International Journal of Innovative Computing, Information and Control
Volume **9**, Number **8**, August **2013**

ICIC International ©2013 ISSN 1349-4198
pp. 3197–3209

ESTIMATING KRIGING-BASED PREDICTIONS WITH PRIVACY

Bulent Tugrul[1] and Huseyin Polat[2]

[1]Department of Computer Engineering
Ankara University
No. 195, Fatih Street, Kecioren, Ankara, Turkey
btugrul@eng.ankara.edu.tr

[2]Department of Computer Engineering
Anadolu University
2 Eylul Campus, Eskisehir, Turkey
polath@anadolu.edu.tr

Received June 2012; revised October 2012

Abstract. *Kriging is a well-known prediction method. It interpolates the value of an unmeasured location from nearby measured locations. In a traditional Kriging interpolation, a client (an entity that is looking for a prediction for a specific location) asks help from a server (an entity that holds enough measurements collected for Kriging interpolations in a region). Predictions are estimated based on location data and measurements, which are considered confidential data. Neither the client nor the server wants to reveal their private data to each other. Although Kriging is increasingly becoming popular and widely used for estimating predictions, it fails to protect confidentiality. Thus, clients and servers might hesitate to participate in Kriging interpolations. In this study, we investigate how to provide Kriging-based predictions without violating data owners' privacy. We propose a scheme, which helps the clients and the servers perform Kriging interpolations while protecting their confidentiality. In other words, our method does not allow them from deriving information about each other's private data. We show that the proposed scheme protects privacy and it does not cause any accuracy losses. We also analyze it with respect to inevitable additional costs, which do not affect online performance. Our analyses show that the proposed scheme is able to provide accurate predictions efficiently while preserving privacy.*

Keywords: Kriging, Geo-statistics, Confidentiality, Accuracy, Performance, Prediction

1. Introduction.

Collecting data and providing useful outcomes from such data after mining are very common. There are various data mining and statistical functionalities utilized to extract meaningful outcomes from collected data. Estimating predictions from known measurements is among such functionalities and it is receiving increasing attention. In order to provide predictions, data collected for recommendation purposes are used together with a prediction algorithm. Kriging is an interpolation method, which is widely employed to estimate the value of an unmeasured location from known measurements observed at nearby locations [1]. In addition to inverse distance weighted (IDW) interpolation, Kriging is widely used interpolation technique for obtaining meta-models [2]. Although there are different types of Kriging models, the most popularly used one is referred to as ordinary Kriging.

Kriging interpolation might be utilized in various areas. Examples of such application areas are including but not limited to mine reservoirs, petroleum industry, environmental sciences, agriculture, and so on [3]. In addition to them, Kriging can be used to generate topographic maps and estimate air pollution. Kriging has been very popular since the work conducted by Matheron [4], where it is assumed that the closer points have more effect on unknown measurement. To put it in another way, effects of known measurements are inversely proportional with distance in both IDW and Kriging interpolations. Kriging consists of two major steps: creating a semi-variogram model from collected measurements and making prediction for unobserved location.

In a traditional Kriging interpolation, one party, referred to as the server (S), collects measurements (P_j values) for some measured locations (G sample points, where $j = 1, 2, \ldots, G$) in a given region (R). After observing such measurements, S investigates such P_j values to generate a model. It then starts providing predictions based on such model. Another party, referred to as the client (C), might seek a prediction (P_q value) for a specific location q (unmeasured location q). C sends a query (Q) including the coordinates of the unmeasured location q (x_q, y_q) to S. After receiving the coordinates, S estimates P_q using Kriging model based on its observed measurements in the nearby locations of the unmeasured location. It finally returns the estimated prediction P_q to the client C. The process depicted in Figure 1 can be summarized, as follows:

1) The client C sends the query Q to the server S including (x_q, y_q) coordinate values for the unmeasured location q.
2) The server S utilizes the model generated previously from known measurements (P_j values) for some locations $j = 1, 2, \ldots, G$ to estimate the prediction P_q.
3) It finally returns P_q (prediction for unmeasured location q) to the client C.

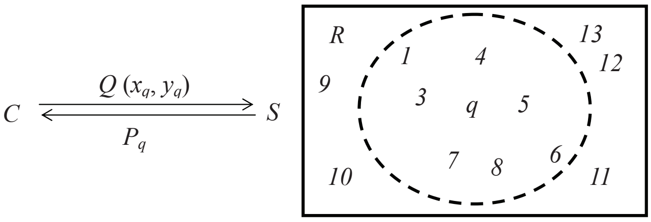


Figure 1. An example of traditional Kriging interpolation

As seen from Figure 1, in a given region R, the server S estimates the prediction P_q value for unmeasured location q based on the known measurements for $G = 8$ nearby locations of q using the Kriging interpolation, which is explained in Section 3. Notice that each location j is represented with coordinate values (x_j, y_j) and the related measurement P_j.

Providing various data mining and statistical services with privacy concerns is increasingly becoming popular [5, 6, 7, 8]. With increasing popularity of confidentiality, performing various functionalities while hiding private data has been taken increasing attention. Without any privacy concern, it is an easy job to perform different services based on available data. However, it is challenging to offer the same services while reserving confidentiality. Likewise, although providing Kriging-based predictions is comparable easy without privacy concerns, it becomes a difficult task to do Kriging interpolation without violating data owners' privacy. The reason for this phenomenon can be explained the conflicting nature of accuracy and privacy.

Data or measurements collected for Kriging interpolations and their related coordinates are considered the server S' confidential data. Hence, it does not want to disclose its private data. S utilizes such measurements in order to provide predictions in return of some benefits. Since it spends considerable efforts for observing such measurements including budget, labor, and so on, it wants to make money or at least compensate for what it has spent. Thus, such collected measurements are also considered its valuable assets. In case of data disclosure, it might lose competitive edge over other rival companies. Due to these privacy and financial concerns, S wants to hide its confidential data while performing

Kriging interpolations. Like the server S, the client C also has concerns about her privacy. The location for which C is looking for a prediction and the estimated prediction for that location are considered confidential. Based on the outcome of the Kriging interpolation, C plans investments. Thus, she does not want to reveal the unmeasured location and the related estimated prediction to S. The problem is *how to estimate Kriging-based predictions without disclosing the server's and the client's confidential data to each other.*

Without privacy-preserving measures, S and C do not feel comfortable. If we provide a scheme preserving their privacy, they feel more comfortable to involve Kriging interpolations. Therefore, we propose a scheme, which helps S and the C perform Kriging interpolations while preserving their confidentiality. Our method keeps confidential data private and it is able to provide predictions with decent accuracy. The purpose of our scheme is (1) to offer Kriging-based predictions with decent accuracy, (2) to estimate them efficiently, and (3) to provide them while preserving privacy.

The contributions of our study can be summarized, as follows:

1) We define the problem of estimating Kriging-based predictions with privacy, as briefly described previously.
2) We propose a naïve and an enhanced solution to this problem. Our enhanced scheme helps S and C perform Kriging interpolations without disclosing their private data to each other. To the best of our knowledge, Kriging has not been investigated in the literature with respect to preserving privacy. Our work happens to be the first one studying Kriging interpolation with privacy.
3) We analyze our scheme in terms of accuracy, privacy, and performance. Our analysis shows that the proposed method does not cause any accuracy losses due to privacy measures (it is able to achieve the same accuracy level as the one without privacy concerns), protects data owners' confidential data against each other, and it is able to provide predictions efficiently.
4) Due to privacy and financial concerns, S and C might hesitate to involve in Kriging interpolation services. Our scheme helps them feel more comfortable to join such interpolations. Any party with privacy concerns can use our scheme.

2. Related Work.

The study conducted by Tobler [9] has initiated the studies related to geo-statistics. Since then geo-statistics has been receiving increasing attention. Geo-statistics interpolation methods can be grouped as deterministic and geo-statistical methods [10]. IDW is a widely used deterministic method while Kriging is the main tool used as a geo-statistical method [11]. Hence, in order to estimate predictions for unmeasured locations, Kriging and IDW interpolations are widely used. Kriging is used in various application areas. Krige [1], who developed Kriging, proposes to use Kriging in order to predict the ore reserves. Armstrong [3] explains the application of geo-statistics in mine reservoirs to calculate capacity of mine reservoir and errors. Kriging is also used to predict air pollution [12]. Shad et al. [12] utilize Kriging for air pollution prediction, where the authors employ a genetic algorithm to optimize membership functions to improve accuracy. Kriging-based techniques are used to predict and analyze soil properties by Sun et al. [13]. The authors perform some experiments and demonstrate that their approach provides highly accurate outcomes for some specific cases. They also develop a software program to perform local regression Kriging automatically. In addition to analyzing soil properties and air pollution prediction using Kriging, Kriging is also utilized to estimate soil contamination [14]. Largueche [14] investigates whether Kriging is a useful tool to estimate the spatial distribution of ground pollutants in contaminated land. The author also discusses the identification of areas that should be subjected to remedial actions. Kaymaz [15] proposes to apply Kriging to structural reliability problems. The author investigates the use of Kriging for such problems and compares it with response surface method. Ali et al. [16] apply Kriging to the spatial interpolation of local disease rates. Their approach helps researchers incorporate the pattern of spatial dependence into the mapping of risk values.

Confidential information is becoming more important with the spread of data mining methods. In addition, some laws force companies to keep their data secret. Privacy-preserving and secure multi-party computation methods give us opportunities to conduct data mining methods without revealing information to other parties. Agrawal and Srikant [17] propose randomized data perturbation methods to hide sensitive information. The authors show that accurate predictive models can be created from a large number of perturbed data items. Evfimievski [18] discusses perturbation levels against privacy levels and presents some methods to measure privacy. Li and Sarkar [19] propose a perturbation method for categorical data to prevent disclosure of private data. Their scheme is based on two steps consisting of linear programming and swapping. In [20], the authors propose geometric data perturbation for preserving confidential data and discuss different aspects of such method. Li and Wang [5] propose a classification method based on singular value decomposition with privacy. Meskine and Bahloul [6] study and analyze privacy-preserving k-means algorithms and classify them based on data distribution, where they discuss advantages and disadvantages of each proposed protocol.

Privacy-preserving data mining methods on vertically or horizontally distributed data have been widely studied [18, 21, 22, 23, 25, 26]. The authors in [18, 21, 22, 23, 25, 26] study performing various data mining functionalities like association rule mining, clustering, outlier detection, and scoring on distributed data while preserving privacy.

Performing statistical analysis with privacy is also becoming popular. Xiao and Tao [8] propose dynamic anonymization to construct privacy-preserving statistical database. Du and Atallah [27] develop protocols to perform statistical analysis in cooperative environments. Drosatos and Efraimidis [28] propose a privacy-preserving scheme to analyze ubiquitous health data. They show that it is possible to conduct distributed statistical analysis on health data with privacy. In [29], it is discussed how to estimate statistical information in financial and commercial systems while keeping individual value private. Yao et al. [30] propose secure protocols for estimating harmonic and geometric mean and mode. They consider applications of secure multi-party computation in statistics.

Providing predictions while preserving privacy is also receiving increasing attention. Polat and Du [7] show how to estimate predictions using collaborative filtering without jeopardizing customers' confidentiality. Sakuma and Arai [31] discuss how to conduct online prediction from expert observes while preserving privacy. In [32], the authors provide a method to protect customers' privacy in churn prediction. They mask users' data and conduct churn prediction on perturbed data.

Our work is different from the above mentioned ones with some respects. First of all, to the best of our knowledge, our study is the first one describing the problem of Kriging interpolation with privacy. Second, we are the first one investigating how to offer Kriging-based predictions while preserving confidentiality. Finally, although there are various studies concerning Kriging or preserving privacy while conducting various data mining tasks, there is no one, which studies Kriging with privacy.

3. Background: Kriging-based Interpolation.

As described previously, in a traditional Kriging-based interpolation, there are two parties (S and C). S owns measurements (P_j values) for some G sample locations with their related coordinates in a given region R. C asks a prediction for some location q from S. The steps of such interpolation are given, as follows:

1) C sends the coordinates of q (x_q, y_q) in R for which she is looking for a prediction to S.
2) S first computes distances between measured locations in R using Euclidean distance measure. Given two measured locations, i and j, the distance between them (d_{ij}) can be calculated, as follows:
$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{1}$$
3) Then, S calculates semi variances (V values) between measured locations, i and j, as follows:
$$V_{ij} = 0.5 \times [P_i - P_j]^2. \tag{2}$$
4) S then groups sample points using binning and finds average semi variances and distances for each bin.
5) Next, S plots average semi variances versus average distances; and finds the formula to estimate semi variance at any given distance. Semi variances can be represented, as follows:
$$Semi\ \ variance = f(distance), \tag{3}$$
where f is a function representing the relationship between semi variances and distances. The relationship is usually linear and estimated semi variances can be found by multiplying the observed ones with distances.
6) S creates $\mathbf{\Gamma}$ matrix, which is an $(G+1) \times (G+1)$ symmetric matrix including the estimated semi variances between any two locations using Equation (3). Notice that the row and correspondingly the last column are filled with 1s, except the diagonal entry, which is set at 0.
7) Then, S finds $\mathbf{\Gamma^{-1}}$ matrix, which is again $(G+1) \times (G+1)$ symmetric matrix including γ values.
8) Next, S computes distances between q and each measured location using Equation (1). It then creates matrix $\mathbf{g}$, which is an $(G+1) \times 1$ matrix including the semi variances estimated between q and each measured location using Equation (3).
9) S then solves the Kriging weights (λ matrix), as follows:
$$\lambda = \mathbf{\Gamma^{-1}} \times \mathbf{g} \tag{4}$$
in which λ is a $(G+1) \times 1$ matrix.
10) Finally, S estimates the final prediction P_q for unmeasured location by multiplying the weight for each measured location and the related measure or value; and adds them together. If we consider λ (remove the last weight representing the weight for the unmeasured location) and $\mathbf{P}$ (including known measurements or values for G locations) as vectors of length G, then P_q can be estimated by finding the scalar product of λ and $\mathbf{P}$, as follows:
$$P_q = \lambda \cdot \mathbf{P} = \sum_{i=1}^{G} \lambda_i \times P_i. \tag{5}$$

4. **Kriging-based Interpolation Schemes with Privacy.** Although Kriging is widely used in many applications for prediction purposes, it fails to protect private data. Due to privacy risks, participating parties might not feel comfortable and they may decide not to involve in Kriging interpolation. Hence, we propose a scheme allowing the servers and the clients to perform Kriging without divulging their confidential data to each other.

4.1. **Confidential data.** The aim of the privacy-preserving data mining schemes is to protect private data. Hence, before we present our proposed scheme, we first need to determine confidential data that should be protected against involving parties. Private data items can be categorized, as follows:

1) Confidential data held by the server S
   (a) *Coordinate values of the sample locations* $((x_j, y_j)$ *values*$)$
   (b) *Observed measurements* $(P_j$ *values*$)$
2) Confidential data held by the client C
   (a) *Coordinate values of the location* q $((x_q, y_q))$
   (b) *The estimated prediction* (P_q)

4.2. **Problem definition.** On one hand, the server S holds measurements together with their related coordinates and wants to provide Kriging interpolation services in return of some benefits. On the other hand, the client C wants to obtain an estimated prediction for a specific location without making measurements. Both S and C do not want to reveal their confidential data to each other due to privacy concerns. Thus, the problem is *how these two parties perform Kriging interpolation without divulging their confidential data to each other. How does* S *provide accurate predictions efficiently without violating its privacy and* C*'s confidentiality.*

4.3. **Proposed schemes.** As explained previously, S first needs to create a model (formula for determining semi variances – Equation (3) and $\mathbf{\Gamma^{-1}}$ matrix) using its data for a given region R in order to estimate a prediction for any unmeasured location q. However, it needs q's coordinates to generate matrix $\mathbf{g}$ so that it can estimate matrix λ and determine P_q. In the following, we describe two schemes assuming that S has already created the model given R. In other words, we explain how S creates $\mathbf{g}$, estimates λ, and determines P_q without jeopardizing privacy constraints. In the following, we describe our proposed naïve solution and then explain our enhanced method in detail.

4.3.1. *First solution – Naïve scheme.* Our proposed schemes' major concern is to protect confidential data of involving parties. Therefore, during Kriging-based interpolation process, sample locations (their coordinates) and their related measurements and unmeasured location q (its coordinates) and the estimated prediction P_q should not be disclosed to C and S, respectively. We propose the following naïve scheme in order to estimate Kriging-based predictions without jeopardizing data owners' privacy. Our naïve scheme is based on randomness (creating bogus locations) and 1 out of n oblivious transfer (OT) protocol. OT refers to a protocol, where at the beginning of the protocol one party, S has n inputs I_1, I_2, …, I_n and at the end of the protocol the other party, C, learns one of the inputs I_i for some $1 \leq i \leq n$ of her choice, without learning anything about the other inputs and without allowing S to learn anything about i. An efficient OT protocol is proposed by Naor and Pinkas [33], where it could be achieved with poly-logarithmic (in n) communication time. The steps of our naïve method can be listed, as follows:

1) S first creates a model (formula for determining semi variances or Equation (3) and $\mathbf{\Gamma^{-1}}$ matrix) using its data for the given region R.
2) C generates $n-1$ bogus locations in order to mask her real location.
3) She hides the unmeasured location q among such fake locations; and sends the coordinates of n locations including q to S.
4) For each location $j = 1, 2, \ldots, n$; S performs the following steps:
   (a) First, it estimates distances between j and each measured location using Equation (1).
   (b) It then computes semi variances using Equation (3).

   (c) Next, it creates **g** matrix.
   (d) It then estimates the weights using Equation (4).
   (e) Finally, it computes P_j using Equation (5).

5) After estimating predictions for all n locations, C utilizes OT protocol in order to get the prediction for her real location q only. Due to OT protocol, which is shown to be secure [33], S cannot learn which prediction is obtained by C; and C cannot know other predictions rather than P_q. OT protocol allows C learns one of the n inputs (P_q) held by S without learning anything about the other inputs and without allowing S to learn anything about q.

Due to bogus locations, S cannot learn the real location q. However, it can guess it with probability of $1/n$ because there are n possibilities. With increasing n, such probability becomes smaller. Similarly, for S, the probability of guessing the estimated prediction is $1/n$ because it estimates predictions for n locations and one of them is for the real location. S does not want any client obtains more than one prediction during a single process due to financial reasons. Service suppliers provide estimated predictions in return of some benefits. To prevent C from receiving predictions for more than one location, OT protocol is utilized. OT protocol forces C to get estimated prediction for her real location only; and at the same time, it prevents S from learning which prediction is obtained by C. Due to aggregate outcome (estimated prediction), C cannot derive useful information about locations and their measurements held by S from received prediction.

In addition to our naïve scheme, we also propose the following scheme, referred to as improved scheme (IS). Details of our IS are described in the following.

4.3.2. *Second solution – Improved scheme (IS).* As explained previously, given R, S first can create a model using its data. It then needs to estimate distances between q and each sample point it holds in the region R. After that it is supposed to estimate semi variances in order to create **g** matrix. It finally needs to estimate P_q. Our second scheme is based on homomorphic encryption (HE). We utilize the HE scheme proposed by Paillier [34] to hide confidential data. If we assume that ξ is an encryption function and K is a public key, and x_{j1} and x_{j2} are private data values, then Paillier's HE scheme allows us to compute $\xi_K(X) = \prod_{j=1}^{n}(\xi_K(x_{j1}))^{x_{j2}}$ values. The steps of our IS are, as follows:

1) The first step is calculating distances between q and each sample location. Such distances between q and each location $j = 1, 2, \ldots, G$ can be computed using Equation (1) while preserving confidentiality, as follows:
   (a) Equation (1) can be written, as follows:

$$d_{jq} = \sqrt{{x_j}^2 + {y_j}^2 + {x_q}^2 + {y_q}^2 - 2 \times (x_j x_q + y_j y_q)} = \sqrt{S_j + C_q - 2x_j x_q - 2y_j y_q}. \quad (6)$$

   As seen from Equation (6), S and C can compute S_j and C_q, respectively without needing each other. However, to estimate $x_j x_q$ and $y_j y_q$ values, they need to collaborate.
   (b) Using HE scheme, C finds $\xi_{KC}(-2x_q)$, $\xi_{KC}(-2y_q)$, and $\xi_{KC}({x_q}^2 + {y_q}^2)$ encrypted values, where ξ represents encryption function and KC is C's public key.
   (c) She then sends such encrypted values to S. Since the related private key is known by C only, S cannot learn x_q and y_q values.
   (d) For each location $j = 1, 2, \ldots, G$, using HE scheme, S determines $\xi_{KC}(D_{jq}) = \xi_{KC}(-2x_q)^{x_j} \times \xi_{KC}(-2y_q)^{y_j} \times \xi_{KC}({x_q}^2 + {y_q}^2)^1 \times \xi_{KC}({x_j}^2 + {y_j}^2)^1 = \xi_{KC}({x_j}^2 + {y_j}^2 + {x_q}^2 + {y_q}^2 - 2x_j x_q - 2y_j y_q)$.
   (e) To find distances, square root of such encrypted values must be computed. Thus, S then sends $\xi_{KC}(D_{jq})$ values for all $j = 1, 2, \ldots, G$ to C.

(f) C then decrypts $\xi_{KC}(D_{jq})$ values using the related private key, obtains D_{jq} values; and finds d_{jq} distance values between q and each location j by taking the square roots of D_{jq} values. Since D_{jq} values are aggregate values, C cannot learn the related x_j and y_j values from them. Even if she knows the distances, she cannot determine the true coordinates. She only learns that any location j is on the circle whose center is q and radius is d_{jq}.

(g) Using HE scheme, C encrypts d_{jq} values using her public key KC and sends $\xi_{KC}(D_{jq})$ to S.

2) S finds estimated semi variances in encrypted form for location q using Equation (3) and HE property; and creates the g matrix including the encrypted values, $\xi_{KC}(g_j)$ values for all $j = 1, 2, \ldots, G$.

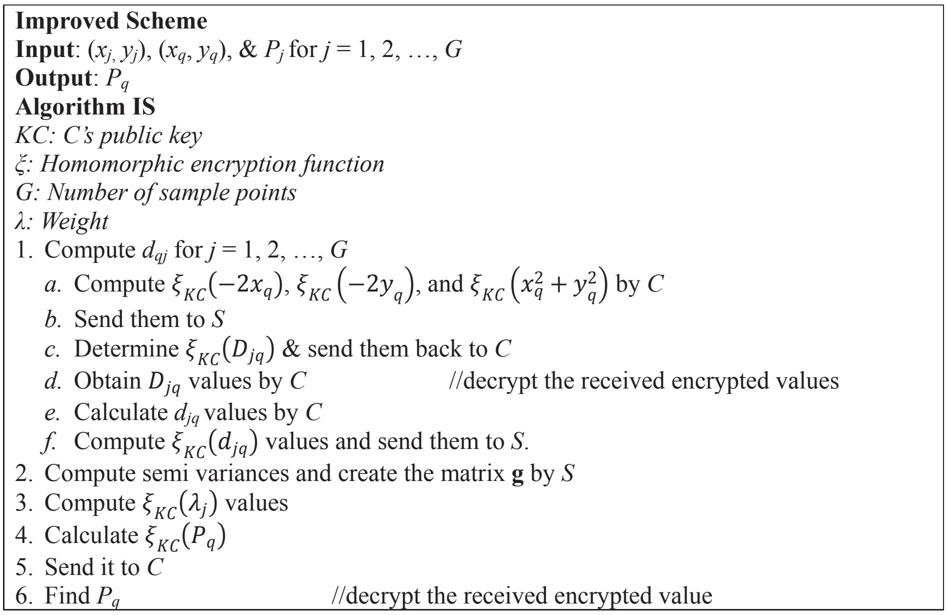
3) Now, S needs to compute weights or λ values using Equation (4). Using HE scheme, S computes $\xi_{KC}(\lambda_j) = \xi_{KC}(g_1)^{\gamma_{j1}} \times \xi_{KC}(g_2)^{\gamma_{j2}} \times \ldots \times \xi_{KC}(g_G)^{\gamma_{jG}} = \xi_{KC}(g_1 \times \gamma_{j1} + g_2 \times \gamma_{j2} + \ldots + g_G \times \gamma_{jG})$.

4) S then can estimate the prediction P_q using Equation (5), as follows: $\xi_{KC}(P_q) = \xi_{KC}(\lambda_1)^{P_1} \times \xi_{KC}(\lambda_2)^{P_2} \times \ldots \times \xi_{KC}(\lambda_G)^{P_G} = \xi_{KC}(\lambda_1 \times P_1 + \lambda_2 \times P_2 + \ldots + \lambda_G \times P_G)$.

5) Finally, S sends $\xi_{KC}(P_q)$ to C. Due to encryption, S cannot know the estimated prediction.

6) Since C knows the related decryption key, she decrypts the received encrypted value and gets P_q. Due to aggregate estimation; C cannot derive information about S's confidential data.

To recapitulate our enhanced scheme, we give the pseudo-code of the IS in Figure 2.



```
Improved Scheme
Input: $(x_j, y_j)$, $(x_q, y_q)$, & $P_j$ for $j = 1, 2, …, G$
Output: $P_q$
Algorithm IS
KC: C's public key
$\xi$: Homomorphic encryption function
G: Number of sample points
$\lambda$: Weight
1. Compute $d_{qj}$ for $j = 1, 2, …, G$
    a. Compute $\xi_{KC}(-2x_q)$, $\xi_{KC}(-2y_q)$, and $\xi_{KC}(x_q^2 + y_q^2)$ by C
    b. Send them to S
    c. Determine $\xi_{KC}(D_{jq})$ & send them back to C
    d. Obtain $D_{jq}$ values by C                //decrypt the received encrypted values
    e. Calculate $d_{jq}$ values by C
    f. Compute $\xi_{KC}(d_{jq})$ values and send them to S.
2. Compute semi variances and create the matrix g by S
3. Compute $\xi_{KC}(\lambda_j)$ values
4. Calculate $\xi_{KC}(P_q)$
5. Send it to C
6. Find $P_q$                  //decrypt the received encrypted value
```

Figure 2. Pseudo-code of the proposed improved scheme

5. Analysis of the Improved Scheme.

There are basically two evaluation criteria for prediction algorithms. They are called performance and accuracy. Performance means that how effectively a prediction algorithm can estimate predictions. It can be measured with respect to off-line and online costs like storage, computation, and communication (number of communications and amount of transferred data) costs. Hence, performance analysis can be done in terms of off-line and online costs. Compared with online costs, off-line costs are not that critical for overall performance. Online efficiency requirements differ for various applications. For some applications, there are very hard online performance requirements. For example, recommender systems should be able to return many recommendations to their customers simultaneously in a very short time during an online interaction. However, performance requirements might be soft for some applications like geo-statistics. Online time limitations are not that rigid in geo-statistical predictions. For example, if one petroleum company looks for oil reserves in a given region, it might ask prediction from those that owns enough measurements in that region. Since investments in energy take some time and considerable amount of budgets, oil companies spend some time to get reliable and accurate predictions. Obtaining dependable and precise predictions is much more important than receiving predictions in a short time. Therefore, it can be said that online performance constraints are soft in Kriging-based interpolations. Performance criterion covers the time needed to perform a single prediction, number of communications spent for a prediction (and/or amount of transferred data), and amount of storage space is needed. Resources spent for interpolations should be minimized for performance reasons.

The second criterion accuracy is related to how accurate the estimated predictions are. Accuracy is measured in terms of the closeness between the estimated predictions and their true values. Estimated interpolations should be as close as to their observed values. Since predictions are estimated values based on available observed measurements, their values should be as close as possible to their expected values. Therefore, predictions generated by our proposed scheme with privacy concerns should be as close as possible their true values.

In addition to performance and accuracy, privacy is another evaluation metric, which is used to investigate privacy-preserving prediction schemes. Privacy-preserving algorithms should be able to protect confidential data. Privacy requirements state that involving parties in interpolation processes cannot derive useful information about each other's private data. Thus, privacy, in this context, means that confidential data should be hidden to those but the intended parties. In other words, our proposed privacy-preserving scheme should be able to hide confidential data held by S and C against each other.

We analyze our enhanced method in terms of privacy, accuracy, and efficiency. We basically analyze additional costs due to privacy concerns even though online performance requirements are not that rigid. We also concern with accuracy losses due to privacy-preserving measures because accuracy and privacy are conflicting goals. Finally, we want to show that our scheme does not violate privacy constraints.

5.1. **Accuracy analysis.** In privacy-preserving prediction schemes, privacy measures usually make accuracy worse due to the conflicting nature of confidentiality and preciseness. However, in our scheme, privacy-preserving methods do not cause any loss in accuracy. In other words, *predictions estimated by our method with privacy concerns are the same as the ones provided by traditional Kriging scheme without confidentiality fears.* Since we utilize cryptographic techniques preserving data originality, accuracy is not affected. Thus, our scheme is able to provide the same predictions while preserving confidentiality.

5.2. **Performance analysis.** We investigate our scheme in terms of supplementary costs due to privacy concerns. We first analyze additional storage costs. Our scheme does not cause any extra storage costs. Involving parties (S and C) do not need additional spaces required to save data caused by confidentiality measures. Thus, storage costs will not be affected by privacy concerns.

As shown in Figure 1, in a traditional Kriging-based prediction process, number of communications is *two* only because C and S communicate two times only. However, number of communication in our scheme increases due to privacy measures. As described in Figure 2, number of communications is *four* in our scheme. In other words, number of communications increases *two* times due to our proposed scheme. Amount of transferred data is also important. In a conventional interpolation, C sends coordinates of location q and S returns a prediction. If we assume that four bytes are needed to save a coordinate and four bytes are enough to store an estimated prediction, then amount of sent data from C to S is about *eight* bytes while it is about *four* bytes from S to C in a traditional scheme. In our scheme, C first sends three encrypted values to S. The size of an encrypted value is imperative. As explained in [35], the size of an encrypted value produced by block cipher encryption can be computed as *size of plain text* + *block size* − (*size of plain text* **mod** *block size*). For example, if we assume that size of plain text is four bytes, block size is 16 bytes, and then we need 16 bytes for an encrypted value. Thus, amount of sent data during this communication is about 48 bytes. After computing G encrypted aggregates, S then sends them to C. Assuming again that 16 bytes are needed for a single encrypted value; amount of sent data is about $16G$ bytes. During the second turn, C sends G encrypted values to S. Thus, amount of transferred data is again $16G$ bytes. And finally, S returns an encrypted value to C. Hence, amount of sent data is about 16 bytes. To sum up, like number of communications, amount of transferred data also increases due to privacy measures.

Supplementary computation costs caused by our scheme are also inevitable. In addition to multiplications and additions, our method includes encryptions, decryptions, and exponentiations because of privacy measures. Number of encryptions is in the order of $O(G)$ and similarly, number of decryptions is in the order of $O(G)$. On the other hand, number of exponentiations is in the order of $O(G^2)$. Notice that G is a constant representing number of measured locations in the region R. Cryptographic functions are usually costly operations. In order to find out the running times of cryptographic operations, benchmarks for the CRYPTO++ toolkit from http://www.cryptopp.com/ can be used [36].

Although our scheme does cause some extra communication (in terms of number of communications and amount of transferred data) and computation costs, they are not critical due to the nature of Kriging-based interpolation schemes. Unlike some real time applications, online performance requirements are softer for Kriging-based prediction methods.

5.3. **Privacy analysis.** Our privacy requirements state that private data values should not be disclosed during prediction process. Notice that the measured locations (or their coordinates) and the related measurements are confidential for S. Similarly, the unmeasured location (or its coordinates) and the estimated prediction are private for C. The parties cannot learn each other's confidential data during our scheme. C sends her location coordinates in encrypted form rather than plain form. Since the related decryption key is known by C only, S cannot decrypt the received values and learn coordinates. After performing required computations using HE scheme, S sends encrypted aggregates to C. Since C knows decryption key, she can decrypt the received values and find distances between q and each measured location. Although C learns the distances, she cannot learn the true coordinates. For each measured point j, given q and the related distance d_{jq}, the only information that C can derive is that j is in somewhere on the circle whose center is q and its radius is d_{jq}.

S returns the estimated prediction in encrypted form. Since the decryption key is known by C only, S cannot decrypt it and learn the prediction. When C obtains the prediction, which is an aggregate value, she cannot learn the measurements of G sample points. Paillier [34] shows that HE is semantically secure for inference of input values. In other words, the parties cannot derive any information from the exchanged encrypted values. Our method prevents C from learning the measured location coordinates and their related measurements. It also prevents S from deriving useful information about the unmeasured location coordinates and the estimated prediction.

6. Conclusions and Future Work.

Kriging interpolation is popularly used in many applications. However, it fails to preserve data privacy. In this study, we basically achieved the followings:

1) We defined the problem of Kriging interpolation with privacy.
2) We proposed two solutions (naïve and enhanced schemes) to this problem.
3) We analyzed the enhanced method with respect to accuracy, performance, and privacy.

Our study is the first one investigating how to offer Kriging-based predictions while preserving privacy. The followings can be derived from our analysis of the improved scheme:

1) Since we used cryptographic techniques (which happen to preserve the originality of the confidential data) in order to hide private data, our scheme did not cause any accuracy losses.
2) Although our scheme caused additional costs, they are not critical for the online performance, because online performance requirements are not rigid in Kriging interpolation.
3) Our scheme did preserve data privacy because we used cryptographic methods, which were proved to be secure.

Kriging has many applications in disciples like mining, remote sensing, hydrogeology, environmental sciences and natural resources. In these disciples, privacy is increasingly becoming a serious issue. With increasing popularity of preserving privacy, privacy-preserving data mining and statistical analysis are also receiving increasing attention. Hence, our scheme can be used by those providing Kriging interpolation to some clients in the above mentioned applications. Our method helps both the servers and the clients hide their private data against each other while still allowing them to perform Kriging interpolations. The proposed scheme helps them overcome privacy concerns so that they can participate in Kriging-based predictions.

Kriging and inverse distance weighted interpolations are two popularly used geo-statistical techniques. Since our study is the first one discussing Kriging with privacy problem, there is no contemporary scheme like inverse distance weighted interpolation with privacy in order to compare our method with the existing ones. Therefore, we are planning to study how to provide inverse distance weighted-based predictions while preserving privacy. We will also want to compare such a scheme with the one proposed in this paper with respect to accuracy, performance, and privacy. We are planning to expand our study to conduct Kriging in two or multi-party scenarios in which it is assumed that data are distributed among various parties. Although we stated that additional costs are not critical, we will study how to improve overall performance. In some applications, supplementary costs might become vital. Thus, such limitations should be eliminated.

REFERENCES

[1] D. G. Krige, A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, vol.52, no.6, pp.119-139, 1951.

[2] V. R. Joseph, Y. Hung and A. Sudjianto, Blind Kriging: A new method for developing metamodels, *ASME Journal of Mechanical Design*, vol.130, no.3, 2008.

[3] M. Armstrong, *Basic Linear Geostatistics*, Springer, Berlin, 1998.

[4] G. Matheron, Principles of geostatistics, *Economic Geology*, vol.58, pp.1246-1266, 1963.

[5] G. Li and Y. Wang, A privacy-preserving classification method based on singular value decomposition, *The International Arab Journal of Information Technology*, vol.9, no.6, pp.529-534, 2012.

[6] F. Meskine and S. N. Bahloul, Privacy-preserving k-means clustering: A survey research, *The International Arab Journal of Information Technology*, vol.9, no.2, pp.194-201, 2012.

[7] H. Polat and W. Du, Privacy-preserving collaborative filtering, *International Journal of Electronic Commerce*, vol.9, no.4, pp.9-35, 2005.

[8] X. Xiao and Y. Tao, Dynamic anonymization: Accurate statistical analysis with privacy preservation, *Proc. of the ACM SIGMOD Conf. on Management of Data*, Vancouver, BC, Canada, pp.107-120, 2008.

[9] W. R. Tobler, Smooth pycnophylactic interpolation for geographical regions, *Journal of the American Statistical Association*, vol.74, no.367, pp.519-530, 1979.

[10] S. Ly, C. Charles and A. Degré, Geostatistical interpolation of daily rainfall at catchment scale: The use of several variogram models in the Ourthe and Anbleve catchments Belgium, *Hydrology and Earth System Sciences*, vol.15, pp.2259-2274, 2011.

[11] G. Y. Lu and D. W. Wong, An adaptive inverse-distance weighting spatial interpolation technique, *Computers and Geosciences*, vol.34, no.9, pp.1044-1055, 2008.

[12] R. Shad, M. S. Mesgari, A. Akbar and A. Shad, Predicting air pollution using fuzzy genetic linear membership Kriging in GIS, *Computers, Environment and Urban Systems*, vol.33, no.6, pp.472-481, 2009.

[13] W. Sun, B. Minasny and A. McBratney, Analysis and prediction of soil properties using local regression-kriging, *Geoderma*, vol.171-172, pp.16-23, 2011.

[14] F. Z. B. Largueche, Estimating soil contamination with Kriging interpolation method, *American Journal of Applied Sciences*, vol.3, no.6, pp.1894-1898, 2006.

[15] I. Kaymaz, Application of Kriging method to structural reliability problems, *Structural Safety*, vol.27, no.2, pp.133-151, 2005.

[16] M. Ali, P. Goovaerts, N. Nazia, M. Z. Haq, M. Yunus and M. Emch, Application of Poisson Kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh, *International Journal of Health Geographics*, vol.5, no.45, 2006.

[17] R. Agrawal and R. Srikant, Privacy preserving data mining, *Proc. of the ACM SIGMOD Conf. on Management of Data*, Dallas, TX, USA, pp.439-450, 2000.

[18] A. Evfimievski, Randomization in privacy-preserving data mining, *ACM SIGKDD Explorations Newsletter*, vol.4, no.2, pp.43-48, 2002.

[19] X. B. Li and S. Sarkar, Privacy protection in data mining: A perturbation approach for categorical data, *Information Systems Research*, vol.17, no.3, pp.254-270, 2006.

[20] K. Chen and L. Liu, Geometric data perturbation for privacy preserving outsourced data mining, *Knowledge and Information Systems*, vol.29, no.3, pp.657-695, 2011.

[21] J. Vaidya and C. Clifton, Privacy-preserving association rule mining in vertically partitioned data, *Proc. of the 8th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, pp.639-644, 2002.

[22] J. Vaidya and C. Clifton, Privacy-preserving k-means clustering over vertically partitioned data, *Proc. of the 9th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, Washington, DC, USA, pp.206-215, 2003.

[23] M. Kantarcioglu and C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data, *IEEE Transactions on Knowledge and Data Engineering*, vol.16, no.9, pp.1026-1037, 2004.

[24] A. Inan, S. V. Kaya, Y. Saygin, E. Savas, A. A. Hintoglu and A. Levi, Privacy-preserving clustering on horizontally partitioned data, *Data and Knowledge Engineering*, vol.63, no.3, pp.646-666, 2007.

[25] J. Vaidya and C. Clifton, Privacy-preserving kth element score over vertically partitioned data, *IEEE Transactions on Knowledge and Data Engineering*, vol.21, no.2, pp.253-258, 2009.

[26] L. T. Dung and H. T. Bao, A distributed solution for privacy-preserving outlier detection, *Proc. of the 3rd International Conference on Knowledge and Systems Engineering*, Washington, DC, USA, pp.26-31, 2011.

[27] W. Du and M. J. Atallah, Privacy-preserving cooperative statistical analysis, *Proc. of the 17th Annual Computer Security Applications Conference*, New Orleans, LA, USA, pp.102-110, 2001.

[28] G. Drosatos and P. S. Efraimidis, Privacy-preserving statistical analysis on ubiquitous health data, *Lecture Notes in Computer Science*, vol.6863, pp.24-36, 2011.

[29] A. Kiayias, B. Yener and M. Yung, Privacy-preserving information markets for computing statistical data, *Lecture Notes in Computer Science*, vol.5628, pp.32-50, 2009.

[30] Y. Yao, Y. Luo, L. Huang, W. Jing, W. Yang and W. Xu, Privacy-preserving technology and its applications in statistics measurements, *Proc. of the 2nd International Conf. on Scalable Information Systems*, Suzhou, China, pp.74-81, 2007.

[31] J. Sakuma and H. Arai, Online prediction with privacy, *Proc. of the 27th International Conf. on Machine Learning*, Haifa, Israel, pp.935-942, 2010.

[32] S. Xu, S. Lai and M. Qiu, Privacy-preserving churn prediction, *Proc. of the ACM Symposium on Applied Computing*, Honolulu, Hawaii, USA, pp.1610-1614, 2009.

[33] M. Naor and B. Pinkas, Oblivious transfer and polynomial evaluation, *Proc. of the 31st ACM Symposium on Theory of Computing*, Atlanta, GA, USA, pp.245-254, 1999.

[34] P. Paillier, Public key cryptosystems based on composite degree residuosity classes, *Lecture Notes in Computer Science*, vol.1592, pp.223-238, 1999.

[35] *www.obviex.com/Articles/CiphertextSize.aspx.*

[36] J. Canny, Collaborative filtering with privacy, *Proc. of the IEEE Symposium on Security and Privacy*, Oakland, CA, USA, pp.45-57, 2002.