PRIVACY-PRESERVING TREND SURFACE ANALYSIS

Salih Demir & Bulent Tugrul
Ankara University, Department of Computer Engineering, Golbasi, Ankara, Turkey

ABSTRACT

Trend Surface Analysis (TSA), which is a spatial interpolation method, is one of the essential instruments used by natural and environmental scientists to produce a prediction value for an unmeasured location. TSA explores the optimal polynomial surface that passes through sampled data. Data have become the most valuable asset of institutions. As a result of technological developments, the privacy of data has become more important. TSA was applied to so many problems without considering confidentiality of data in the literature. In traditional TSA scheme, there are two parties; client and server. The client requests a prediction value for a specific coordinate where it will spend its money and time for future investment. The server is the data owner which spent a significant amount of money and time to obtain such valuable asset. We propose a privacy-preserving solution to provide TSA-based spatial analysis without violating the confidentiality of both parties' data. We analyse our scheme in terms of privacy and performance to show that our solution provides accurate prediction model without violating privacy.

Key words: *Spatial interpolation, Trend surface analysis, Privacy, Prediction, Accuracy*.

1. INTRODUCTION

Data is seen as the new oil in the 21st Century. Companies and governmental organizations collect data from mobile phones, search queries and social media sites. Storing vast amount of data does not provide any useful information. Data should be processed to unveil useful and valuable information. Statistics and data mining methods are applied to reach such information.

Spatial interpolation has a crucial role in planning, risk assessment, and decision making. Collecting data for all points in a geographic region is impossible and expensive. Spatial interpolation methods are used to predict values for unmeasured locations. The First Law of Geography stated by Tobler [1] says that all things are related to each other but near things are more related than distant things. Spatial interpolation methods are used mainly in natural and environmental studies. Especially mine engineers prefer these methods to determine the optimal path of ore. In the first place Kriging [2] comes to mind as the most known spatial interpolation methods; however, Inverse Distance Weighting (IDW), TSA and many more are the other choices for engineers and researchers [3]. IDW [4] follows Tobler's principal. It produces predictions using near measurements around the unknown location. TSA is an interpolation method that basically tries to find the optimal polynomial surface that passes through sampled data points in a specified region. It is used to detect general tendencies of the sampled data. It has two options: global and local. The local methods determine the best surface using the only certain number of sampled data around unmeasured points. The global methods use entire sampled data to find the best surface for the interested region.

In traditional spatial interpolation scheme, there are two parties. The first party is called Servers (S) which store all sampled data in the specified region. Servers can build a prediction model based on TSA. The servers invest a huge amount of money and time to gather such valuable data. Therefore, they want to hide their data and prediction model from other servers and clients. The second party is called Clients (C). Clients need a prediction value for the specific location. A client might be interested in this specific location for its future investments. That is why it wants to hide the exact coordinate from the server and other clients. The traditional scheme is depicted in the figure below. However, this scheme does not provide any privacy measure. Privacy-preserving schemes for IDW and Kriging methods were studied [5, 6]. In addition to this, this study will propose a privacy-preserving scheme for TSA. In this way, confidential data possessed by servers and clients will be protected.

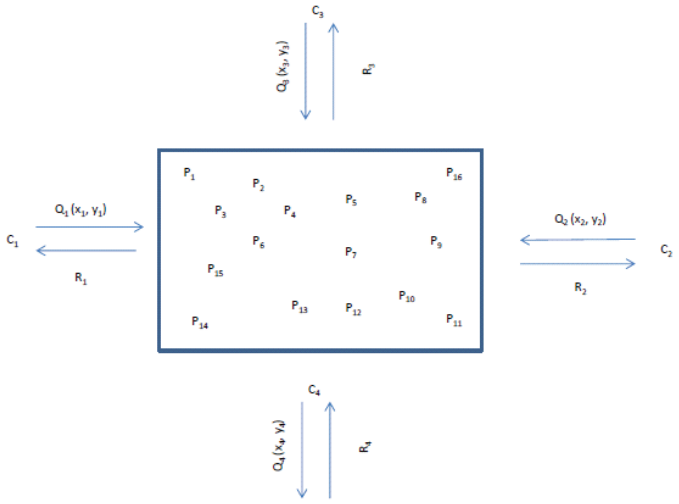

Figure 1 Traditional TSA scheme

The paper is organized as follows; Sect. 2 explains related studies in the literature. After extensively presenting the privacy-preserving solution in Sect. 3, we analyze our solution in terms of additional storage, communication and computation costs, privacy and accuracy in Sect. 4. Finally, there will be conclusions and future work section.

2. RELATED WORK

Data analysis and information extraction have become increasingly prevalent as a result of technological developments. Companies and governmental organizations have realized that they need to collaborate to reveal more accurate models. The data they have will determine their future positions. Therefore, steps taken for information security are becoming more vital. There are many studies in the literature for this purpose.

Data is the private property of companies. Data mining is the process of extracting valuable knowledge from big amount of data. Data collected for different purposes may change hands over time. Lindell and Pinkas [7] discuss how to apply some of the data mining methods on the union of two databases owned by competing companies. Pinkas [8] shows cryptographic techniques to provide a secure computation of data mining algorithms. Verykios et al. [9] provide a general description of data analysis based on privacy-preserving scheme. Clifton et al. [10] define and create privacy measures for future studies.

Spatial interpolation methods are used to create a prediction method for a specific region. Due to time and budget concern, it is impossible to collect measurements for all points. Therefore samples are taken from the field. Sampling strategy affects the accuracy of the interpolation models. Zhang et al. [11] study this fact in their paper. The results of interpolation methods may vary according to geographic properties of the field. Scientists try to apply all interpolation methods and choose the best one according to the accuracy of the model. Chen et al. [12] run several experiments to determine fishery resources density in the Yellow sea. They concluded that IDW and Ordinary Kriging give the best results.

TSA is one of the broadly accepted and applied spatial interpolation methods. Oldham and Sutherland [13] introduced TSA for the first time. Agterberg [14] describes the basics of TSA and discusses applied case studies in environmental studies. Grohmann [15] performed TSA on data from morphometric parameters isobase and hydraulic gradient. He concluded that sixth and second order surface expressions gave the best accuracy. Yao et al. [16] use TSA and several other spatial interpolation methods to figure out groundwater level in northern China. Gao and Prasad [17] have developed a mobile app to find position in a closed areas such as buildings and factories. They utilized TSA as a spatial interpolation tool.

The author has proposed studies in the area of privacy-preserving spatial interpolation. In the literature, previous spatial interpolation studies did not take privacy into consideration. We have completed the necessary studies on privacy preserving IDW and Kriging. The author published three studies on Privacy-preserving Kriging. The first paper [6] deals with how to build a secure kriging interpolation model based on single-party architecture. In this scheme there are two parties; a client and server. The server holds all the necessary sampled data to create a kriging model. The model and sampled data are assumed as the private data of the server. The client requests a prediction value for a specific coordinate from the server. The final result of the prediction value and coordinate are the private data of the client. The other two papers propose two and multi-party architecture [18, 19]. In addition to these

studies, they published the work on privacy-based IDW for single-party architecture [6]. Two-party and multi-party solutions have been studied and are in the review process.

3. BACKGROUND

Spatial interpolation is concerned with spatial data [3]. IDW, Kriging and TSA are the most known and used methods of spatial interpolation. The purpose of all interpolation methods is to produce a prediction value for an unmeasured location. Methods used in spatial interpolation can be grouped into two: deterministic and geostatistical. Deterministic methods create models based on similarity of sampled data. On the contrary, statistical properties of sampled data are taken into account by geostatistical methods. Furthermore, deterministic methods can be divided into two groups as global and local, depending on the number of sampled data used to produce a prediction.

Trend surface is a kind of multiple regression forms where the predictors are the spatial coordinates [20]. A general prediction function can be defined as follows;

$$z(x_t, y_t) = f(x, y) + \epsilon \quad (1)$$

where $z(x_t, y_t)$ is the prediction value at target point (x_t, y_t), f denotes a function that is based on sampled data coordinates and ε is the error term of the model. The sampled data are expressed as polynomial equations such as plane, quadratic or cubic by the least squares method. Thus for a plane, regression equation would be

$$z = b_0 + b_1 x + b_2 y, \quad (2)$$

and for a quadratic surface

$$z = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 y^2 + b_5 xy. \quad (3)$$

Higher degree expressions can be used for more complex surfaces [21].

There are two types of cryptographic algorithms; symmetric (private) and asymmetric (public) [22]. Symmetric algorithms are built on two operations; confusion and diffusion. Symmetric algorithms use only one key which is shared by both users who want to build a secure communication. However, each user has two keys in asymmetric algorithms. The first key is known as public which can be shared with all users. The second one is the private key which must be kept secret by the owner of the key. Asymmetric algorithms are built on three mathematical problems; integer factorization, discrete logarithm and elliptic curves. RSA is the well-known and most used public cryptography algorithm which utilizes big prime numbers and modular arithmetic.

Homomorphic encryption built on public cryptography allows addition and multiplication operation to be carried out on cipher text. There are several such algorithms in the literature. We employ [23] cryptosystems which satisfies two homomorphic properties. The first property known as the homomorphic addition of plaintexts is expressed as;

$$D(E(m_1, pk) \cdot E(m_2, pk)(mod\ n^2)) = m_1 + m_2\ (mod\ n). \quad (4)$$

The first property states that multiplication of two ciphertexts will be equal to the sum of their corresponding plaintexts. Furthermore, the second property is expressed as;

$$D(E(m_1, pk)^{m_2}\ (mod\ n^2)) = m_1\, m_2\ (mod\ n) \quad (5)$$

The second property states that encrypted plaintext raised to the power of another plaintext will be equal to the product of the two plaintexts.

4. PRIVACY-PRESERVING TREND SURFACE ANALYSIS (PPTSA)

As explained above in traditional TSA there is no privacy concern. The clients and servers may be in an unprotected state. However future of organizations depends on the data they possess. Therefore clients and data owners must conduct their scientific calculations without revealing their private data. We will propose PPTSA which provides privacy for both the clients and servers. Our solution follows these steps;

i. The client decides the target coordinates $(x_t,\ y_t)$ where it needs a prediction.
ii. The server builds the optimal trend surface model using the sampled data.
iii. After finalizing the prediction model, the server creates an input sequence according to the model.

Assume that the model built by the server is

$$z = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 y^2 + b_5 xy,$$

in that case, corresponding input sequence will be $[1, x, y, x^2, y^2, xy]$ (For the sake of clarity, solution of a quadratic equation is explained in details, but a higher degree equation can be solved in the same manner). The

server must not send the coefficients $[b_0, b_1, b_2, b_3, b_4, b_5]$ of the prediction model in a clear or readable form; otherwise the client learns the model.

iv. The client calculates the corresponding value of each input sequence $[1,\ x_t,\ y_t,\ x_t^2,\ y_t^2,\ x_t y_t]$.
v. Then the client encrypts each input sequence with its encryption key $[\varepsilon_c(1), \varepsilon_c(x_t),\ \varepsilon_c(y_t),\ \varepsilon_c(x_t^2), \varepsilon_c(y_t^2), \varepsilon_c(x_t y_t)]$ using Paillier homomorphic scheme and sends back to server. As long as the server does not know the corresponding decryption key, it cannot decrypt the values and learn the target coordinate.
vi. The server first calculates each expression using the model coefficients $[\varepsilon_c(1)^{b_0},\ \varepsilon_c(x_t)^{b_1},\ \varepsilon_c(y_t)^{b_2},\ \varepsilon_c(x_t^2)^{b_3}, \varepsilon_c(y_t^2)^{b_4}, \varepsilon_c(x_t y_t)^{b_5}]$.
vii. Then it multiplies all the encrypted values $[\varepsilon_c(1)^{b_0} \cdot \varepsilon_c(x_t)^{b_1} \cdot \varepsilon_c(y_t)^{b_2} \cdot \varepsilon_c(x_t^2)^{b_3} \cdot \varepsilon_c(y_t^2)^{b_4} \cdot \varepsilon_c(x_t y_t)^{b_5}]$ to get the prediction value at the target coordinate which is equal to the encrypted form of "$b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 y^2 + b_5 xy$" which it then sends back to the client.
viii. Finally, the client gets the prediction value in encrypted form $\varepsilon_c\left(z(x_t, y_t)\right)$ and decrypts it with the key known by the client only.

5. ANALYSIS OF THE PROPOSED SOLUTION

5.1 Supplementary Cost Analysis

We analyze PPTSA in terms of additional storage, communication and computation cost due to privacy concerns. In traditional trend surface analysis client stores the target coordinate and prediction value only. However, in PPTSA the client also need to store the input sequence sent by the server and their corresponding values. The length of the sequence depends on the model built by the server. To make it more clear, if the model is in quadratic form, input sequence and corresponding values will be $[1, x, y, x^2, y^2, xy]$ and $[1, x_t, y_t, x_t^2,\ y_t^2, x_t y_t]$. On the other hand, the server stores sampled data, the model and the prediction value in traditional trend surface analysis. Additionally the server needs to store clear $[1, x, y, x^2, y^2, xy]$ and encrypted form $[\varepsilon_c(1), \varepsilon_c(x_t),\ \varepsilon_c(y_t),\ \varepsilon_c(x_t^2), \varepsilon_c(y_t^2), \varepsilon_c(x_t y_t)]$ of input sequence in PPTSA.

In traditional TSA the client sends the target coordinate to the server and the server calculates the prediction value and sends back to the client. Therefore the total number of communication between client and server is two. However, our proposed solution increases the number of communication due to privacy measures. Plain $[1, x, y, x^2, y^2, xy]$ and encrypted $[\varepsilon_c(1), \varepsilon_c(x_t),\ \varepsilon_c(y_t),\ \varepsilon_c(x_t^2), \varepsilon_c(y_t^2), \varepsilon_c(x_t y_t)]$ forms of input sequence must be exchanged between the server and client.

The client does not do any calculations in traditional TSA. All the necessary calculations are performed by the server. However, in PPTSA, the client needs to encrypt the input sequence and decrypt prediction value calculated by the server in encrypted form. In traditional TSA, the server gets the target coordinate from the client. The server builds the model in advance, therefore it has all the necessary data to produce the prediction value. On the other hand, the server needs to calculate $[\varepsilon_c(1)^{b_0}, \varepsilon_c(x_t)^{b_1}, \varepsilon_c(y_t)^{b_2},\ \varepsilon_c(x_t^2)^{b_3}, \varepsilon_c(y_t^2)^{b_4}, \varepsilon_c(x_t y_t)^{b_5}]$ and $[\varepsilon_c(1)^{b_0} \cdot \varepsilon_c(x_t)^{b_1} \cdot \varepsilon_c(y_t)^{b_2} \cdot \varepsilon_c(x_t^2)^{b_3} \cdot \varepsilon_c(y_t^2)^{b_4} \cdot \varepsilon_c(x_t y_t)^{b_5}]$ to produce the prediction value in PPTSA.

5.2 Accuracy Analysis

Privacy preserving schemes may worsen the accuracy of the prediction model due to the methods used to provide privacy. Adding random data to sampled data is one of the examples of such methods. However, we deployed Paillier cryptosystem which is a kind of homomorphic encryption algorithms. Paillier cryptosystem satisfies two properties; homomorphic addition and multiplication of plaintexts. Paillier cryptosystem is also a deterministic encryption method which means that doing the calculation on cipher-texts does not affect the result. Therefore, the client gets exact prediction value as if there is no privacy measure. Consequently, PPTSA produces same prediction value as in traditional scheme. In addition to this, it also offers to protect confidential data of both client and server.

5.3 Privacy Analysis

Our proposed solution should be verified to ensure that both client's and server's data are not disclosed to each other and third parties during calculation of prediction value. The target coordinate and prediction value is accepted as the private data of the client. Besides, the server must keep the coefficients of the prediction model and sampled data secret from the client. The degree of the model does not reveal sensitive information. All the necessary data transfers and calculations to produce a prediction value at target coordinate are done in encrypted form. As Paillier [23] stated that HE is secure as long as the client hides its private key from other users.

6. CONCLUSION AND FUTURE WORK

TSA is a fundamental spatial interpolation method. However, it does not ensure the privacy of both clients and servers. We proposed a privacy-preserving solution to protect private data of both parties. PPTSA guarantees that both parties' data will be kept secret. Therefore they do not need to worry about their private data. We analyzed our solution in terms of privacy and performance. We proved that PPTSA builds a prediction model as accurate as a traditional model in which there is no privacy. However, it increases storage, communication and computation cost in negligible.

In the proposed architecture data is collected by only one server. On the contrary, data may be collected for the same or neighbor region by two or more servers. We are planning to study how to produce TSA-based predictions on data partitioned between competing servers which are willing to cooperate to build more accurate prediction models while preserving their and client's privacy. Furthermore, there are more than ten spatial interpolation methods in the literature. Remaining methods may be explored to create a comparative spatial analysis.

7. ACKNOWLEDGEMENTS

Part of this study was supported by a grant of Ankara University Scientific Research Projects (BAP - 15B0443011).

8. REFERENCES

[1] W. R. Tobler, Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, **74**(367), 519-530 (1979).

[2] D. G. Krige, A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, **52**(6), 119-139 (1951).

[3] J. Li, A.D. Heap, Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, **53**, 173-189 (2014).

[4] D. Shepard, A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pp. 517-524 (1968).

[5] B. Tugrul, H. Polat, Estimating kriging-based predictions with privacy. *International Journal of Innovative Computing, Information and Control*, **9**(8), 3197-3210 (2013).

[6] B. Tugrul, H. Polat, Privacy-preserving inverse distance weighted interpolation. *Arabian Journal for Science and Engineering*, **39**, 2773-2781 (2013).

[7] Y. Lindell, B. Pinkas, Privacy preserving data mining. In *Annual International Cryptology Conference*, pp. 36-54 (2000).

[8] B. Pinkas, Cryptographic techniques for privacy-preserving data mining. *ACM Sigkdd Explorations Newsletter*, **4**(2), 12-19 (2002).

[9] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, Y. Theodoridis, State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, **33**(1), 50-57 (2004).

[10] C. Clifton, M. Kantarcioglu, J. Vaidya, Defining privacy for data mining. In *US National Science Foundation Workshop on Next Generation Data Mining*, pp. 126-133 (2002).

[11] H. Zhang, L. Lu, Y. Liu, W. Liu, Spatial Sampling Strategies for the Effect of Interpolation Accuracy. *ISPRS International Journal of Geo-Information*, **4**(4), 2742-2768 (2015).

[12] Y. Chen, X. Shan, X. Jin, T. Yang, F. Dai, D. Yang, A comparative study of spatial interpolation methods for determining fishery resources density in the Yellow Sea. *Acta Oceanologica Sinica*, **35**(12), 65-72 (2016).

[13] C.H.G. Oldham, D.B. Sutherland, Orthogonal polynomials: Their use in estimating the regional effect. *Geophysics*, **20**(2), 295-306 (1955).

[14] F.P. Agterberg, Trend surface analysis. In *Spatial statistics and models*, pp. 147-171 (1984).

[15] C.H. Grohmann, Trend-surface analysis of morphometric parameters: A case study in southeastern Brazil. *Computers & geosciences*, **31**(8), 1007-1014 (2005).

[16] L. Yao, Z. Huo, S. Feng, X. Mao, S. Kang, J. Chen, J. Xu, T.S. Steenhuis, Evaluation of spatial interpolation methods for groundwater level in an arid inland oasis, northwest China. *Environmental earth sciences*, **71**(4), 1911-1924 (2014).

[17] S. Gao, S. Prasad, Employing spatial analysis in indoor positioning and tracking using wi-fi access points. In *Proceedings of the Eighth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, pp. 27-34 (2016).

[18] B. Tugrul, H. Polat, Privacy-preserving kriging interpolation on partitioned data. *Knowledge-Based Systems*, **62**, 38-46 (2014).
[19] B. Tugrul, H. Polat, Privacy-Preserving Kriging Interpolation on Distributed Data. In *International Conference on Computational Science and Its Application*, pp. 695-708 (2014).
[20] R. Webster, M.A. Oliver, *Geostatistics for environmental scientists*. John Wiley & Sons (2007).
[21] Z. Sen, "Spatial modeling principles in earth sciences". Springer (2009).
[22] C. Paar, J. Pelzl, *Understanding cryptography: A textbook for students and practitioners*. Springer Science & Business Media (2009).
[23] P. Paillier, Public-key cryptosystems based on composite degree residuosity classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 223-238 (1999).