.$$

Because the log function is monotonically increasing, the maximum of the likelihood function corresponds to the maximum of the logarithm of the likelihood function. This transformation in the log space is employed because it improves the numerical stability of calculations [11]. The physical interpretation of the MLE technique is to find hyper-parameter values that maximize the prior probability of observations.

One of the convenient features of the Gaussian process is that the log-likelihood has a computationally efficient analytical formulation

$$\log p(\mathbf{y}|\mathbf{l}^{\{Y\}},\mathcal{P}) = \log \mathcal{N}(\mathbf{0},\mathbf{\Sigma_{YY}}) = -\frac{1}{2}\mathbf{y}^\intercal\mathbf{\Sigma_{YY}^{-1}}\mathbf{y} - \frac{1}{2}\log|\mathbf{\Sigma_{YY}}| - \frac{M}{2}\log 2\pi.$$

The MLE hyper-parameters are found by maximizing the log-likelihood. The maximum corresponds to the values for which the log-likelihood derivative is zero

$$\frac{d}{d\mathcal{P}_j}\log p(\mathbf{y}|\mathbf{l}^{\{Y\}},\mathcal{P}) = \frac{1}{2}\mathbf{y}^\intercal\mathbf{\Sigma_{YY}^{-1}}\frac{d\mathbf{\Sigma_{YY}}}{d\mathcal{P}_j}\mathbf{\Sigma_{YY}^{-1}}\mathbf{y} - \frac{1}{2}\mathrm{tr}(\mathbf{\Sigma_{YY}^{-1}}\frac{d\mathbf{\Sigma_{YY}}}{d\mathcal{P}_j}) = 0.$$

General purpose gradient-based MLE algorithms are implemented in the packages *GPML* [14] and *GPStuff* [15].

3. New extensions to Gaussian Process Regression

This section describes two new extensions to GPR. The first extension presented in Subsection 3.1 allows modelling observations that are obtained from aggregated soil samples. The second extension is presented in Subsection 3.2 and allows modelling site featuring discrete physical discontinuities.

3.1. Gaussian Process Regression Using Aggregated Samples

It is common to perform laboratory analyses for contamination concentration on soil samples obtained from borehole drillings. In order to obtain a sufficient volume of soil for the purpose of analysis, the soil core extracted over a vertical length from 0.5 m to 1 m is aggregated as a single sample. When the vertical length-scale parameter is smaller than the vertical length of an aggregated sample, it is an indication that different locations comprised within the same sample have a correlation coefficient close to zero. In that case, one should expect a high heterogeneity within the sample itself. For example, one end of a sample could be contaminated while the other end is not.

The standard Gaussian Process Regression method cannot handle aggregated samples. We propose to extend GPR by employing the formulation employed in state-space models [11, 19] to provide a formulation that explicitly consider the effect of aggregated samples. Note that although the formulation employed already exists in the field of state-space models, it has not been applied to GPR. The resulting formulation employs an observation matrix $\mathbf{F}$ that allows considering a concentration observation y_i as an average value over a set of virtual covariate locations $\mathbf{l}^{\{Y*\}}$ that are discretized over the vertical soil sample length. The observation matrix $\mathbf{F}$ is employed to compute the covariance matrix, $\mathbf{\Sigma_{YC}}$ and $\mathbf{\Sigma_{YY}}$, so that

$$\mathbf{\Sigma_{YC}} = \mathbf{F}\mathbf{\Sigma_{Y*C}}$$

and

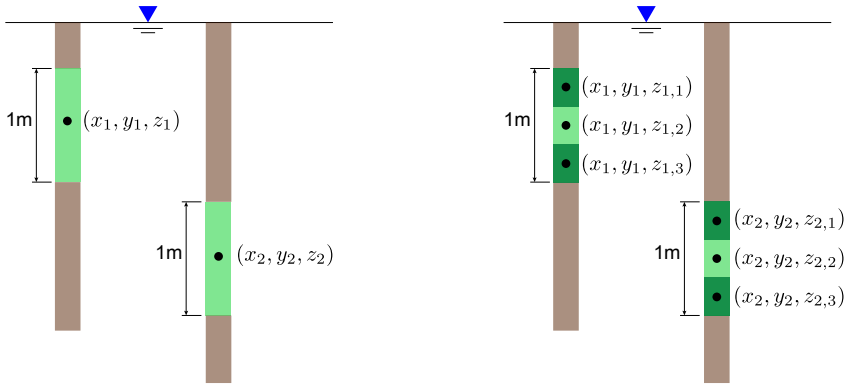
$$\mathbf{\Sigma_{YY}} = \mathbf{F}\mathbf{\Sigma_{Y*Y*}}\mathbf{F}^{\intercal} + \mathbf{R}$$

where $\mathbf{\Sigma_{Y*Y*}}$ and $\mathbf{\Sigma_{Y*C}}$ are the covariance matrices computed using virtual covariate locations $\mathbf{l}^{\{Y*\}}$. Note that the matrix $\mathbf{\Sigma_{Y*Y*}} = \rho(\mathbf{l}_i^{\{Y\}}, \mathbf{l}_j^{\{Y\}})\sigma_C^2$ does not include measurements noise which comes from the diagonal matrix $\mathbf{R}$

$$[\mathbf{R}]_{jj} = \sigma_V^2, \ \forall j = 1 : M$$

This special treatment of the measurement noise is necessary because measurement errors applies to the aggregated sample and not on the individual virtual locations.

Figure 2 presents a comparative example of the discretization procedure for two boreholes (i.e. $M = 2$), each employed to represent a 1 m long soil sample. In Figure 2a, the basic approach is employed and no discretization is performed. Observed concentrations are assigned to the soil sample centroid.



(a) Basic approach without sample discretization

(b) Refined approach with sample discretization

Figure 2: Comparative example of the discretization procedure for two boreholes, each employed to obtain a 1m-long soil sample where the blue triangle represents the soil surface.

This corresponds to an observation matrix

$$\mathbf{F} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

where in this special case, $\mathbf{\Sigma_{YC}} = \mathbf{\Sigma_{Y*C}}$ and $\mathbf{\Sigma_{YY}} = \mathbf{\Sigma_{Y*Y*}} + \mathbf{R}$. In Figure 2b, each soil sample is discretized vertically into three virtual subsamples. The observation matrix corresponding to this case is

$$\mathbf{F} = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}$$

where each line corresponds to an observation and each column corresponds to a virtual subsample. Here, the weight of each virtual subsample is $\frac{1}{3}$ because there are three discretization points. In this illustrative example, the number of discretization subsamples is arbitrarily set to three. In real situations, the accuracy of the approach increases asymptotically with the number of discretization subsamples.

3.2. Gaussian Process Regression With Hidden Covariates

The standard GPR formulation presented in Section 2 is unable to capture the effect of physical discontinuities in a site. We propose a new method to enhance the standard GPR formulation in order to overcome this limitation. Physical discontinuities may arise from situations such as when site filling is performed using different materials, or due to the presence of physical

barriers that prevent or hinders the migration of contaminants across space. Figure 3 presents an example of site that is separated in two regions by a partially permeable physical barrier. In this example we assume that the

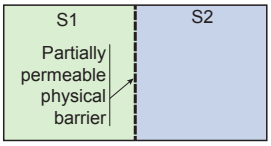


Figure 3: Example of site separated in two regions by a partially permeable physical barrier.

spatial correlation between two locations either within region $S1$ or $S2$ is a function of the spatial distance between the two locations as described in Equation 6. The correlation between any point in the region $S1$ and any point $S2$ is assumed to be described by

$$\rho(\{\mathbf{l}_i, S1\}, \{\mathbf{l}_j, S2\}) = \rho(\mathbf{l}_i, \mathbf{l}_j) \cdot \rho_{S1,S2}$$

where $\rho_{S1,S2} \in (0, 1)$ is the correlation coefficient between the region $S1$ and $S2$. For the special case where $\rho_{S1,S2} = 0$, the two regions are completely independent, and for the other special case where $\rho_{S1,S2} = 1$, both regions are linearly correlated, $\rho(\{\mathbf{l}_i, S1\}, \{\mathbf{l}_j, S2\}) \equiv \rho(\mathbf{l}_i, \mathbf{l}_j)$. In order to provide a correlation function that is applicable to the general case where there can be Q distinct regions, we propose to model the correlation as a function of *virtual distances* d_{rs}, $\forall\{r = 1 : Q, s = r : Q\}$ between each pair of regions defined by indexes r and s. The new correlation function defined for spatial distances (i.e. $\mathbf{l}_i - \mathbf{l}_j$) and virtual distances (i.e. d_{rs}) is

$$\rho(\{\mathbf{l}_i, r\}, \{\mathbf{l}_j, s\}) = \exp\left(-\frac{1}{2}\left((\mathbf{l}_i - \mathbf{l}_j)^\intercal \text{diag}(\boldsymbol{\ell}^2)^{-1}(\mathbf{l}_i - \mathbf{l}_j) + d_{rs}^2\right)\right), \quad (7)$$

where the correlation coefficient between pairs of regions is given by

$$\rho_{rs} = \exp\left(-\frac{1}{2}d_{rs}^2\right).$$

Each virtual distance d_{rs}, $\forall r \neq s$ is now a hyper-parameter. Given that there are Q distinct regions and considering that $d_{rr} \equiv 0$, there are $Q^2/2 - Q$ additional hyper-parameters to be inferred using the MLE formulation presented in Subsection 2.3.

In order to be admissible, covariance functions $[\mathbf{\Sigma}]_{ij}$ described in Equation 5 must be *positive semi-definite*. This criterion is met for any set of virtual distance d_{rs}, for which it is possible to find the coordinates $\mathbf{w}_r = [w_{r,1}, w_{r,2}, \cdots, w_{r,Q-1}]^\intercal$, $\forall\{r = 1 : Q\}$ defining a *simplex* in the *real coordinate space*, $\mathbb{R}^{Q-1}$. A *simplex* is the generalization of a triangle in any number of dimensions [20]. Figure 4 presents examples of simplex in one, two and three dimensions, where the length of edges corresponds to virtual distances d_{rs} and vertices are defined by coordinates $\mathbf{w}_r$. From examples in

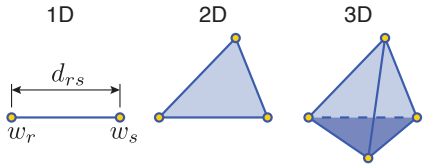


Figure 4: Examples of simplex in one, two and three dimensions, where the length of edges corresponds to virtual distances d_{rs} and vertices are defined by coordinates $\mathbf{w}$.

Figure 4, for $Q = 2$ distinct regions, one has to infer a single virtual distance d_{12} describing a 1D simplex with vertices coordinate $[w_1, w_2]$. For $Q = 3$ distinct regions, one has to infer three virtual distances $\{d_{12}, d_{13}, d_{23}\}$, describing a 2D simplex with vertex coordinates $\{[w_{11}, w_{12}], [w_{21}, w_{22}], [w_{31}, w_{32}]\}$. If such a simplex exists in the real coordinate space, the correlation function employing virtual distances d_{rs} in Equation 7 is equivalent to the correlation function in Equation 6 where virtual distances d_{rs} would be computed from coordinates $\mathbf{w}_r$.

The backward calculation of vertex coordinates $\mathbf{w}_r$ from virtual distances d_{rs} can be done by following the recursive method presented by Erlandson [21]. In the case where a simplex is defined in the complex coordinate space, the resulting covariance matrices are not admissible. Thus, during each optimization step, before accepting a virtual distance change $d_{rs} \rightarrow d_{rs}+\Delta_{d_{rs}}$, it is necessary to check if the simplex defined by the virtual distances $d_{rs}+\Delta_{d_{rs}}$ lies in the real coordinate space. If this criterion is met, the optimization continues, otherwise, the log-likelihood corresponding to $d_{rs} + \Delta_{d_{rs}}$ is set to be equal to the one obtained for d_{rs}. This constrains results in rejecting the parameter change because the gradient-based optimization algorithm described in Section 2.3 only accept a parameter change if it leads to a log-likelihood increase. This procedure ensures that the optimization space is contained to virtual distances d_{rs} leading to positive semi-definite covariance matrices.

4. Empirical Validation

This section compares the performance of the standard GPR model with the two improvements proposed. Subsection 4.1 presents the site and dataset employed for performance comparison purpose, Subsection 4.2 presents the method employed for quantifying the predictive capacity of each model configuration, and Subsection 4.3 presents the comparison procedure along with results. All analyses presented in this section have been performed using the GPML package [14] in which we implemented the new formulations presented in section 3.

4.1. Site and Dataset Description

The site studied has been anonymized in order to satisfy confidentiality agreements. Its size is approximately 475 meters long by 175 meters wide by 9 meters deep. The site historically consisted in several water channels delimited by masonry walls and that were employed for industrial purposes. Around 1950 water channels were filled and the industrial purpose of the site remained until the end of the century.

The site is heterogeneous and across the different layers of boreholes, traces of debris such as bricks, coal, wood and coal cinders have been found. Also, under the canopy or asphalt, there are different layers which are composed either of sand, gravel, silt, clay or sandy silt. Even though the overburden above the bedrock have a thickness of about 18 meters, our model goes up to 9 meters deep as it represents the drilling maximum depth. In this study, the water flow in the site and the contaminant diffusion are neglected. The site is assumed to be in a stationary state over the duration of the site characterization and rehabilitation. The absence of contamination of the water table justifies the absence of migration by this vector. The location of land (S) and former water channel (B) regions are represented in Figure 5.

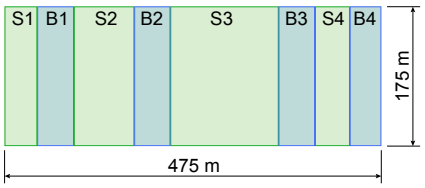


Figure 5: Simplified representation of the site studied. Labels S1-S4 represent land regions and labels B1-B4 represent locations that are former water channels that were filled around 1950.

A total of 116 soil samples have been analyzed for the purpose of characterizing lead contamination concentration. Among several contaminants, such as petroleum hydrocarbons, chrome, copper and zinc, whose level exceeded the thresholds set by the government at various locations, lead contamination was chosen for our study. The detection limit of laboratory analyses is 5 mg/kg. Observations below that threshold are discarded. Each concentration observation is associated with its longitude and latitude coordinates along with a vertical depth interval corresponding to the boundary of the aggregated soil sample analyzed in laboratory.

4.2. Cross-Validation Procedure

The goal of the comparison is to identify the model configuration that has the best capacity at predicting the contaminant concentration. The log-likelihood is employed as criteria to optimize hyper-parameter values. However, the likelihood is not an adequate criterion to compare different model configuration because it is not immune to over-fitting [11, 13]. A robust criterion for comparing the predictive capacity of model configuration is the *Leave One Out Cross-Validation* (LOO-CV) [22]. This procedure consists in predicting iteratively each observation i without including it in the observation set $\mathcal{D}* = \mathcal{D} \setminus \mathcal{D}_i$ for the purpose of estimating $p(c_i|\mathbf{l}^{\{C\}}, \mathcal{D}*)$. Assuming that all observations are independent, the joint probability density of all predictions is obtained as the product of marginal probabilities $p(c_i|\mathbf{l}^{\{C\}}, \mathcal{D}*)$. An equivalent, yet more robust way of performing this calculation is to sum the logarithm of marginal probabilities

$$\log(p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D})) = \sum_{i=1}^{M} \log p(C_i|\mathbf{l}^{\{C\}}, \mathcal{D}*).$$

The model configuration which has the highest $\log(p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D}))$ has the highest capacity at predicting unobserved data. Therefore, if the two methods proposed in this paper improve the predictive capacity of the model, they must return LOO-CV values that are higher than the standard GPR model.

4.3. Comparative study

4.3.1. Compared scenarios

In this comparative study, 4 sets of regions are considered along with five maximal lengths for soil sample discretizations. Table 1 presents the 4 sets of regions employed.

Regarding soil maximal sample length U, five values are considered: $U = \{\infty, 0.75, 0.5, 0.25, 0.125\}\, m$. For any maximal sample length, there

Table 1: Sets of subregions studied.

Q	Subregions	Description
1	$\{B1, B2, B2, B4, S1, S2, S3, S4\}$	All subregions
2	$\{B1, B2, B2, B4\}$, $\{S1, S2, S3, S4\}$	Separation of land and basins
5	$\{S1, S2, S3, S4\}, \{B1\}, \{B2\}, \{B3\}, \{B4\}$	Land together and basins separated
8	$\{S1\}, \{S2\}, \{S3\}, \{S4\}, \{B1\}, \{B2\}, \{B3\}, \{B4\}$	Everything separated

cannot be two points within a soil sample that are separated by a distance larger than U without adding a virtual covariate location $\mathbf{l}^{\{Y*\}}$. Virtual covariate locations are added so that their spread is maximized are uniformly spaced, and are centred on soil subsamples. For the special case where $U = \infty$, there is only one covariate location situated at the centroid of the soil sample. Given that there are four sets of regions and five maximal sample lengths to be tested, a total of 20 scenarios are compared. Note that the special case where $Q = 1$ and $U = \infty$ corresponds to the *standard GPR model*.

Comparisons involving the LOO-CV procedure only employ concentration data that are above the detection limit of $5mg/kg$. Concentration data that are below the detection limit consists in upper-bound observations [23], which are not fully consistent with the measurement error structure employed in this paper.

4.3.2. Parameters Calibration

The GPR hyper-parameters initial values have been selected as follows : The length scale parameters are $\boldsymbol{\ell_x} = \boldsymbol{\ell_y} = 50\,m$ and $\boldsymbol{\ell_z} = 1\,m$. The Gaussian process prior standard deviation initial value is $\sigma_f = 3$, the measurement error prior standard deviation (in the log-space) is assumed to be $\sigma_V = 0.05$ and the initial values for virtual distances $d_{rs} = 0.45$ so that $\rho_{rs} = 0.9\ \forall r \neq s$.

All hyper-parameters are calibrated using the entire dataset. The final hyper-parameter values obtained are only presented for two scenarios: (1) the standard GPR model, $U = \infty$ & $Q = 1$, and (2) the model including both extensions where $U = 0.125\,m$ & $Q = 8$. For the first scenario, length scale parameters are $\boldsymbol{\ell_x} = 51.0\,m$, $\boldsymbol{\ell_y} = 53.9\,m$, $\boldsymbol{\ell_z} = 0.73\,m$ and the standard deviations are $\sigma_f = 2.8,\ \sigma_v = 0.15$. For the second scenario, the length scale parameters are $\boldsymbol{\ell_x} = 56.4\,m$, $\boldsymbol{\ell_y} = 53.7\,m$, $\boldsymbol{\ell_z} = 0.64\,m$ and the standard deviations are $\sigma_f = 2.9,\ \sigma_v = 0.18$. In addition to these parameters Table 2 presents the correlation matrix that is computed from the estimated virtual distances for $Q = 8, U = 0.125$m. The correlation between the water channels

Table 2: MLE estimates for correlation coefficient between regions estimated for $Q = 8$, $U = 0.125$m

Regions	S1	B1	S2	B2	S3	B3	S4	B4
S1	1	0.90	0.92	0.91	0.91	0.90	0.90	0.90
B1		1	0.87	0.92	0.90	0.90	0.90	0.90
S2			1	0.86	0.95	0.90	0.90	0.90
B2				1	0.84	0.90	0.90	0.90
S3					1	0.84	0.90	0.90
B3			sym.			1	95	0.90
S4							1	0.93
B4								1

is similar to the one between the land regions. Correlation coefficients for regions that are far apart from each other remain equal to the prior value. This effect occurs because the influence of a hyper-parameter d_{ij} on the log-likelihood is negligible for regions that are far apart from each other.

4.3.3. Results

The log-likelihood computed using the LOO-CV for each scenario is presented in Table 3. Results indicate that for a maximal sample length U,

Table 3: The results correspond to the log-likelihood $\log(p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D}))$ computed using the LOO-CV procedure.

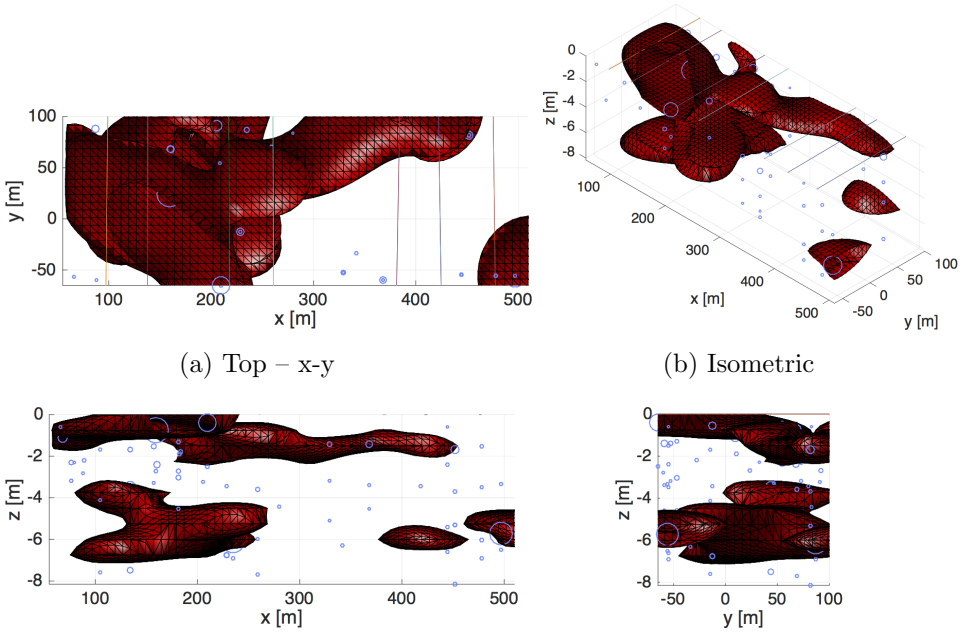
Number of subregions, Q	Max. sample length, U [m] ∞	0.75	0.5	0.25	0.125
1	-173.0	-173.3	-173.4	-173.5	-173.6
2	-175.8	-172.3	-171.7	-171.8	-171.8
5	-177.6	-172.6	-172.1	-171.7	-172.3
8	-175.9	-171.5	-171.0	-171.3	-170.9

increasing the number of sub-regions Q tends to increase the log-likelihood $\log(p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D}))$. When using one subregion, i.e. $Q = 1$, decreasing the maximal sample length U increases $\log(p(\mathbf{c}|\mathbf{l}^{\{C\}}, \mathcal{D}))$. For $Q = 8$ subregions, best results are achieved for the smallest maximal sample length, i.e. $U = 0.125\,m$.

One possible cause explaining that results in Table 3 are not monotonically increasing with Q and decreasing with U, is the usage of a MLE approach

this estimation method does not consider the uncertainty in the estimation of hyper-parameters. Nonetheless, LOO-CV results indicate that models accounting for the effect of physical site discontinuities and aggregated soil samples have a superior predictive capacity than the baseline GPR model.

Figures 6 and 7 compare the tridimensional iso-surfaces $\Pr(C > 55\,mg/kg) > 0.5$ for scenarios with respectively $\{Q = 1, U = \infty\}$ and $\{Q = 8, U = 0.125\,m\}$. Lines in the x-y plane represent the physical discontinuities



(a) Top – x-y

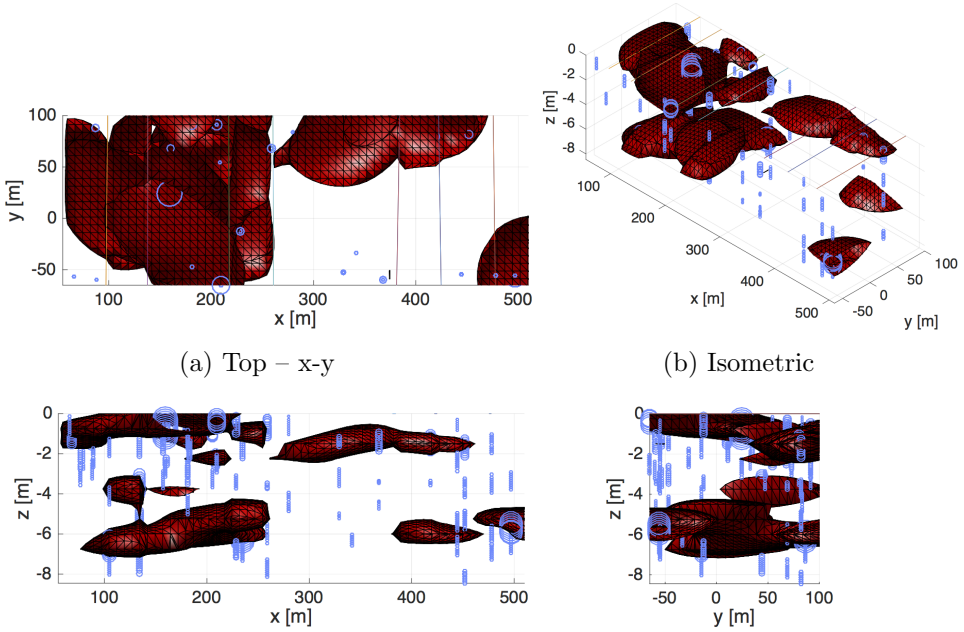
(b) Isometric

(c) Elevation – x-z

(d) Elevation – y-z

Figure 6: Tridimensional iso-surfaces $\Pr(C > 55\,mg/kg) > 0.5$ for the scenario with $Q = 1, U = \infty\,m$. Lines in the x-y plane represent the physical discontinuities between each subregion and dots represent sample locations where the size is proportional to the observed contaminant concentration.

between each subregion and dots correspond to virtual sample locations where their size is proportional to the observed contaminant concentrations. Contrarily to iso-contours in Figure 6, Figure 7 (which include the effect of physical discontinuities) displays sharp transitions of the contaminant concentration between sub-regions. In both cases iso-contours indicate that the contaminant concentration is clustered in layers that have approximately the same thickness as the soil sample length, i.e. $\approx 1\,m$. This confirms the relevance to consider the effect of aggregated soil samples.



(a) Top – x-y

(b) Isometric

(c) Elevation – x-z

(d) Elevation – y-z

Figure 7: Tridimensional iso-surfaces $\Pr(C > 55\, mg/kg) > 0.5$ for the scenario with $Q = 8$, $U = 0.125\, m$. Lines in the x-y plane represent the physical discontinuities between each subregion and dots represent virtual sample locations where the size is proportional to the observed contaminant concentration.

5. Discussion

The results obtained in Section 4.3 demonstrates that two new extensions to the GPR outperform the standard GPR in the presence of aggregated soil samples and site discontinuities. This conclusion can be reached based on the LOO-CV results presented in Table 3. In Figures 6 and 7, the reader can appreciate the effects of the formulation proposed on results. Nevertheless, these contaminant iso-surfaces cannot be employed to compare the performance of model formulations since the true contaminant iso-surfaces remain unknown. Therefore, the most robust comparison metric remains the LOO-CV results. Note that the comparison with other techniques such as ANN remains to be confirmed quantitatively.

The results have shown that the current method is able to treat practical cases where there are site discontinuities in sites as well as aggregated samples. The main limitation of the current models is that parameters are estimated using Maximum Likelihood Estimation (MLE). Future work should quantify the epistemic uncertainty in model parameters $\mathcal{P}$, which is resulting from the usage of a small dataset. This task could be achieved by using a Bayesian approach instead of MLE approach for estimating model parameters.

6. Conclusion

Two improvements to the standard Gaussian Process Regression are proposed in this paper. The first one can take into account the observation method employing aggregated samples. The second method allows considering physical discontinuities between sub-regions within a site. The two new probabilistic formulations proposed outperformed the standard GPR model. Although the gain in the prediction performance is small, the fundamental hypotheses employed to model site discontinuities and soil sampling techniques are in better agreement with experimental conditions, as demonstrated by LOO-CV results.

Acknowledgements

The authors would like to thank Yvon Courchesne from WSP group for his help in the project and from Denis Marcotte for his comment on the manuscript preliminary version. The project was funded by the Fonds de recherche du Quebec - Nature et technologies (FRQNT, Project #2017-NC-197235).

References

[1] J.-P. Boudreault, J.-S. Dubé, D. Marcotte, Quantification and minimization of uncertainty by geostatistical simulations during the characterization of contaminated sites: 3-d approach to a multi-element contamination, Geoderma 264 (2016) 214–226.

[2] P. Goovaerts, Geostatistics in soil science: state-of-the-art and perspectives, Geoderma 89 (1) (1999) 1–45.

[3] P. Goovaerts, Geostatistical modelling of uncertainty in soil science, Geoderma 103 (1) (2001) 3–26.

[4] G. Matheron, Le krigeage universel, Tech. rep., Ecole Nationale Superieure des Mines de Paris (1969).

[5] M. Kanevski, A. Pozdnoukhov, V. Timonin, Machine learning for spatial environmental data, Theory, applications and software (2009) 377.

[6] M. F. Kanevski, Spatial predictions of soil contamination using general regression neural networks, SYSTEMS RESEARCH AND INFORMATION SCIENCE 8 (1999) 241–256.

[7] G. Corani, Air quality prediction in milan: feed-forward neural networks, pruned neural networks and lazy learning, Ecological Modelling 185 (2) (2005) 513–529.

[8] G. Asadollahfardi, M. Madinejad, S. H. Aria, V. Motamadi, Predicting particulate matter (pm2. 5) concentrations in the air of shahr-e ray city, iran, by using an artificial neural network, Environmental Quality Management 25 (4) (2016) 71–83.

[9] M. Zickus, A. Greig, M. Niranjan, Comparison of four machine learning methods for predicting pm10 concentrations in helsinki, finland, Water, Air and Soil Pollution: Focus 2 (5-6) (2002) 717–729.

[10] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016.

[11] K. Murphy, Machine learning: a probabilistic perspective, The MIT Press, 2012.

[12] D. MacKay, Introduction to gaussian processes, NATO ASI Series F Computer and Systems Sciences 168 (1998) 133–166.

[13] C. Rasmussen, C. Williams, Gaussian processes for machine learning, the MIT Press 2 (3) (2006) 4.

[14] C. Rasmussen, H. Nickisch, Gaussian processes for machine learning (gpml) toolbox, The Journal of Machine Learning Research 11 (2010) 3011–3015.

[15] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, A. Vehtari, Bayesian modeling with gaussian processes using the gpstuff toolbox, arXiv preprint arXiv:1206.5754.

[16] T. Burgess, R. Webster, Optimal interpolation and isarithmic mapping of soil properties, Journal of Soil Science 31 (2) (1980) 333–341.

[17] E. Pebesma, Multivariable geostatistics in s: the gstat package, Computers & Geosciences 30 (7) (2004) 683–691.

[18] N. Cressie, Block kriging for lognormal spatial processes, Mathematical Geology 38 (4) (2006) 413–443.

[19] M. West, J. Harrison, Bayesian Forecasting and Dynamic Models, Springer Series in Statistics, Springer New York, 1999.

[20] J. R. Munkres, Elements of algebraic topology, Vol. 2, Addison-Wesley Menlo Park, 1984.

[21] E. Erlandson. Computing simplex vertex locations from pairwise object distances [online] (March 2016) [cited October 24th 2016].

[22] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The elements of statistical learning: data mining, inference and prediction, The Mathematical Intelligencer 27 (2) (2005) 83–85.

[23] P. Gardoni, A. Der Kiureghian, K. Mosalam, Probabilistic capacity models and fragility estimates for reinforced concrete columns based on experimental observations, Journal of Engineering Mechanics 128 (10) (2002) 1024–1038.