Kriging Analysis In The Spatial Domain For Dispersion Models

Taduri Vamshi Krishna and Debashis Dutta, *Member, IAENG*

Abstract— **Statistical data analysis on toxicity for air pollution problems is studied for the atmosphere with the various pollutant's observations. In this paper Kriging techniques are explained for pollutant concentration, also forecast by Kriging methods. Approximate function called as system function is obtained by these techniques. An approach is mentioned for calculating weights for the given data. This model can easily be extended to higher dimensional air pollution data.**

Index Terms— **Kriging, Spatial analysis, Support Vector Regression (SVR), Variogram.**

I. Introduction

In air pollution modeling we observe the dispersion by statistical fitting techniques. Then we forecast the concentration of pollutant through which we take some monitoring actions. One of the fitting techniques is spatial interpolation known as Kriging. Most of the applications of this technique have been used in mining, hydrology and contamination [2]. Reference [6] has used it for acid precipitation data. Understanding the plume concentration of any contaminant is a multi faceted problem. Often data is limited and modeling exercise is tedious. These techniques are for capitalizing on the measurements of the level of contaminant in the atmosphere. Hence we are explaining and analyzing the procedure with mathematical and statistical methods. Kriging procedures are applied to specific data with the help of Variogram methods.

Kriging is a form of weighted average estimator. The weights are assigned on the basis of model fitted to variogram function which represents spatial structure in the variable. The most frequently used form of kriging is ordinary kriging (OK). OK estimates linear weighted dynamic averages of the n available observations over which estimation is made. Bayesian probability density for interpolation function extrapolates ([9]) to the density of the least squares linear model. This density can be used to implement cardinal interpolation [5] since cardinal points are of basic importance for sampled data. Generally we consider the three points i.e. mean, mean plus standard deviation and mean minus standard deviation of the interpolator probability density function. Variance function extrapolates a quadratic function for the least squares linear model. The standard natural cubic spline interpolator extrapolates to the lines that diverge greatly from the least squares line. The same divergent extrapolation behavior is commonly found in models obtained by Gaussian process such as Kriging models.

A case study problem is considered in section III and kriging methods have been implemented.

Manuscript received September 29, 2008.

Taduri Vamshi Krishna is with the Department of Mathematics, National Institute of Technology, Warangal, Andhra Pradesh, India (e-mail: tvkmath@ gmail.com).

Dr. Debashis Dutta, is with the Department of Mathematics, National Institute of Technology, Warangal, Andhra Pradesh, PIN-506004, India. (Phone: 91 9849548445, Fax: 91 870 2459547, e-mail: dduttamath@gmail.com). Dr. Dutta is thankful to the Institute for giving financial support.

II. MATHEMATICS OF KRIGING

The basic objective of kriging is to derive an empirical model for the stochastic component of an observation. This model allows estimating the value of variable points of the observations.

Let kriging assume the variable $C(z)$

$$C(x) = C_m + \varepsilon(z) \quad \textbf{(2.1)}$$

Where C_m is mean of given sampled data and $\varepsilon(z)$ is noise in the given data.

Kriging assumes that Forecasting estimate of $C(z)$ is $\overline{C(z)}$, it is also known as system function and is approximated as follows:

$$\overline{C(z)} = \sum_{j=1}^{n} \lambda_j C_j \quad \textbf{(2.2)}$$

Where λ_j are weights, C_j are observations of pollutant's concentration.

λ_j's are determined by insisting that ensemble averaged variance of $\overline{C(z)}$.

Ensemble average bias between the estimates $\overline{C(z)}$ and true value of C_m is zero.

$$< \left(\overline{C(z)} - C_m\right) > = 0 \quad \textbf{(2.3)}$$

Substituting (2.2) in (2.3)

$$< \left(\sum_{j=1}^{n} \lambda_j C_j - C_m \right) > = 0 \quad \textbf{(2.4)}$$

Since ($< C_j > = < C_m > = C_m$) then (2.4) reduces to

$$\sum \lambda_j = 1 \quad \textbf{(2.5)}$$

So Variance term is defined as follows which is to be minimized

$$V = V(\lambda_j) = < \left(\sum_{j=1}^{n} \lambda_j C_j - C_m \right)^2 > \quad \textbf{(2.6)}$$

Subject to (2.5)

Theorem 1: As (2.6) is an inner product space then

$$V = V(\lambda\ ) = -\sum \lambda_i \sum \lambda_j \gamma_{ij} + 2\sum \lambda_j \gamma_{jm}$$

And γ_{ij} is semi variogram and is a function of separation distance. $d = (x_i - x_j)$,

$$\gamma_{ij} = f\left(\frac{< \left(C_i - C_j\right)^2 >}{2} \right)$$

Proof: $V = V(\lambda_j) = < \left(\sum_{j=1}^{n} \lambda_j C_j - C_m \right)^2 >$

$$= \sum \lambda_i \sum \lambda_j C_i C_j - 2C_m \sum \lambda_j C_j + C_m^2 \quad \textbf{(2.7)}$$

Let $\sum \lambda_j C_i C_j = \frac{1}{2} \sum \lambda_j \left(C_i^2 + C_j^2 - (C_i - C_j)^2 \right)$

$$\sum \lambda_i \sum \lambda_j C_i C_j =$$

$$\frac{1}{2}\left[\sum \lambda_i \sum \lambda_j C_j^2 + \sum \lambda_j \sum \lambda_i C_i^2 - \sum \lambda_i \sum \lambda_j (C_i - C_j)^2 \right]$$

$$= \left[\sum \lambda_i C_i^2 - \frac{1}{2} \sum \lambda_i \sum \lambda_j (C_i - C_j)^2 \right] \quad \textbf{(2.8)}$$

Since $\sum \lambda_i = \sum \lambda_j = 1$

Substituting (2.8) in (2.7)

$$V =$$

$$\sum \lambda_j C_i^2 - \frac{1}{2} \sum \lambda_i \sum \lambda_j (C_i - C_j)^2 - 2C_m \sum \lambda_j C_j + C_m^2$$

$$= \sum \lambda_j \left(C_j - C_m \right)^2 - \sum \lambda_i \sum \lambda_j \frac{\left(C_i - C_j\right)^2}{2}$$

$$= 2\sum \lambda_j \gamma_{jm} - \sum \lambda_i \sum \lambda_j \gamma_{ij} \qquad \square$$

Since the best estimator will minimize the deviations, the estimation of these variogram from the given data is explained in section IV.

Let us introduce Lagrange function for (2.6) and (2.5)

$$G(\lambda, \mu) = -\sum \lambda_i \sum \lambda_j \gamma_{ij} + 2\sum \lambda_j \gamma_{jm} + \mu\left(\sum \lambda_j - 1\right) \quad \textbf{(2.9)}$$

Here μ is Lagrange multiplier.

Hence critical points (λ, μ) are obtained by solving normal equations

Partial derivative with respective to λ_k

$$\frac{\partial}{\partial \lambda_k} G(\lambda, \mu) =$$

$$\frac{\partial}{\partial \lambda_k}\left(-\sum \lambda_i \sum \lambda_j \gamma_{ij} + 2\sum \lambda_j \gamma_{jm} + \mu\left(\sum \lambda_j - 1\right)\right) = 0 \quad \textbf{(2.10)}$$

$$\Rightarrow \quad -2\sum \lambda_j \gamma_{kj} + 2\gamma_{km} + \mu = 0$$

[Since $\gamma_{ij} = \gamma_{ji}$]

$$\Rightarrow \quad \sum \lambda_j \gamma_{ij} - \frac{\mu}{2} = \gamma_{im} \qquad \textbf{(2.11)}$$

i, j=1 to n

$$\frac{\partial}{\partial \mu} G(\lambda, \mu) =$$

$$\frac{\partial}{\partial \mu}\left(-\sum \lambda_i \sum \lambda_j \gamma_{ij} + 2\sum \lambda_j \gamma_{jm} + \mu\left(\sum \lambda_j - 1\right)\right) = 0$$

This reduces to (2.5) i.e., $\sum \lambda_j = 1$

Here there are n+1 unknowns $\{\lambda_1, \lambda_2, ...\lambda_n, \mu\}$ and n+1 equations.
Hence the system comprising of (2.11) and (2.5) is consistent. Solving the above system using MatLab /Mathematica we can obtain $\{\lambda_1, \lambda_2, ...\lambda_n\}$.
These are known as weight for the given data.
Now we can fit a function by any Numerical Interpolation technique for the data (λ_i, z_i, C_i)
Criteria for convergence of the system are $\lambda_j \geq 0$. **(2.12)**
So we can use Two-Phase simplex procedure for (2.11), (2.5) and (2.12).
Then we shall obtain positive λ_j.

III. A CASE STUDY

Let us consider the data of pollution concentration (Table I).

Table I

Pollution data from [7]

Observation	z	C(x,z)
1	0	0
2	0.1	0.881779
3	0.2	0.588789
4	0.3	0.388469
5	0.4	0.251239
6	0.5	0.158012
7	0.6	0.095877
8	0.7	0.055684
9	0.8	0.030714
10	0.9	0.015966
11	1	0.007764

Semi- Variograms [8] are calculated using

$$\gamma_{ij} = f\left(\frac{< (C_i - C_j)^2 >}{2}\right) = d(i \sim j)$$

d(0)=0;
d(1) = 0.008882;
d(2) = 0.028210;
d(3)= 0.052278;
d(4)= 0.079510;
d(5) = 0.110776;
d(6)= 0.148990;
d(7) = 0.199562;
d(8) = 0.271806;
d(9)= 0.381951;

Equations (2.11) and (2.5) are written in augmented form as follows:

$[A : B]$ and constrained to (2.12)

$$\left[\begin{array}{ccccccccc:c}
d(0) & d(1) & d(2) & d(3) & d(4) & d(5) & d(6) & d(7) & d(8) & -0.5 : d(1) \\
d(1) & d(0) & d(1) & d(2) & d(3) & d(4) & d(5) & d(6) & d(7) & -0.5 : d(2) \\
d(2) & d(1) & d(0) & d(1) & d(2) & d(3) & d(4) & d(5) & d(6) & -0.5 : d(3) \\
d(3) & d(2) & d(1) & d(0) & d(1) & d(2) & d(3) & d(4) & d(5) & -0.5 : d(4) \\
d(4) & d(3) & d(2) & d(1) & d(0) & d(1) & d(2) & d(3) & d(4) & -0.5 : d(5) \\
d(5) & d(4) & d(3) & d(2) & d(1) & d(0) & d(1) & d(2) & d(3) & -0.5 : d(6) \\
d(6) & d(5) & d(4) & d(3) & d(2) & d(1) & d(0) & d(1) & d(2) & -0.5 : d(7) \\
d(7) & d(6) & d(5) & d(4) & d(3) & d(2) & d(1) & d(0) & d(1) & -0.5 : d(8) \\
d(8) & d(7) & d(6) & d(5) & d(4) & d(3) & d(2) & d(1) & d(0) & -0.5 : d(9) \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 : 1
\end{array}\right]$$

The solution of the above problem is obtained as $\lambda_2 = 1$ and all remaining variables are zero.
Fig.1 and Fig. 3 give the variation between Kriging values with exact data and interpolated data respectively. It concludes that the peak estimation shown as $\overline{C(z)}$ is same. Fig. 2 analyzes and it is observed that they are highly correlated.

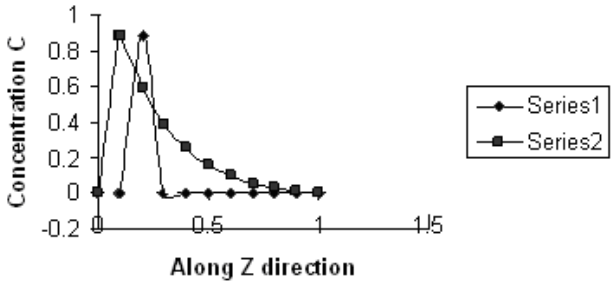

Fig. 1: variation between exact data (Series 1) and Kriging values (Series 2).

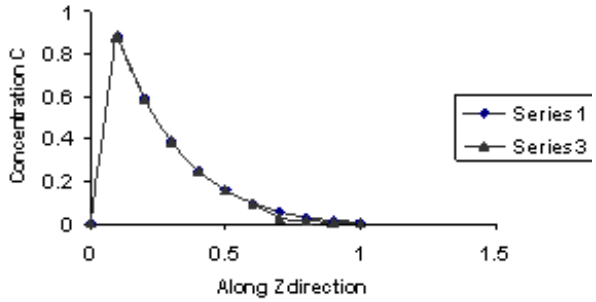

Fig. 2: Variation between Exact data (Series 1) values with interpolated values (Series 3)

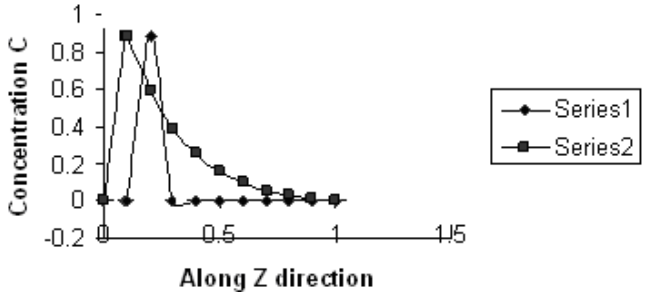

Fig. 3: variation between interpolated (Series 1) and Kriging values (Series 2)

IV. ESTIMATING PARAMETERS IN VARIOGRAM AND SVR

i) Let $\overrightarrow{\gamma(h)}$ be a vertical form of a discrete exponential semi variogram [4] containing k estimates of the computed semi variogram for increasing value of lag h.

$$\overrightarrow{\gamma(h)} = [\gamma(h_1), \gamma(h_2), \gamma(h_3) \ldots\ldots\ldots\ldots \gamma(h_k)] \quad \textbf{(4.1)}$$

Let $\overrightarrow{\gamma(h,\theta)}$ be a vector with value for semi variogram with unknown parameter $\theta_i, i = 1,2,..k$.

Using least square principle, the best set of parameters between the given values and their predicted values by the model is

$$S = \left[\overrightarrow{\gamma(h)} - \gamma(h,\theta)\right]^T M \left[\overrightarrow{\gamma(h)} - \gamma(h,\theta)\right] \quad \textbf{(4.2)}$$

Where M is variance matrix.

If M is identity matrix, then

$$S = \sum_{i=1}^{k} \left[\overrightarrow{\gamma(h_i)} - \gamma(h,\theta_i)\right]^2 \quad \textbf{(4.3)}$$

If M is diagonal matrix, then

$$S = \sum_{i=1}^{k} W_i \left[\overrightarrow{\gamma(h_i)} - \gamma(h,\theta)\right]^2 \quad \textbf{(4.4)}$$

$$M = [e_{ij}] where \quad e_{ij} = \begin{cases} W_i & for\ i = j \\ 0 & otherwise \end{cases}$$

Equation (4.4) is weighted least squares approximation, through which best fit can be obtained. This best fit is useful for predicting or extrapolating the values.

$\gamma(h,\theta)$ is a two parameter family of curves . It may be spherical, exponential, power, Gaussian, cubic, etc.

ii) A new paradigm [3] analyzing and learning from data is called support vector machines (SVM). Now it is generalized for regression called support vector regression (SVR). We have set of data points generated from the atmospheric pollution say $P(z, C(z))$.

The objective is to minimize the risk functional $R[C]$

$$R[C] = \int Q(C - \overline{C}, z) \quad dP(z, C) \quad \textbf{(4.5)}$$

Where Q is loss function/error function

Also the empirical expression is

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} Q_i (C - \overline{C}, z)_i \quad \textbf{(4.6)}$$

Hence loss function can be defined as ε-insensitive loss

$$Q(C - \overline{C}, z) = \begin{cases} \left|C - \overline{C}\right| - \varepsilon & if \quad \left|C - \overline{C}\right| > \varepsilon \\ 0 & otherwise \end{cases} \quad \textbf{(4.7)}$$

Establishing the connection between maximum likelihood estimator and risk minimization

$$P(z, C) = P(C / z) P(z)$$

$$= P(C - \overline{C}) P(z) \quad \textbf{(4.8)}$$

So we can write likelihood estimator

$$P(z_1, C_1; z_2, C_2; \ldots\ldots\ldots\ldots; z_n, C_n)$$

$$= \prod_{i=1}^{n} P(C - \overline{C})_i \, P(z_i) \qquad \textbf{(4.9)}$$

A distribution function $P(C - \overline{C})$ is obtained by Variogram methods. Consider exponential distribution.

$$P(z_1, C_1; z_2, C_2; \ldots\ldots\ldots\ldots; z_n, C_n)$$

$$= e^{-\sum_{i=1}^{n} Q(C-\overline{C})_i} \prod_{i=1}^{n} P(z_j) \qquad \textbf{(4.10)}$$

Hence maximum likelihood minimizes risk functional $R[C]$ [1].

V. CONCLUSIONS

In this paper pollution concentration is approximated by Kriging function. The expected peak pollutant concentration has been forecast. If we simulate the system with the proposed methods, we can also find the best fitting values. Variogram estimates are discussed. However, the above kriging techniques can be extended to higher dimensional air pollution data.

REFERENCES

[1] Jae Myung., Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology, Vol. 47, pp90-100, 2003.

[2] Lloyd C.D, P.M. Atkinson, Assessing uncertainty in estimates with ordinary and indicator Kriging, Computers & Geosciences, Vol. 27, pp 929-937, 2001.

[3] Proceedings, DST/NRDMS Sponsored work shop on Soft Computing Techniques for Spatial Data Analysis, Nov 16-17, 2006, Dept. of Computer and Information Sciences, University of Hyderabad.

[4] Rachel K Boeckenhauer, Dennis D.Cox, Katherine B.Ensor,Philip Bedient , and Anthony W Holder, Statistical estimation and visualization of ground water contamination data. National Risk Management Research Lab, Office of R&D, US EPA/600/R-00/034, August 2000.

[5] Steven C. Gustafson, David R. Parker, Richard K Martin, Cardinal Interpolation. IEEE Transactions on pattern analysis and machine intelligence, Vol. 29, No.9, September 2007.

[6] Venkatram Akul, On this use kriging in the spatial analysis of acid precipitation data. Atmospheric Environment, Vol. 22, No9, pp1963-1975, 1988.

[7] Vamshi Krishna T., K. Appala Raju and Debashis Dutta, A Mathematical Model for Atmospheric Dispersion in an Environment with Pollutants, pp213-221, Proceedings 48th ISTAM Dec (18-21, 2003).

[8] Vamshi Krishna T, Debashis Dutta, Variogram Analysis of Spatial outcomes of Geo-pollution Models, Int J of Logic Based Intelligent Systems, Vol. 2, no 1, pp.97-107, 2008.

[9] Xiaodong Jian, Ricardo A.Olea, Yun-sheng Yu. Semivariogram modeling by weighted least squares. Computer & Geosciences, Vol. 22, No.4, pp 387-397, 1996.

*****