Cost-Effective Strategy for the Investigation and Remediation of Polluted Soil Using Geostatistics and a Genetic Algorithm Approach

Yong-Qiang Cui*, Minoru Yoneda, Yoko Shimada, Yasuto Matsui

Department of Environmental Engineering, Graduate School of Engineering, Kyoto University, Kyoto, Japan
Email: *cui@risk.env.kyoto-u.ac.jp

Abstract

The geostatistical technique of Kriging has extensively been used for the investigation and delineation of soil heavy metal pollution. Kriging is rarely used in practical circumstances, however, because the parameter values are difficult to decide and relatively optimal locations for further sampling are difficult to find. In this study, we used large numbers of assumed actual polluted fields (AAPFs) randomly generated by unconditional simulation (US) to assess the adjusted total fee (ATF), an assessment standard developed for balancing the correct treatment rate (CTR) and total fee (TF), based on a traditional strategy of systematic (or uniform) grid sampling (SGS) and Kriging. We found that a strategy using both SGS and Kriging was more cost-effective than a strategy using only SGS. Next, we used a genetic algorithm (GA) approach to find optimal locations for the additional sampling. We found that the optimized locations for the additional sampling were at the joint districts of polluted areas and unpolluted areas, where abundant SGS data appeared near the threshold value. This strategy was less helpful, however, when the pollution of polluted fields showed no spatial correlation.

Keywords

**Cost-Effective, Soil Pollution, Heavy Metals, Investigation and Remediation, Kriging,
Genetic Algorithm**

---

*Corresponding author.

1. Introduction

Soil heavy metal pollution is becoming an increasingly severe global environmental problem, especially in developing countries such as China, due to the toxic, non-biodegradable, and persistent nature of heavy metals (HMs) [1]-[5]. According to China's latest national soil survey report, 16.1% of Chinese soil is polluted and 82.8% of the polluted land is contaminated by toxic inorganic pollutants and HMs [6]. Remediation of polluted soil by HMs improves not only the quality of soil to keep food and water safe for human health, but also the economic and commercial value of the land after remediation. Yet when we deal with a real-world polluted site, we invariably struggle to solve complex issues such as the number of samples to be taken, the actual site of sampling, the remediation method to use, and where the additional samples should be located in order to trade off the total fee (TF: fees for investigation, analysis, transportation, remediation, etc.) and correct treatment rate (CTR: the percentage of the area or volume of treated polluted soil and the total polluted soil before treatment). We therefore seek to develop a cost-effective strategy for trading off the TF and CTR in this study. Somebody must always take responsibility for obtaining representative samples before analysis. Practitioners of chemometrics often fail to heed that the error for representative sampling is typically far larger (10 - 100 times, even as much as 100 - 1000 times) than that incurred in pre-analysis treatment and analysis [7]-[9]. Sampling design involves a sampling approach and a sample density. Two principal sampling patterns are always recommended to collect metal data in soils: random sampling and systematic grid sampling (SGS) [10] [11]. Yet the former, random sampling, is inadequate and inefficient in reflecting site conditions and spatial differences in contamination levels [10] [12]. On one hand, the site soil is heterogeneous and the sampling method tends to produce clustering. [13] [14]. On the other hand, the large numbers of samples required for random sampling can incur high costs in time and money [10]. For this study, we therefore select the SGS approach, which gives an optimum between minimum variance and cost, as an optimal strategy [15] [16]. Sample density by SGS depends on the sampling area and budget. In the soil sampling guidelines used in Europe (ESSG), the number of samples ranges from 2/ha to 144/ha, depending on the country, and the sampling density can reach levels as high as $20/400 \text{ m}^2$ (Finland) and $16/100 \text{ m}^2$ (Switzerland) [17]. Even when sparsely distributed across sites, soil sampling data cannot provide all of the suitable information for the effective delineation of contaminated areas (information on the distribution of heavy metals at unsampled locations can be lacking, for example).

Kriging interpolation has been used to model and map spatial distributions of soil heavy metal contamination for more than three decades [18]-[20]. Kriging is more advanced than other interpolation methods such as the inverse distance weighted method, since Kriging considers two sets of distances (the distance between a location of interest and a sample location, and the distance between two sample locations) [21]. Yet Kriging-based delineation can be confounded by the variation of the Kriging estimation (e.g. the smoothing effect). This confounding effect can lead to both false positive and false negative delineation, which in turn can lead to financial risks and environmental risks due to over- or under-estimation at unsampled locations [19] [22]-[24]. The variogram, which is essential for Kriging, requires adequate sampling data to generate more precise estimation. Webster and Oliver have proposed that variograms require at least 100 data for computation, while those computed on fewer than 50 data are of little value [25]. When fewer than 50 or even 100 observations are used, the Kriging estimation will lack any clear advantage and lead to a large proportion of false delineations [26] [27]. The data, moreover, must be normally distributed to produce the best possible estimate [28]. The observed data sets from real polluted sites always diverge from the normal distribution (e.g. unduly high skewness and outliers), hence the standard geostatistical methods are inappropriate for datasets. A logarithmic or Box–Cox data transform is often applied before analysis in this situation, and a Q-Q normal plot is also used to test the normality of the transformed data [29]-[31].

The sampling at the beginning of a contaminated site investigation is mainly designed not for geostatistical interpolations, but for risk assessment based on legal regulations or intuition (e.g. the deviating color of soil) [32]. Additional sampling therefore becomes necessary at a later stage. A number of sampling approaches to improve the reliability of Kriging-based delineation have been proposed. An adaptive cluster sampling (ACS) approach, for example, was proposed for additional sampling to reduce Kriging errors and misclassification rates [19] [24]. The remediation cost uncertainty or expected total loss from misclassifications of the contamination status of the soil has also been used as performance criteria for additional or later-phase sampling [32] [33]. Many studies have focused on conditional simulation (CS) or sequential Gaussian simulation to generate probability maps for the assessment of soil pollution [18] [22] [23] [32]-[37]. Even though CS is not aimed at mini-

mizing local variance error, the simulation can better reproduce the spatial variability modeled by the same sample variogram [34]. Additional sampling locations are commonly suggested at sites where the conditional probability of exceedance is around 50%, while the number of additional samples taken depends on the decision-maker's stance on risk (a risk-prone investigator prefers taking fewer additional samples and a risk-averse decision-maker prefers taking more) [32] [33] [38]. However, these approaches or performance criteria do not lead to an optimal trade-off between TF and CTR simultaneously.

The commonly used validation methods such as cross-validation tend to waste some of the existing sampling data or require more independent samples to evaluate the effectiveness of the variogram model. Cross-validation only validates the prediction accuracy at sampling sites and cannot reflect the accuracy at unsampled sites. [39]. Contaminant concentrations are usually impossible to determine at every single location of a real-world site unless the surveyor is willing to pay the prohibitive costs in time and money required to take exhaustive samples. The costs would clearly impede a whole field-scale validation. A complete validation becomes feasible, however, using something like stochastic unconditional simulation (US), a technique capable of generating a perfectly known "synthetic" reality. This type of validation has been widely applied within different fields of application (e.g. nuclear waste, petroleum engineering, the environment) [22]. Geostatistical techniques have been repeatedly proven to be among the best available methods in the delineation of polluted soil, yet most of the world's soil remediation projects seem to avoid the use of geostatistics in the field. According to a survey in Belgium, for example, 84% of the 32 Dutch soil remediation firms rarely or never use geostatistics in practical circumstances [36]. This clearly underscores the need to report convincing results from research to demonstrate the advantages of integrating spatial statistics into the practice of soil remediation.

In this study we randomly generated a series of assumed actual polluted fields (AAPFs) by US and then used them to validate a cost-effective strategy based on the main two performance criteria: the CTR and TF. We then developed a cost-effective strategy using SGS and Kriging and determined relatively optimal locations for additional sampling through GA optimization.

2. Materials and Methods

2.1. Assumed Actual Polluted Fields (AAPFs) Generated by Unconditional Simulation (US)

Methods based on conditional simulation (CS) generate a series of realizations (equally probable solutions) of pollution distribution through original sampling data from one polluted field. In the present research we developed another approach based on US. First, we randomly generated a series of AAPFs. Next, we used the sampling data (logarithm) on each AAPF to generate the realization by Kriging. US honors the overall mean, variance, and spatial correlation while disregarding observations at specific sampling locations [40]. As a consequence, an AAPF generated by US keeps a certain spatial correlation and can show the concentration (logarithm data) at every single location. The parameters of a semivariogram based on the exponential model to generate AAPFs are set as follows: sill = 1; correlation scale = 30 m; nugget = 0; mean = 1. The total area of each AAPF is 90 m × 90 m = 8100 m^2. The threshold value is changeable, and different threshold values can generate AAPFs with different pollution rates (PRs), as shown in **Figure 1**.
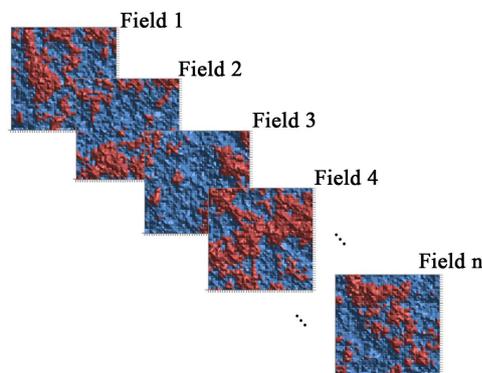

Figure 1. AAPFs randomly generated by US. Red area: polluted soil (>=Threshold value); Blue area: unpolluted soil (<Threshold value).

2.2. Preliminary Grid Sampling

Each AAPF is regularly or uniformly divided into N = n × n square meshes, and a sample is taken at the center of each mesh. The total number of samples is equal to the total number of the square meshes (N), depending on the number of samples, or number of meshes (n) of each side of the AAPF (**Figure 2**). In this study we set n in a range from 1 to 45, so the total number of samples ranges from 1 to 2025.

2.3. Remediation Method

Excavation, solidification/stabilization, soil washing, electro-remediation, and phytoremediation have all been used as *in-situ* and *ex-situ* techniques to reduce the impact of metals in soil. Notwithstanding, the remediation of contaminated soils by HMs is still recognized as one of the most difficult problems to solve. A few technologies are available for this purpose, but all are costly [41]-[43]. The policy in Japan is usually to move all of the "possibly polluted" soil (the excavation method): in total, about 86% of all remediation projects in Japan for HM-polluted soil adopt this method [44]. The excavation method is costly, but it removes almost all of the total contaminants, as well as the risk of contamination to neighboring soils, waters, and plants. It also can be performed in relatively little time without introducing other chemicals or requiring further monitoring or treatment on site. If we can use certain methods or techniques to reduce the cost, excavation can become an ideal method for the actual remediation of polluted soil. In this study we treat the excavation method as the optimal strategy.

2.4. Performance Criteria

To apply the optimal strategy to real-world remediation projects, the parameters used in the strategy should be as close as possible to real-world conditions and reflect the issues that most directly concern stakeholders or decision-makers: TF and CTR. TF consists of a sampling fee (SF: includes the preliminary sampling fee and additional sampling fee) and remediation fee (RF: includes the analysis fee, transport fee, treatment fee, etc.). One remediation scenario is assumed before the calculation of TF and CTR shown in **Figure 3**. The calculation formula is as follows:

$$TF = SF + RF = \text{number of samples} \times 17.5 + ARA \times 4.5 \times 0.5 \qquad (1)$$

$$CTR = \frac{CRA}{APA} \times 100\% \qquad (2)$$

where 17.5 and 4.5 are the unit prices of sampling and remediation selected with actual market unit prices, respectively. We had two reasons for selecting a soil depth of 0.5. First, HMs generally accumulate more in soil between 0 and 5 cm than at other depths, and most of the studies so far performed have focused on the sampling
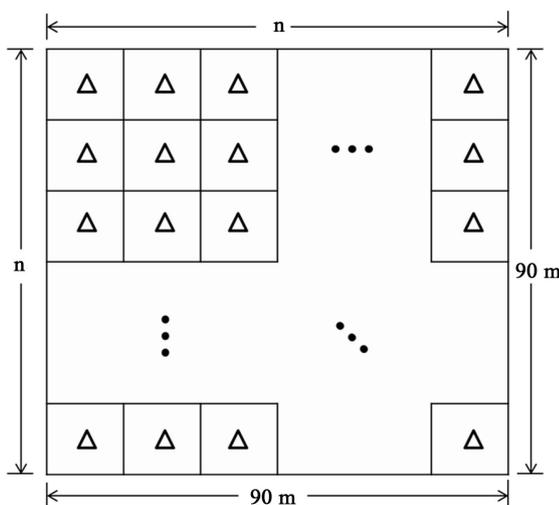

Figure 2. Grid sampling on an AAPF. n: sample or mesh number of each side of the square field; Δ: sampling point at the center of each mesh.
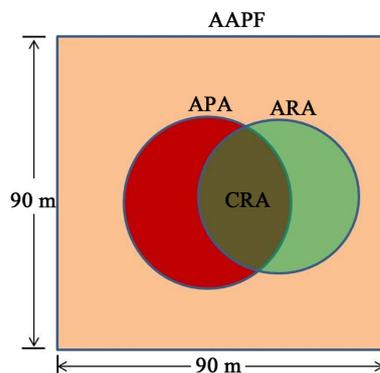
Figure 3. One assumed remediation scenario on an AAPF. APA: the Actual Polluted Area of the AAPF (red area); ARA: the Actual Remediated Area of the AAPF, (green area); CRA: the Correctly Remediated Area of the APA of the AAPF (area where the red and green areas overlap).

of surface soil (≤20 cm) [10] [45]-[48]. Besides, HMs tend to accumulate in the soil layers rich in organic matter and clay at depths from 0 to 50 cm, because HMs have a high affinity with organic matter and clay that prevents them from rapidly leaching [49] [50]. These layers were also a target of focus to quantify the effect of spatial variability of soil contaminants in cost calculations for soil remediation [51]. Second, more layers (>50 cm) result in more variation, because remediation costs increase quickly with increasing depth and uncertainty increases as more layers are used due to increasing scarcity of available data at deeper depths [51].

To reduce the number of performance criteria in the GA optimization procedure for additional samples and optimally operate the trade-off between TF and CTR, the adjusted total fee (ATF) is introduced with the following function:

$$ATF = \frac{TF}{CTR^{\partial}}(\partial \geq 0) \tag{3}$$

where the value of ∂ decides the weight between TF and CTR, that is, where ATF focuses more on TF when the value of ∂ gets smaller (or TF has more influence on the final ATF value than on CTR) and focuses more on CTR when the value of ∂ gets bigger. When the value of ∂ is fixed, the strategy with a lower ATF value is considered better. In a real-world project, the optimal strategy is always expected to reach a high level of accuracy at a given stage of investigation. In this research, the value of ∂ is subjectively set as 2. The performance criterion ATF, however, is imperfect, insufficient, and subject to its own constraint as an assessment standard, since the strategies may have the same ATF values but different TFs and CTRs (e.g. two strategies have the same ATFs (such as $ATF_1 = ATF_2$), whereas the TFs and CTRs may differ (such as $TF_1 < TF_2$, $CTR_1 < CTR_2$). In this case, the ATF (e.g. ATF_2) with a high CTR (CTR_2) is selected as the relatively better strategy in this study.

2.5. Geostatistical Interpolation (Kriging Estimation)

We performed this study using ordinary Kriging (OK), a method illustrated in detail in earlier literature [52]-[54]. To implement Kriging here, the exponential semivariogram model is adopted for the spatial interpolation at unsampled locations using log-transformed sampling data, and a semivariogram is calculated for log-transformed data to minimize the effect of extreme outliers [10]. Log-normal Kriging has theoretical limits, primarily due to the sensitivity of the back transformation from logarithmic data space, but the method yields the best results when data are highly skewed or relatively sparse [37].

2.6. Optimization of Additional Samples by a Genetic Algorithm (GA) Approach

Genetic algorithms (GAs) are a subset of heuristic evolutionary algorithms that generate globally optimal or near-optimal solutions for diverse, complex, and globally distributed problems by mimicking the process of natural evolution. While heuristic optimization methods cannot ensure global optimal solutions, they have the advantage of searching discrete solution-spaces globally. Gradient-based search methods lack this capability, as

they require continuous solution-spaces and run the risk of getting stuck in local optimal solutions [55]. The three fundamentals used for running GA programming are the representation of the parameters to be optimized (or genes), the genetic operator, and the objective function [56]. In this study, additional samplings are represented as genes, and the ATF is the objective function. A detailed description of the GA process is shown in **Figure 4**.

All of the statistic and optimization analyses in this study are performed with programs created and developed in VBA for Microsoft Excel (2010).

3. Results and Discussion

3.1. Preliminary Grid Sampling Analysis

Figure 5(A)-(c) shows the relationship between the ATF and number of samples by summarizing 100 randomly generated AAPFs. The general trend of the average ATF (μ) first decreases and then increases as the number of samples increases. The standard deviations (SDs: σs) decrease sharply as the number of sample increases while the number of samples is still small, but change little afterwards. When the number of samples is quite small, the ATF and SD fluctuate largely. Hence, taking a small number of samples from a real-world polluted field always introduces a high risk of uncertainty for both the TF and CTR. From **Figure 5(A)-(d)**, a partial enlarged drawing around the turning point of **Figure 5(A)-(c)**, we can see a somewhat lower ATF when the number of samples ranges from 81 to 169, and is lowest when the number of samples is 121. Based on the performance criterion of ATF in this research, SGS with 121 samples, that is, more samples than required by Japanese law (=81), is performed on a 90 m × 90 m square polluted field. Other parameters used for the generation of AAPFs by US such as the sill, correlation scale, and sample price (SP) are also discussed to elucidate the change between the ATF and number of samples.

When sill values vary in the manner shown in **Figure 5(A)**, the optimal numbers of samples for keeping a relatively lower ATF does not widely differ (from 81 - 169), and the relative lowest ATF appears when number of samples is near 121. Hence, the optimal number of samples for keeping a lower ATF is not closely related to the sill.
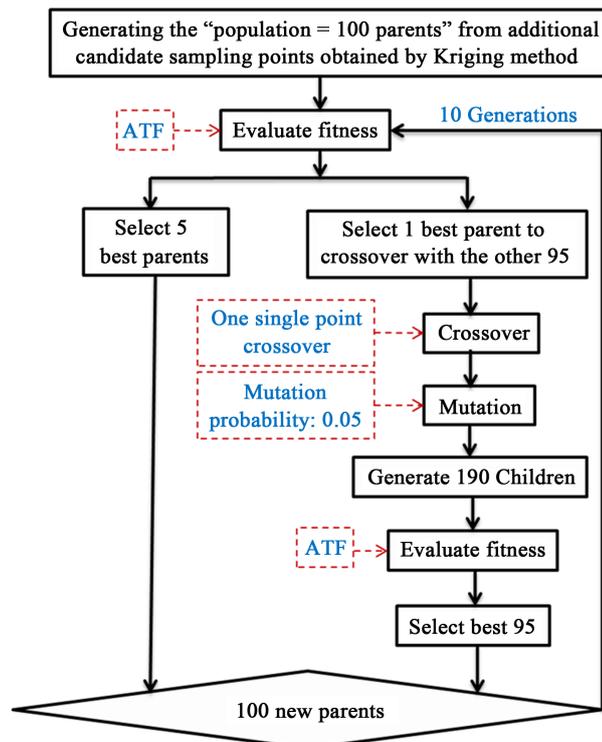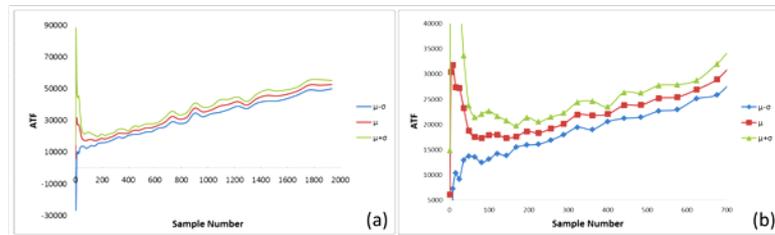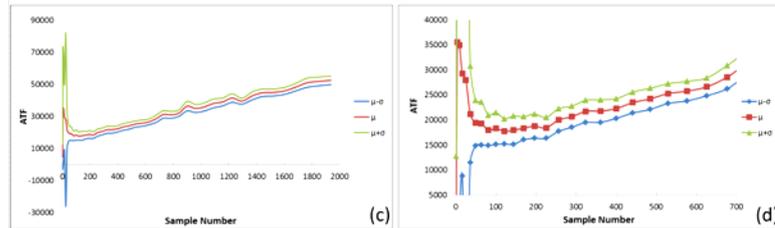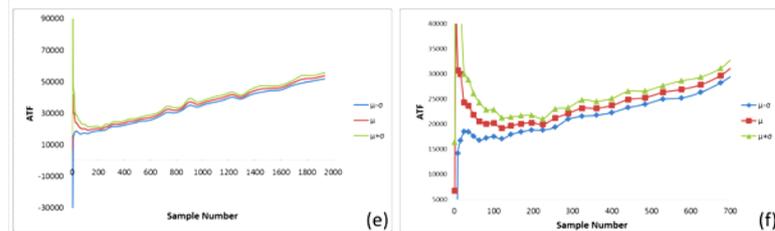

Figure 4. Flowchart of the GA process used in this study.

Sill=0.5, Correlation scale=30m, SP=17.5(Ten thousand Yen)

Sill=1, Correlation scale=30m, SP=17.5 (Ten thousand Yen)

Sill=1.5, Correlation scale=30m, SP=17.5 (Ten thousand Yen)

(A)

Sill=1, Correlation scale=15m, SP=17.5 (Ten thousand Yen)
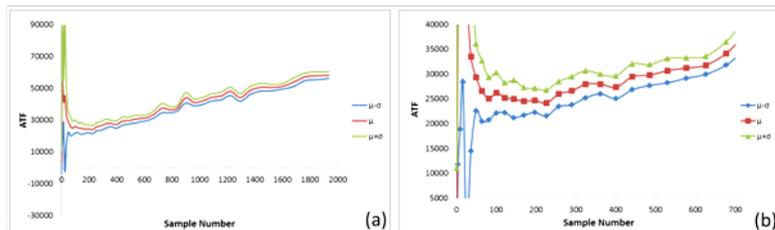
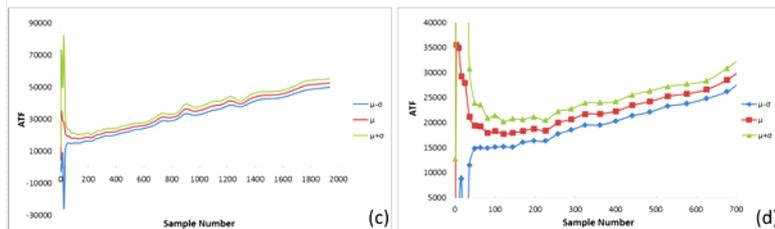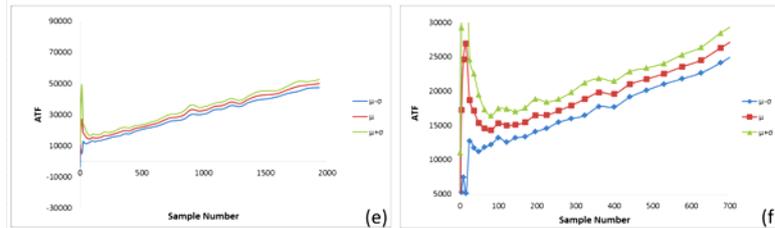**Sill=1, Correlation scale=30m**, SP=17.5 (Ten thousand Yen)

Sill=1, Correlation scale=45m, SP=17.5 (Ten thousand Yen)

(B)

Sill=1, Correlation scale=30m, **SP=10** (Ten thousand Yen)

Sill=1, Correlation scale=**30m**, **SP=17.5** (Ten thousand Yen)

Sill=1, Correlation scale=30m, **SP=20** (Ten thousand Yen)

(C)

Figure 5. (A) Relationship between the ATF and sample number determined by summarizing 100 randomly generated AAPFs while changing the values of the sill. Figures (b), (d), and (f) are partial enlargements around the turning points of figures (a), (c), and (e). μ is th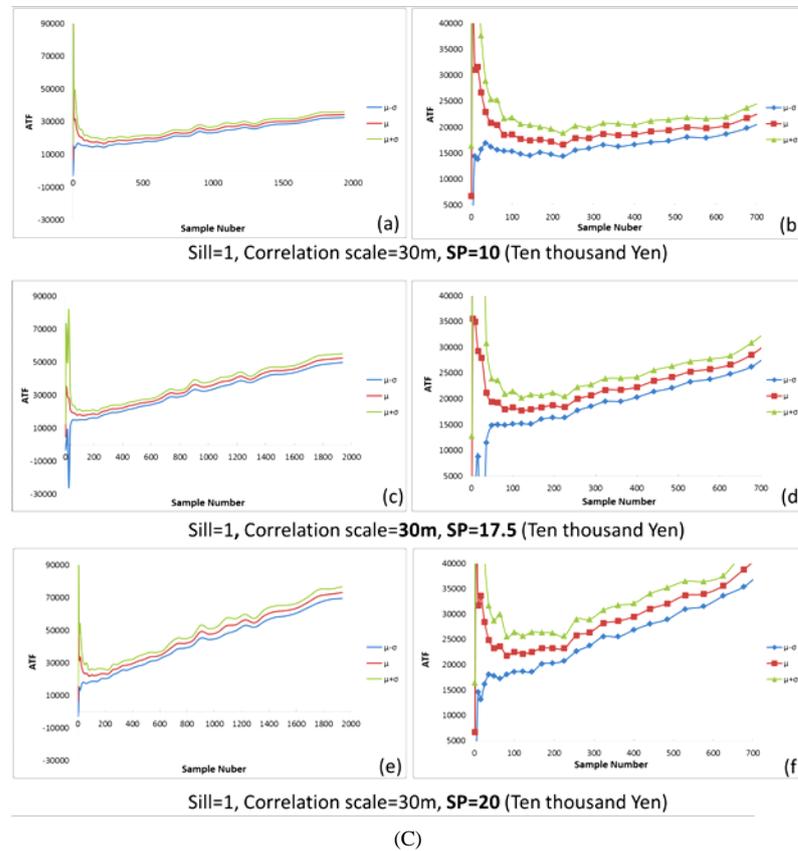e average value and σ is the corresponding standard deviation; (B) Relationship between the ATF and sample number determined by summarizing 100 randomly generated AAPFs while changing the values of the correlation scale. Figures (b), (d), and (f) are partial enlargements around the turning points of figures (a), (c), and (e). μ is the average value and σ is the corresponding standard deviation; (C) Relationship between the ATF and sample number determined by summarizing 100 randomly generated AAPFs while changing the sample price (SP). Figures (b), (d), and (f) are partial enlargements around the turning points of figures (a), (c), and (e). μ is the average value and σ is the corresponding standard deviation.

When the values of correlation scale differ in the manner shown in **Figure 5(B)**, the SDs of ATF do not show much obvious change. The optimal number of samples for a relatively lower ATF decreases as the correlation scale increases, because of a lower correlation scale value when the sill is fixed signifies that the AAPF only has spatial correlation on a smaller scale, which requires denser sampling data to give a more accurate estimation.

While SP changes, the optimal number of samples with the relative lowest ATF varies. As **Figure 5(C)** illustrates, the optimal number of samples with the relative lowest ATF appears around 225 when the SP decreases from 17.5 to 10, and approaches 81 when the sample price increases from 17.5 to 20. Additionally, the SD of the ATF generally increases as the SP increases, because the same uncertainty with higher SP will result in a larger change in the ATF value. We therefore propose that more samples can be taken in the real-world treatment of a polluted field when the sample price decreases, and *vice versa*.

3.2. Geostatistical (Kriging) Analysis

Based on the result of preliminary grid sampling analysis (optimal number of samples, 121; assumed threshold value, 1.5), one AAPF (shown in **Figure 6(A)**) from a total of 150 randomly generated AAPFs is selected as one example to continually perform the geostatistical analysis. The analysis basically consists of three steps: (1) exploratory data analysis; (2) structural analysis (variogram model selection, parameter determination, and model validation); and (3) prediction (Kriging estimation error and simulations) [35].
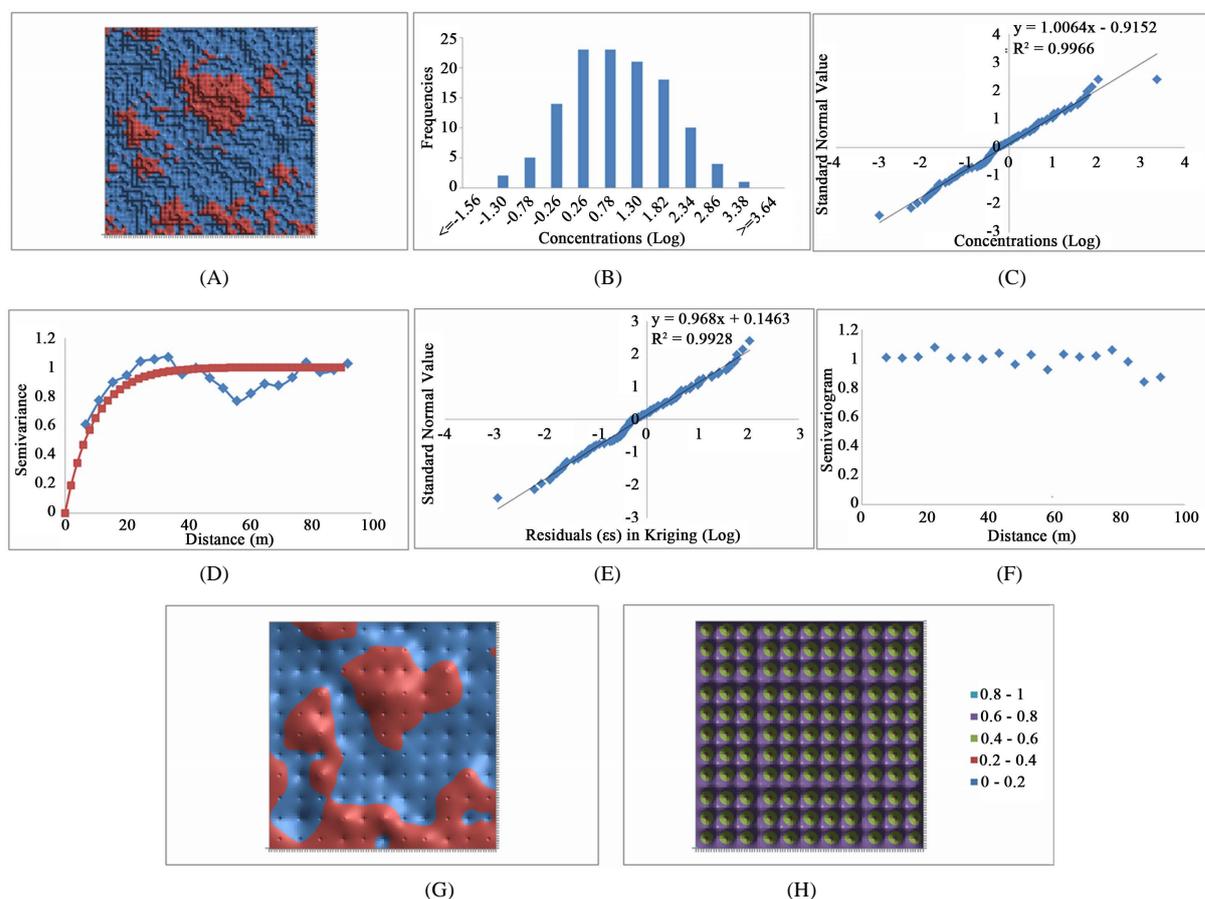
Figure 6. (A) One randomly generated AAPF; (B) Histogram of concentrations on the randomly selected AAPF; (C) Normal plot of concentrations (Log); (D) Sample semivariogram for concentrations (blue line) and the exponential model (red line) fitted to the sample semivariogram; (E) Normal plot of residuals (εs) in Kriging; (F) Sample semivariogram for the residuals (εs) in Kriging; (G) Kriging estimation of pollution distribution (CI = 75%); (h) Kriging Standard Deviation (SD) of 121 preliminary grid sampling data.

A histogram of the sampling data on this randomly selected AAPF is illustrated in **Figure 6(B)**. Although OK requires no data to distribute normally, predictions are usually performed better if the sample histogram is displayed without strong skewness. The use of normally distributed data, is essential, however, if one plans to use the Kriging SD to build confidence intervals (CIs) for the estimation (e.g. in this study) [27] [35]. From both **Figure 6(B)** and **Figure 6(C)**, we see that the sampling data generally conform well with the symmetrical shape of the normal distribution.

The sample semivariogram calculated to account for the spatial structure of concentrations (see **Figure 6(D)**) is well fitted to the exponential model (sill = 1, correlation scale = 28.5 m). The goodness of fit was evaluated by the residuals (εs) in Kriging, which is introduced in detail in Kitanidis' book on the application of geostatistics to hydrogeology [57]. When we follow the same calculation with **Figure 6(C)**, we learn from **Figure 6(E)** that the (εs) generally display the normal distribution well. Since the (εs) are orthonormal (*i.e.* uncorrelated with each other and normalized to have unit variance), the (εs) should only have the sample semivariogram with a nugget-effect value of about 1 at the short-correlation-scale distance. At a long-correlation-scale distance, on the other hand, they might show heavily fluctuated scatter points from 1. Clearly, as shown in **Figure 6(F)**, the sample semivariogram values of (εs) are very close to 1 at short-correlation-scale distance, which illustrates again that the exponential modelfits the sample semivariogram better. **Figure 6(G)** and **Figure 6(H)** respectively show the estimation values of Kriging (CI = 75%) and the corresponding SD distributions. As **Figure 6(H)** reveals, the SDs around the existing sampling locations are small and will increase as the distance (between the unsampled locations to be estimated by Kriging and the existing sampling locations) increases.

By continuously using different Confidence Intervals (CIs from 50% to 95% in 5% increments, and 99%) on this AAPF after the structural analysis, the Kriging generates solutions (realizations) for the treatment areas with different corresponding ATFs until the best solution with the lowest ATF can be obtained. The best CI values, however, differ for different AAPFs. By analyzing the best CI values for different AAPFs, we find some relation between the best CI values and pollution rates (PR: APA/ total area of AAPF*100%, as shown in **Figure 3**). An exploration of this relation with the 150 AAPFs (PR: 4% - 84%) reveals a stepwise reduction of CIs as the PR of the AAPFs increase (**Figure 7(A)**). In real-world polluted sites, meanwhile, there is no way of knowing the PR beforehand (as with the AAPF), because the concentration of HM is difficult to determine at each single location. This obstacle makes it quite difficult to use the PR to suggest the corresponding best CI value used in the Kriging. We do, however, have the option of using the PR reflected by the sampling data (SDPR) as a proxy for the PR, since the SDPR and PR have a linear relation (see **Figure 7(B)**). After using the SDPR instead of PR, we can obtain the relation between the SDPR and corresponding best CI. When we use this approach to obtain the relation shown **Figure 7(C)**, we still find a stepwise reduction of the CI value as the SDPR increases.

The best CI range and best average CI value (calculated from the best CI range) can be suggested for AAPFs with different SDPR ranges, which are generally summarized in **Table 1**. We also find suggestions for the best CI range and best average CI value when the SDPRs are at 0% - 4% and 84% - 100%, but these are subjectively based on the trend of the best average CI values suggested in the other SDPR ranges in **Table 1**. **Figure 8(A)** compares the results between the ATF of the preliminary grid sampling analysis and the ATF of the Kriging analysis. The ATFs that use the corresponding best CI value for each AAPF in the Kriging estimation (green line) can be substituted for the ATFs that use the corresponding best average CI value (red line), since they fluctuate at almost the same level and have similar values. Both of the ATFs (red and green line) may be far smaller than the ATFs based only on the preliminary sampling data (blue line), although the differences among them become smaller as the PR rate of the AAPF (AAPF number) increases. If the CTR is considered apart from the performance criterion ATF, the CTRs that use the best CI value (green line) and best average CI value (red line) fluctuate at the same level and have similar values, as shown in **Figure 8(B)**. Both of the CTRs are greatly improved compared with the CTR using only the preliminary sampling data (blue line), although the differences among them become smaller as the PR (or SDPR) rate of the AAPF (or AAPF number) increases.
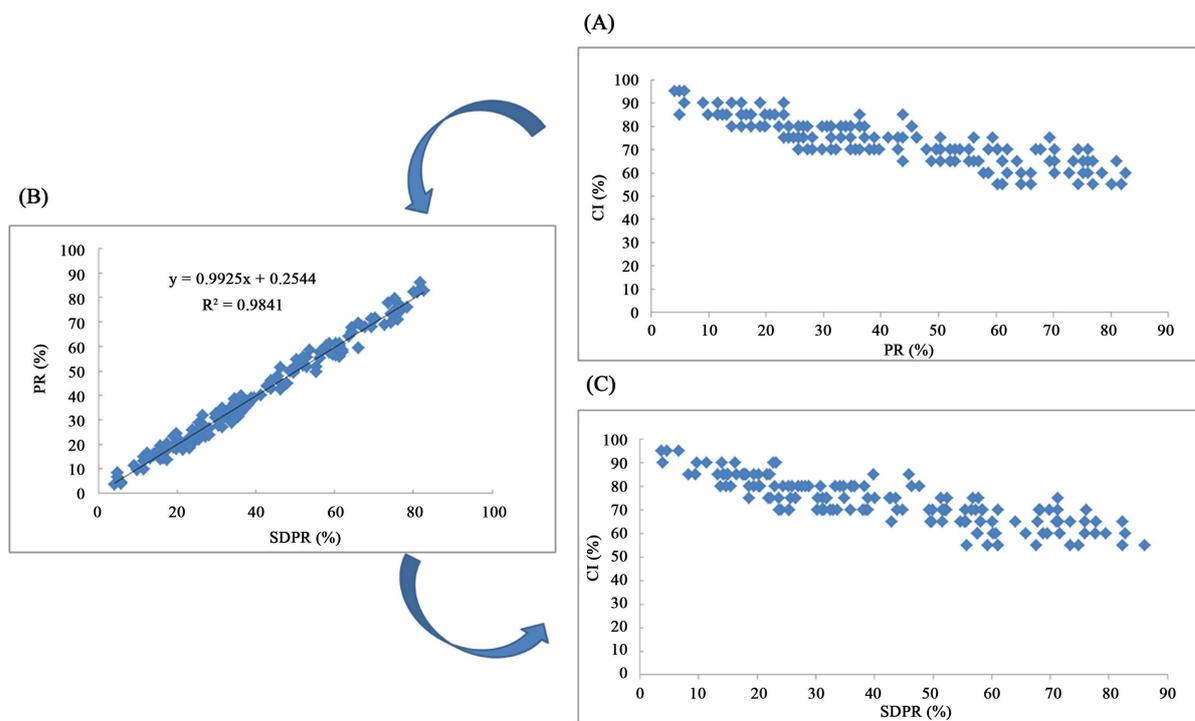

Figure 7. (A) Relation between PR and the corresponding best CI used in Kriging; (B) Relation between PR and SDPR; (C) Relation between SDPR and the corresponding best CI used in Kriging.
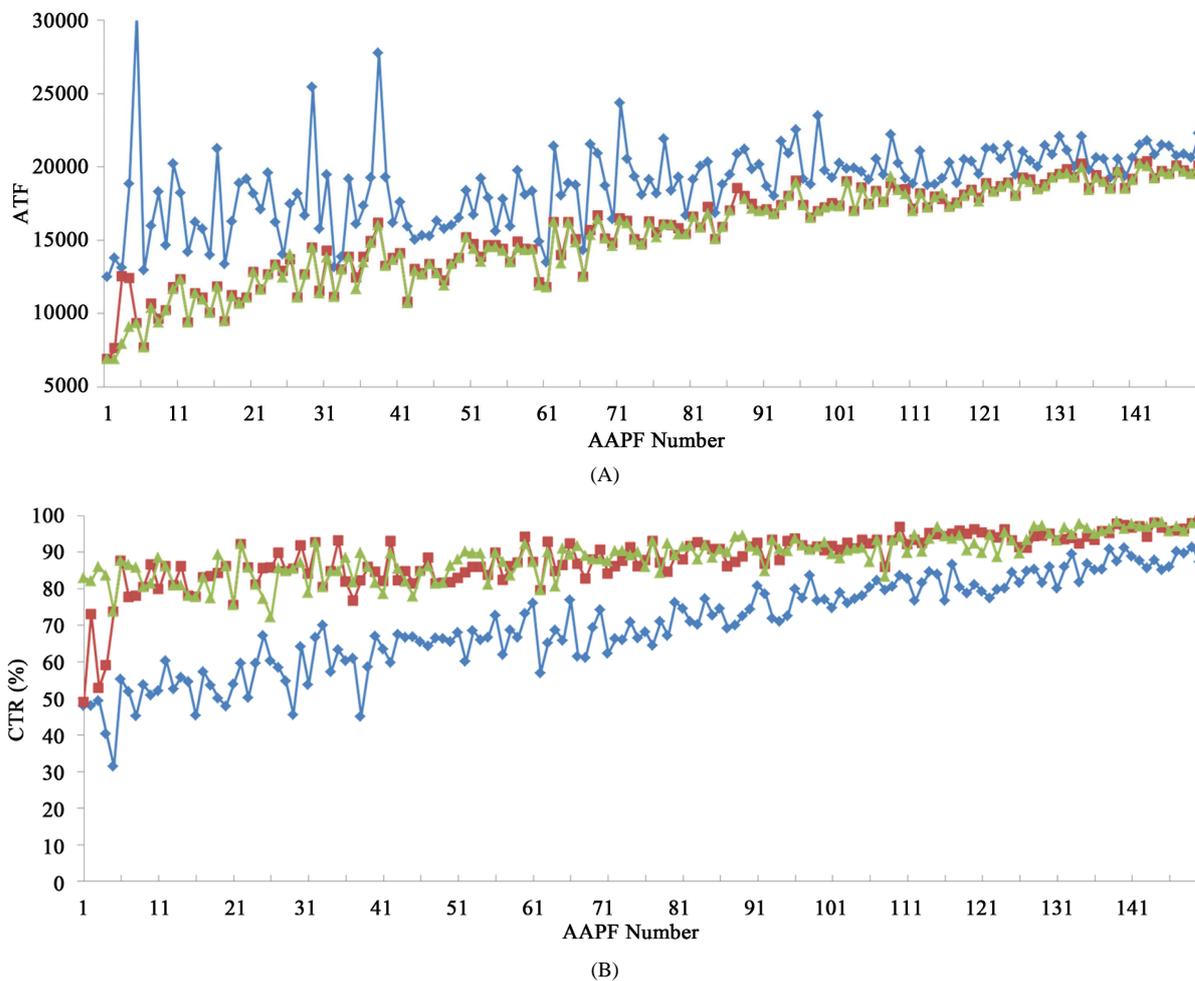
(A)

(B)

Figure 8. The different ATFs (A) and corresponding CTRs (B) for each of the 150 AAPFs. Blue line: ATF (CTR) for the preliminary grid sampling method. Green line: ATF (CTR) after Kriging estimation by selecting the best CI value for each AAPF. Red line: ATF (CTR) after Kriging estimation by selecting the best average CI value. The AAPF number is ranked in ascending order of PR or SDPR (larger AAPF number, with a larger PR or SDPR value).

Table 1. The best CI values of Kriging suggested for different SDPR ranges.

SDPR Ranges (%)	0 - 4[a]	4 - 14	14 - 24	24 - 44	44 - 64	64 - 84	84 - 100[a]
Best CI range for Kriging (%)	95 - 99[a]	85 - 95	80 - 90	70 - 80	65 - 75	55 - 70	50 - 60[a]
Best average CI values for Kriging (%)	95[a]	90	85	75	70	60	55[a]
Best CI range for additional samples (%)	95 - 99[a]	90 - 95	90 - 95	80 - 85	75 - 80	65 - 70	55 - 60[a]

[a]These values are extrapolated based on the trend of best CI values suggested in the other SDPR ranges.

Based on the analysis above, the combination of preliminary grid sampling and the Kriging method can result in more accurate estimations of pollution distribution with relatively higher CTRs, which in turn can result in more cost-effective strategies with relatively lower ATFs for decision-makers for remediation. As a result, **Table 1** can serve as a reference when using Kriging to estimate the pollution distribution on a single polluted field before making a cost-effective plan for remediation.

3.3. Genetic Algorithm (GA) Optimization Analysis

Ten AAPFs were randomly selected in each SDPR range (except the first range (1% - 4%) and last range (84% -

100%)) based on **Table 1**. Hence, a total of 50 AAPFs were used to conduct the optimization work. Five candidate areas contoured by two neighboring Kriging CI values (from 50% to 95% in 5% increments, and 99%) around the best average CI value (**Table 1**) were selected to conduct the GA optimization on each randomly selected AAPF. Assume, for example, that the SDPR of one randomly selected AAPF is 8%, which belongs to the SDPR range 4% - 14%, so that the corresponding best average CI value suggested is 90% based on **Table 1**. The 5 section candidate areas with CI ranges (90% - 95%, 85% - 90%, 95% - 99%, 80% - 85%, and 75% - 80%) around this best average CI value (90%) can thus be selected to operate the GA optimization and then search the relatively optimal additional samples, which keep the ATF of the whole strategy lowest. The additional number of samples should preferably be kept small to run the optimization procedure at a fine resolution [36]. In this research we selected 23 additional samples because the ATF calculated by the 121 preliminary grid sampling data together with the 23 additional sampling data can be compared with the ATF calculated by 121or 144 preliminary grid sampling data either in combination with the Kriging estimation or without it combined with Kriging estimation or not. Besides, the additional sampling number 23 has no obvious effect on the semivariogram structure when using the Kriging estimation.

Here we use additional samples taken on 5 candidate areas (contoured by CI ranges: 75% - 80%, 80% - 85%, 85% - 90%, 90% - 95%, and 95% - 99%) of 10 randomly selected AAPFs (SDPR Range: 4% - 14%) as one example to show the changes of the ATFs after GA optimization. **Figure 9(A)** illustrates that the relative lowest ATF appeared in the CI range of 90% - 95%, whereas none of the ATFs after GA optimization on the 5 candidate areas showed much difference. The corresponding CTRs, however, have the general decreasing trend shown in **Figure 9(B)**. Compared with both **Figure 9(A)** and **Figure 9(B)**, we can always obtain a relatively lower ATF with a higher CTR in the CI range of 90-95%. We can also obtain a relatively lower ATF with a higher CTR in the CI range of 95% - 99%, but this CI range always contributes to a larger candidate area and increases the difficulty in selecting additional sampling locations. Hence, we selected 90% - 95% as the best CI range for the candidate area for additional sampling. **Table 1** shows the best CI range suggested for the additional sampling based on the same evaluation of the ATFs on other randomly selected AAPFs with different SDPR ranges (14% - 24%, 24% - 44%, 44% - 64%, 64% - 84%). When the SDPR ranges are at 0% - 4% and 84% - 100%, the best CI ranges are subjectively suggested to be 95% - 99% and 55% - 60%, respectively, based on the trend of the best CI ranges values suggested in other SDPR ranges.
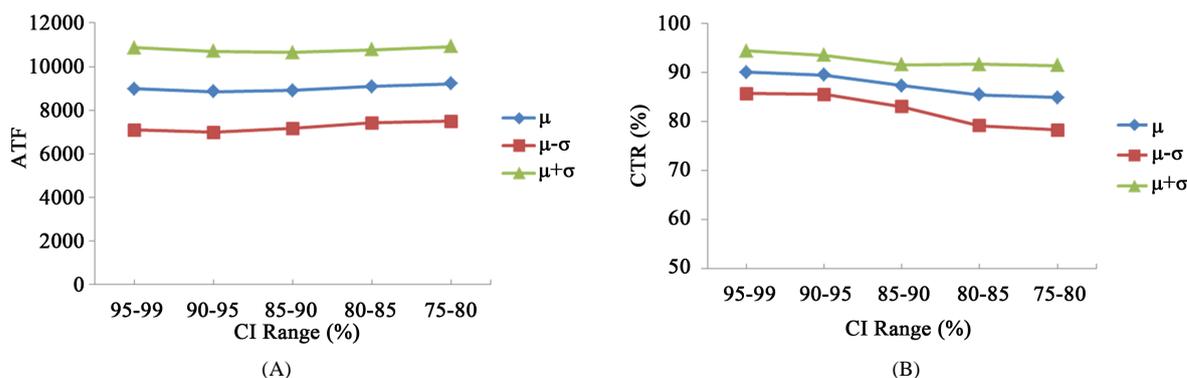

Figure 9. ATFs (a) and CTRs (b) after GA optimization on 5 candidate areas (CI Ranges: 75% - 99%) of 10 randomly selected AAPFs (SDPR Range: 14% - 24%). μ is the average value and σ is the corresponding standard deviation.

As a result, the corresponding ATFs and CTRs for the 50 AAPFs were compared with the ATFs and CTRs calculated by the other three approaches, after GA optimization using the best CI range suggested for additional samples in **Table 1**: 144 preliminary grid sampling data, 144 preliminary sampling data combined with Kriging estimation, and 121 preliminary sampling data combined with Kriging estimation. As shown in **Figure 10(A)**, the ATFs after Kriging estimation (blue line and green line) are considerably lower than the ATF (red line) calculated only from the preliminary sampling data. We find, however, that the ATFs (purple line) after GA optimization and Kriging estimation are always the lowest, even though this may be difficult to tell from the visual plot in **Figure 10(A)**. As the SDPR of the AAPFs increases (or the AAPF number increases), the ATFs generated by these four approaches become inconspicuous and bunched more closely together, since an AAPF with a
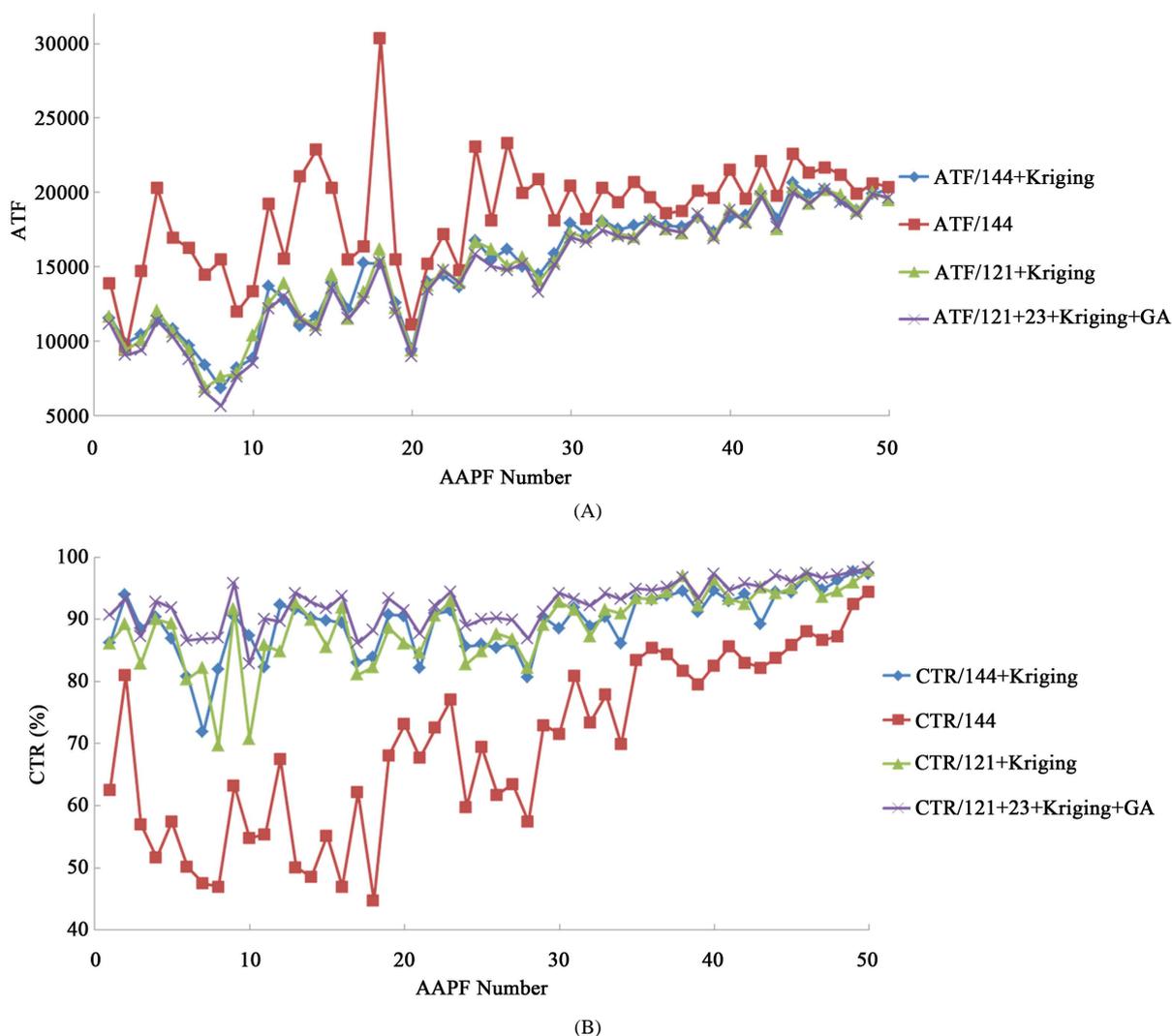
(A)

(B)

Figure 10. Comparison of the ATFs (A) and CTRs (B) calculated by different combinations of preliminary sampling data, Kriging estimation, and GA optimization. The 121 and 144 in the legend are the preliminary numbers of samples and 23 is the additional number of samples. The AAPF numbers from 1 to 50 in 10 increments correspond to the SDPR ranges of 4% - 14%, 14% - 24%, 24% - 44%, 44% - 64%, and 64% - 84%, respectively.

higher SDPR always offers relative less space for optimization. Even though the ATFs (blue, green, and purple lines) are closer to each other, the structures of the ATFs (the ratios or weights of the TFs and CTRs) vary. From **Figure 10(B)** we can see that the corresponding CTR (purple line) is always the highest. We also find, in both **Figure 10(A)** and **Figure 10(B)**, that the ATFs and CTRs (blue lines and green lines) show relatively little difference. From this we can infer that a traditional increase of grid sampling data will not necessarily generate a more cost-effective effort. Through the additional sampling data optimized by GA, we can lower the ATF and increase the CTR simultaneously. We also need to narrow the candidate area for additional samples and consider their detailed locations, which is invariably difficult when dealing with real polluted sites. Limitations in the GA optimization procedure in this study may preclude identification of the optimal additional sampling locations, since the termination was set at 10 generations in consideration of computing power (CPU limitation of computer) and time. Even so, the optimized additional samples by GA in this study suggested the relative optimal locations.

In a study on multi-phase sampling for soil remediation surveys by Marchant *et al.*, the additional samples were generally dispersed on the boundaries between areas which, according to the first phase, either required or did not require remediation [33]. Our result shown in **Figure 11** is similar. The additional samples (green bub-
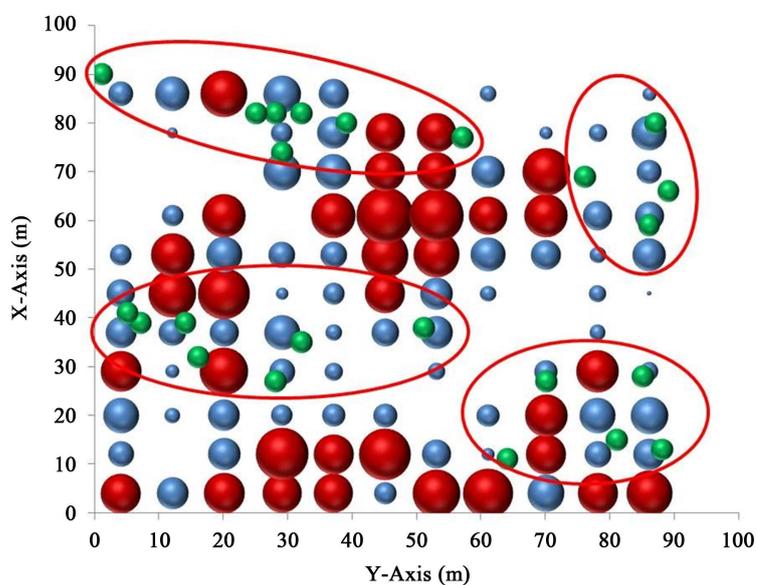
Figure 11. Distribution of 121 preliminary samples (red and blue bubbles) and 23 additional samples (green bubbles) on the AAPF used in **Figure 6**. The sizes of the red and blue bubbles represent the concentrations of samples above and below the threshold value, respectively, while the green bubbles only represent the additional sample location without considering the size.

bles inside the red circles) are always selected at the joint locations of polluted and unpolluted areas, where there are abundant sampling data near the threshold value, whereas the additional samples are generally dispersed over the full scale of the AAPF rather than concentrated in any one area.

In conclusion, our study combined the grid sampling method, Kriging estimation, and GA optimization for additional samples into an overall strategy for investigating and remediating polluted fields. Many different AAPFs generated by US are used in place of real-world polluted sites to give relative optimal suggestions to engineers or decision-makers, such as parameters to use in Kriging estimation (as shown in **Table 1**) and suggested locations for additional samples (as shown in **Figure 11**), to develop more cost-effective (balanced TF and CTR) strategies before dealing with the real-world polluted sites. The study thus exemplifies the advantages of integrating geostatistical interpolation and GA optimization into the practice of soil investigation and remediation, even though the model relies on certain assumptions and the strategy is unhelpful when the pollution of a polluted field shows no spatial correlation. We are now considering ways to use and validate this strategy in practical circumstances in the future.

Acknowledgements

References

[1] Qu, C.-S., *et al*. (2012) Human Exposure Pathways of Heavy Metals in a Lead-Zinc Mining Area, Jiangsu Province, China. *PLoS ONE*, **7**, e46793. http://dx.doi.org/10.1371/journal.pone.0046793

[2] Li, Z., Ma, Z., Kuijp, T. and Vander, J. (2014) A Review of Soil Heavy Metal Pollution from Mines in China: Pollution and Health Risk Assessment. *Science of the Total Environment*, **468-469**, 843-853. http://dx.doi.org/10.1016/j.scitotenv.2013.08.090

[3] Chen, R., Sherbinin, A.D.E., Ye, C. and Shi, G. (2014) China's Soil Pollution: Farms on the Frontline. *Science (Letters)*, **344**, 691.

[4] Yang, H., Huang, X., Thompson, J.R. and Flower, R.J. (2014) China's Soil Pollution: Urban Brown Fields. *Science (Letters)*, **344**, 691-692.

[5] Zhao, F., Ma, Y., Zhu, Y., Tang, Z. and Mcgrath, S.P. (2015) Soil Contamination in China: Current Status and Mitiga-

tion Strategies. *Environmental Science & Technology*, **49**, 750-759. http://dx.doi.org/10.1021/es5047099

[6]   (2014) The Ministry of Environmental Protection; The Ministry of Land and Resources Report on the National Soil Contamination Survey. http://www.mep.gov.cn/gkml/hbb/qt/201404/t20140417_270670.htm

[7]   Petersen, L., Minkkinen, P. and Esbensen, K.H. (2005) Representative Sampling for Reliable Data Analysis: Theory of Sampling. *Chemometrics and Intelligent Laboratory Systems*, **77**, 261-277. http://dx.doi.org/10.1016/j.chemolab.2004.09.013

[8]   Fortunati, G. and Pasturenzi, M. (1994) Quality in Soil Sampling. *Química Analítica*, **13**, S5-S20.

[9]   Markert, B. (1995) Quality Assurance of Plant Sampling and Storage. In: Quevauviller, P., Ed., *Quality Assurance in Environmental Monitoring Sampling and Sample Pretreatment*, VCH Weinheim, New York, 215-254.

[10] Andronikov, S.V., Davidson, D.A. and Spiers, R.B. (2000) Variability in Contamination by Heavy Metals: Sampling Implications. *Water*, *Air*, & *Soil Pollution*, **120**, 29-45. http://dx.doi.org/10.1023/A:1005261522465

[11] Kabir, E., *et al*. (2012) Current Status of Trace Metal Pollution in Soils Affected by Industrial Activities. *The Scientific World Journal*, **2012**, 1-18. http://dx.doi.org/10.1100/2012/916705

[12] Chu, H.-J., Lin, Y.-P., Jang, C.-S. and Chang, T.-K. (2010) Delineating the Hazard Zone of Multiple Soil Pollutants by Multivariate Indicator Kriging and Conditioned Latin Hypercube Sampling. *Geoderma*, **158**, 242-251. http://dx.doi.org/10.1016/j.geoderma.2010.05.003

[13] Hooda, P.S. and Glavinandp, R.J. (2005) A Practical Examination of the Use of Geostatistics in the Remediation of a Site with a Complex Metal Contamination History. *Soil and Sediment Contamination*, **14**, 155-169. http://dx.doi.org/10.1080/15320380590911814

[14] Lord, D.W. (1987) In Reclaiming Contaminated Land. Springer, Netherlands, 62-113. http://dx.doi.org/10.1007/978-94-011-6504-4_4

[15] Flatman, G.T. and Yfantis, A.A. (1984) Geostatistical Stragegy for Soil Sampling: The Survey and the Census. *Environmental Monitoring and Assessment*, **4**, 335-349. http://dx.doi.org/10.1007/BF00394172

[16] Royal, A.G. (1981) A Practical Introduction to Geostatistics.

[17] Theocharopoulos, S., *et al*. (2001) European Soil Sampling Guidelines for Soil Pollution Studies. *Science of the Total Environment*, **264**, 51-62. http://dx.doi.org/10.1016/S0048-9697(00)00611-2

[18] Lin, Y.-P., Chu, H.-J., Huang, Y.-L., Cheng, B.-Y. and Chang, T.-K. (2010) Modeling Spatial Uncertainty of Heavy Metal Content in Soil by Conditional Latin Hypercube Sampling and Geostatistical Simulation. *Environmental Earth Sciences*, **62**, 299-311. http://dx.doi.org/10.1007/s12665-010-0523-5

[19] Juang, K.-W., Liao, W.-J., Liu, T.-L., Tsui, L. and Lee, D.-Y. (2008) Additional Sampling Based on Regulation Threshold and Kriging Variance to Reduce the Probability of False Delineation in a Contaminated Site. *Science of the Total Environment*, **389**, 20-28. http://dx.doi.org/10.1016/j.scitotenv.2007.08.025

[20] Milillo, T.M., Sinha, G. and Gardella, J.A. (2012) Use of Geostatistics for Remediation Planning to Transcend Urban Political Boundaries. *Environmental Pollution*, **170**, 52-62. http://dx.doi.org/10.1016/j.envpol.2012.06.006

[21] Ha, H., Olson, J.R., Bian, L. and Rogerson, P.A. (2014) Analysis of Heavy Metal Sources in Soil Using Kriging Interpolation on Principal Components. *Environmental Science & Technology*, **48**, 4999-5007. http://dx.doi.org/10.1021/es405083f

[22] Demougeot-Renard, H. and De Fouquet, C. (2004) Geostatistical Approach for Assessing Soil Volumes Requiring Remediation: Validation Using Lead-Polluted Soils Underlying a Former Smelting Works. *Environmental Science & Technology*, **38**, 5120-5126. http://dx.doi.org/10.1021/es0351084

[23] Qu, M., Li, W. and Zhang, C. (2013) Assessing the Risk Costs in Delineating Soil Nickel Contamination Using Sequential Gaussian Simulation and Transfer Functions. *Ecological Informatics*, **13**, 99-105. http://dx.doi.org/10.1016/j.ecoinf.2012.06.005

[24] Juang, K.-W., Lee, D.-Y. and Teng, Y.-L. (2005) Adaptive Sampling Based on the Cumulative Distribution Function of Order Statistics to Delineate Heavy-Metal Contaminated Soils Using Kriging. *Environmental Pollution*, **138**, 268-277. http://dx.doi.org/10.1016/j.envpol.2005.04.003

[25] Webster, R. and Oliver, M.A. (1992) Sample Adequately to Estimate Variograms of Soil Properties. *Journal of Soil Science*, **43**, 177-192. http://dx.doi.org/10.1111/j.1365-2389.1992.tb00128.x

[26] Lettenmaier, D.P. and Hughes, J. (1981) Data Requirements for Kriging: Estimation and Network Design. *Water Resources Research*, **17**, 1641-1650.

[27] Saito, H. and Goovaerts, P. (2000) Geostatistical Interpolation of Positively Skewed and Censored Data in a Dioxin-Contaminated Site. *Environmental Science & Technology*, **34**, 4228-4235. http://dx.doi.org/10.1021/es991450y

[28] Caro, A., Legarda, F., Romero, L., Herranz, M., Barrera, M., Valiño, F., Idoeta, R. and Olondo, C. (2013) Map on Pre-

dicted Deposition of Cs-137 in Spanish Soils from Geostatistical Analyses. *Journal of Environmental Radioactivity*, **115**, 53-59. http://dx.doi.org/10.1016/j.jenvrad.2012.06.007

[29] McGrath, D., Zhang, C. and Carton, O.T. (2004) Geostatistical Analyses and Hazard Assessment on Soil Lead in Silvermines Area, Ireland. *Environmental Pollution*, **127**, 239-248. http://dx.doi.org/10.1016/j.envpol.2003.07.002

[30] Saby, N.P.A., Marchant, B.P., Lark, R.M., Jolivet, C.C. and Arrouays, D. (2011) Robust Geostatistical Prediction of Trace Elements across France. *Geoderma*, **162**, 303-311. http://dx.doi.org/10.1016/j.geoderma.2011.03.001

[31] Aelion, C.M., Davis, H.T., Liu, Y., Lawson, A.B. and McDermott, S. (2009) Validation of Bayesian Kriging of Arsenic, Chromium, Lead, and Mercury Surface Soil Concentrations Based on Internode Sampling. *Environmental Science & Technology*, **43**, 4432-4438. http://dx.doi.org/10.1021/es803322w

[32] Demougeot-Renard, H., De Fouquet, C. and Renard, P. (2004) Forecasting the Number of Soil Samples Required to Reduce Remediation Cost Uncertainty. *Journal of Environmental Quality*, **33**, 1694-1702. http://dx.doi.org/10.2134/jeq2004.1694

[33] Marchant, B.P., McBratney, A.B., Lark, R.M. and Minasny, B. (2013) Optimized Multi-Phase Sampling for Soil Remediation Surveys. *Spatial Statistics*, **4**, 1-13. http://dx.doi.org/10.1016/j.spasta.2012.11.001

[34] Ersoy, A., Yunsel, T.Y. and Faculty, E. (2008) Geostatistical Conditional Simulation for the Assessment of Contaminated Land by Abandoned Heavy Metal Mining. *Environmental Toxicology*, **23**, 96-109. http://dx.doi.org/10.1002/tox.20314

[35] Ersoy, A., Yunsel, T.Y. and Cetin, M. (2004) Characterization of Land Contaminated by Past Heavy Metal Mining Using Geostatistical Methods. *Archives of Environmental Contamination and Toxicology*, **46**, 162-175.

[36] Verstraete, S. and Van Meirvenne, M. (2008) A Multi-Stage Sampling Strategy for the Delineation of Soil Pollution in a Contaminated Brownfield. *Environmental Pollution*, **154**, 184-191. http://dx.doi.org/10.1016/j.envpol.2007.10.014

[37] Gopalakrishnan, G., Minsker, B.S. and Valocchi, A.J. (2011) Monitoring Network Design for Phytoremediation Systems Using Primary and Secondary Data Sources. *Environmental Science & Technology*, **45**, 4846-4853. http://dx.doi.org/10.1021/es1042657

[38] Meerschman, E., Cockx, L. and Van Meirvenne, M. (2011) A Geostatistical Two-Phase Sampling Strategy to Map Soil Heavy Metal Concentrations in a Former War Zone. *European Journal of Soil Science*, **62**, 408-416. http://dx.doi.org/10.1111/j.1365-2389.2011.01366.x

[39] Xie, Y., Chen, T.-B., Lei, M., Yang, J., Guo, Q.-J., Song, B. and Zhou, X.-Y. (2011) Spatial Distribution of Soil Heavy Metal Pollution Estimated by Different Interpolation Methods: Accuracy and Uncertainty Analysis. *Chemosphere*, **82**, 468-476. http://dx.doi.org/10.1016/j.chemosphere.2010.09.053

[40] Hicks, P.J. (1996) Unconditional Sequential Gaussian Simulation for 3-D Flow in a Heterogeneous Core. *Journal of Petroleum Science and Engineering*, **16**, 209-219. http://dx.doi.org/10.1016/S0920-4105(96)00041-1

[41] Scanferla, P., Ferrari, G., Pellay, R., Ghirardini, A.V., Zanetto, G. and Libralato, G. (2009) An Innovative Stabilization/Solidification Treatment for Contaminated Soil Remediation: Demonstration Project Results. *Journal of Soils and Sediments*, **9**, 229-236. http://dx.doi.org/10.1007/s11368-009-0067-z

[42] Montinaro, S., Concas, A., Pisu, M. and Cao, G. (2012) Remediation of Heavy Metals Contaminated Soils by Ball Milling. *Chemical Engineering Transactions*, **28**, 187-192.

[43] Tica, D., Udovic, M. and Lestan, D. (2011) Immobilization of Potentially Toxic Metals Using Different Soil Amendments. *Chemosphere*, **85**, 577-583. http://dx.doi.org/10.1016/j.chemosphere.2011.06.085

[44] Wada, S. (2010) Dynamics of Heavy Metals in Soils. *Global Environmental Research*, **15**, 15-21. (In Japanese)

[45] Ağca, N. (2014) Spatial Distribution of Heavy Metal Content in Soils around an Industrial Area in Southern Turkey. *Arabian Journal of Geosciences*, **8**, 1111-1123. http://dx.doi.org/10.1007/s12517-013-1240-7

[46] Chen, T.-B., Zheng, Y.-M., Lei, M., Huang, Z.-C., Wu, H.-T., Chen, H., *et al.* (2005) Assessment of Heavy Metal Pollution in Surface Soils of Urban Parks in Beijing, China. *Chemosphere*, **60**, 542-551. http://dx.doi.org/10.1016/j.chemosphere.2004.12.072

[47] Dayani, M. and Mohammadi, J. (2010) Geostatistical Assessment of Pb, Zn and Cd Contamination in Near-Surface Soils of the Urban-Mining Transitional Region of Isfahan, Iran. *Pedosphere*, **20**, 568-577. http://dx.doi.org/10.1016/S1002-0160(10)60046-X

[48] Ettler, V., Mihaljevič, M., Kříbek, B., Majer, V. and Šebek, O. (2011) Tracing the Spatial Distribution and Mobility of Metal/Metalloid Contaminants in Oxisols in the Vicinity of the Nkana Copper Smelter, Copperbelt Province, Zambia. *Geoderma*, **164**, 73-84. http://dx.doi.org/10.1016/j.geoderma.2011.05.014

[49] Nguyen Ngoc, M., Dultz, S. and Kasbohm, J. (2009) Simulation of Retention and Transport of Copper, Lead and Zinc in a Paddy Soil of the Red River Delta, Vietnam. *Agriculture, Ecosystems & Environment*, **129**, 8-16. http://dx.doi.org/10.1016/j.agee.2008.06.008

[50] Hernandez, L., Probst, A., Probst, J.L. and Ulrich, E. (2003) Heavy Metal Distribution in Some French Forest Soils: Evidence for Atmospheric Contamination. *Science of the Total Environment*, **312**, 195-219. http://dx.doi.org/10.1016/S0048-9697(03)00223-7

[51] Broos, M.J., Aarts, L., Van Tooren, C.F. and Stein, A. (1999) Quantification of the Effects of Spatially Varying Environmental Contaminants into a Cost Model for Soil Remediation. *Journal of Environmental Management*, **56**, 133-145. http://dx.doi.org/10.1006/jema.1999.0271

[52] Lin, Y.-P., Yeh, M.-S., Deng, D.-P. and Wang, Y.-C. (2008) Geostatistical Approaches and Optimal Additional Sampling Schemes for Spatial Patterns and Future Sampling of Bird Diversity. *Global Ecology and Biogeography*, **17**, 175-188. http://dx.doi.org/10.1111/j.1466-8238.2007.00352.x

[53] ArcGIS Resource Center. Understanding a Semivariogram: The Range, Sill, and Nugget. http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/Understanding_a_semivariogram_The_range_sill_and_nugget/0031000000mq000000/

[54] Geoff, B. (2005) Introduction to Geostatistics and Variogram Analysis. http://people.ku.edu/~gbohling/cpe940/Variograms.pdf

[55] Maringanti, C., Chaubey, I., Arabi, M. and Engel, B. (2011) Application of a Multi-Objective Optimization Method to Provide Least Cost Alternatives for NPS Pollution Control. *Environmental Management*, **48**, 448-461. http://dx.doi.org/10.1007/s00267-011-9696-2

[56] Kaini, P., Artita, K. and Nicklow, J.W. (2012) Optimizing Structural Best Management Practices Using SWAT and Genetic Algorithm to Improve Water Quality Goals. *Water Resources Management*, **26**, 1827-1845. http://dx.doi.org/10.1007/s11269-012-9989-0

[57] Kitanidis, P.K. (1997) Introdcution to Geostatistics: Application to Hydrogeology.