Advanced Engineering Informatics 33 (2017) 60–67

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

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

ARTICLE INFO

Article history:
Received 4 November 2016
Received in revised form 3 May 2017
Accepted 5 May 2017
Available online 18 May 2017

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 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) [6–9]. 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 for-

* Corresponding author.
E-mail address: james.a.goulet@gmail.com (J.-A. Goulet).

http://dx.doi.org/10.1016/j.aei.2017.05.002
1474-0346/

mulations 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

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

where Y_i is a noise-contaminated observation of the contaminant concentration $\mathrm{c}(\mathbf{l}_i^{\{Y\}})$, and where V_i is a zero-mean Gaussian measurement error such that $V_i \perp\!\!\!\perp V_j, \forall i \neq j$. $\mathrm{c}(\mathbf{l}_i^{\{Y\}})$ 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 Eq. (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],

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

This paper only employs the model formulation in the logarithmic space as described in Eq. (2). The true contaminant concentration $\mathrm{c}(\mathbf{l}_i^{\{Y\}})$, 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 $\mathrm{c}(\mathbf{l}_i^{\{Y\}})$ 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 modeled 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 $\mathbf{l}^{\{C\}} = [\mathbf{l}_1^{\{C\}}, \mathbf{l}_2^{\{C\}}, \ldots, \mathbf{l}_N^{\{C\}}]^\intercal$ 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 Eqs. (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

$$\underbrace{\mathbf{M} = \begin{Bmatrix}\mathbf{M}_{\mathbf{Y}} \\ \mathbf{M}_{\mathbf{C}}\end{Bmatrix}, \mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_{\mathbf{YY}} & \mathbf{\Sigma}_{\mathbf{YC}} \\ \mathbf{\Sigma}_{\mathbf{YC}}^\intercal & \mathbf{\Sigma}_{\mathbf{CC}}\end{bmatrix}}_{\text{Prior knowledge}} \tag{4}$$

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

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

In Eq. (5), subscripts $i,j = 1,2,\ldots,M$ and $k,l = 1,2,\ldots,N$, 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. Fig. 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 exponential is employed in this paper.Note that although the formulation in Eq. (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 $\mathbf{\Sigma}_{\mathbf{YY}}$. This formulation is described in detail by Ras-

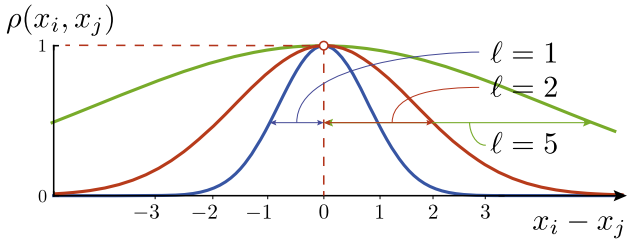

Fig. 1. Examples of unidimensional square-exponential covariance functions for different length-scale parameters.