Southeast-Asian J. of Sciences Vol. 6, No 2 (2018) pp. 111-133

ANALYZING INCOMPLETE SPATIAL DATA IN AIR POLLUTION PREDICTION

Man V.M. Nguyen[1,2], **Nhut C. Nguyen**[3]

[1] *Center of Excellency in Mathematics (CEM)*
Ministry of Education, Thailand
272 Rama VI, Bangkok
[2] *Department of Mathematics, Faculty of Science*
Mahidol University, 272 Rama VI, Bangkok
email: man.ngu@mahidol.ac.th

[3] *Department of Information Technology*
Nguyen Tat Thanh University, Vietnam

Abstract

In air pollution studies at metropolis, as in Bangkok or Saigon, installation of new stations for monitoring dangerous pollution sources is costly. Using statistical models and analyzing data sets collected at good stations to predict air pollution levels at malfunctioning stations, therefore, are highly demanding. We study air pollution prediction by geo-statistical methods with a realistic dataset costly observed in Ho Chi Minh City. Geostatistics includes statistical methods for modeling of spatially continuous phenomena, using data measured at a finite number of locations to build up right models, to estimate and predict values of interest (such as air or water pollutant levels in a geographical region, oil volumes of reservoirs under the ocean bed...) at unmeasured locations. To analyze our multivariate data (of SO2, PM-10 and benzen, where the last two are popular air pollution causes at metropolis) recorded in HCMC since 2003, we start from determining suitable co-kriging models for pollutants to predicting these pollutant concentrations at some un-measured stations in the city.

The paper's key contributions include, firstly, formulating co-kriging models and computing theirs optimal unbiased estimators for air pollution prediction using the valuable observed data with two pollutants; secondly, proposing a computational mechanism (*progressively co-kriging imputation*) to deal with missing data at unmeasured monitoring sites.

Key words: air pollution, co-kriging, imputation, multivariate geostatistical techniques, spatial-temporal data analysis, stationary random process
(2010) Mathematics Subject Classification: 62H11, 62F30, 60G10, 60G60

1 Introduction

Environmental pollution, and air pollution in particular, have become critical concerns from both social and scientific views in the globe, and critically serious in developing countries like Vietnam, Thailand, China or India. Specifically, air pollution in metropolitan areas - caused mostly by construction, transportation and industrial manufacturing - increasingly degrades environment quality, and leads to severe problems for health of dwellers as well.

1.1 Mathematics and Geostatistics for air pollution studies

The use of mathematical models to study air pollution has started since 1859 by Robert A. Smith in calculating the distribution of CO2 concentration in Manchester City under Gaussian models (Smith [11]). The most popular model, the ISCST3 (*Industrial Source Complex Short Term*) model is a *Gaussian dispersion model* used to assess the impact of single sources in the United State's industry. The AERMOD model of the US's EPA (briefed of AERMIC-**A**MS/**E**PA **R**egulatory **M**odel **I**mprovement **C**ommittee- Model) is used in placement for the ISC3 Model to study pollution at complex terrains. More precisely, having initially being focused on the regulatory models that are designed for estimating near-field impacts from a variety of industrial source types, see [15].

Key statistical methods which are popularly employed in these models form a specific class of statistical science, named **geostatistics**. The methods of geo-statistics provide quantitative descriptions of natural variables distributed in space or in *time and space*. Examples of such variables are ore grades in a mineral deposit, concentrations of pollutants in a contaminated site etc. Investment or management decisions are based on studies involving many disciplines besides *geostatistics*, but they illustrate the notion of **spatial uncertainty** and how it affects development decisions.

The modern approach of geostatistics deals with the inherent uncertainty of spatial data in a stochastic way; more precisely is to treat the variable of interest as a random variable, or better *spatial random variable*. This implies that at each point in space, $\boldsymbol{x} \in \mathbb{R}^3$ there is a series of values for a property, $Z(\boldsymbol{x})$, and the one observed, $z(\boldsymbol{x})$, is drawn at random according to some law, from some probability distribution. Statistics come into play because *probability distributions* are the meaningful way to represent the range of possible values of a parameter of interest. In addition, a statistical model is well-suited to the apparent randomness of spatial variations. The prefix "geo" emphasizes the spatial aspect of the problem. Geostatistics, due to Jean-Paul and al. (Jean Paul [6]) briefly includes a few types of problems, ranging from

a) *Structural analysis*, in Section 2.2, with the key tool of **variogram**, statistically describes how the values at two points become different as the separation between them increases;

b) *Survey optimization*, to answer questions related to **sampling patterns** which ensure the best precision; to

c) *Spatial interpolation*, to estimate the values of a **regionalized variable** at places where it has not been measured.

Spatial data analysis (SDA) is the task of reducing spatial patterns of geologic variability to a few clear and useful summaries, (Dung [14]). As a major part of SDA, *spatial interpolation* methods include a group of different approaches, among which computational geometry-based methods have been employed a great deal in many environmental sectors, not only for modeling and predicting air pollutants. First of all, *polygon methods* (as nearest neighbor, triangulation method...) have advantages such as easy to use, quick calculation in 2D; but also possesses many disadvantages as discontinuous estimates, edge effects/sensitive to boundaries, and difficult to realize in 3D. Secondly, the *inverse distance method* allows some flexibility for adapting to different estimation problems. This method can handle **anisotropy**;[1] but its weaknesses include difficulties encountered when points to estimate coincide with data points ($d = 0$, weight is undefined), susceptible to clustering. Environmentalists, especially petroleum geologists, also use polynomial-based methods like splines or trend surfaces, see [13, Chapter 3] for more info.

1.2 Study area and its environmental problems

The study area is Ho Chi Minh City metropolis (HCMC) in South Vietnam, shown in Figure 1. With an approximated area of $2096\ km^2$, and around 10 million people. the city has a tropical climate, specifically a tropical wet and dry climate, with an average humidity of $78 - 82\%$, and average temperature of 28^0C (or 82^0F).

Air pollution sources are diverse in HCMC. Consequently, the city has seriously faced environmental pollution, mostly caused by the rapid population growth, the slowly upgraded infrastructure, and last but least, its backward management mechanism. In HCMC metropolis, main sources of pollution include not only rubbish and the above mentioned relevant issues, but also daily dweller's traffic and construction, of which, air pollution caused by traffic activities highly accounts for about 70%, due to 2010 data of Vietnamese Ministry of Transport [10]. The main means of transportation within the city are buses, cars, taxis, motorbikes, and bicycles. The growing number of cars tends to cause gridlock and severely contributes to air pollution.

We study air pollution prediction by geo-statistical methods with a realistic dataset observed at air monitoring stations scattering in and around HCMC. The building of air quality monitoring stations is essential, but also difficult

[1] this term is applied both to a random function and to it's variogram when the values of the variogram depend on both the distance and the direction

because of expensive installation costs, no good information of selected areas for installation in order to achieve precise results.

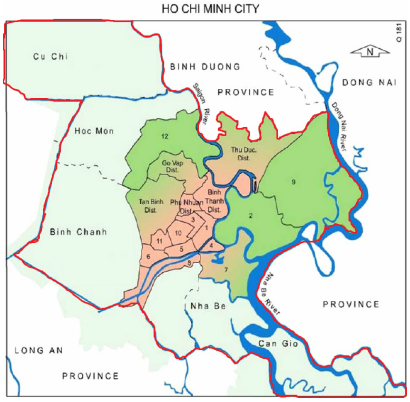

Figure 1: Location of the study area ([16])

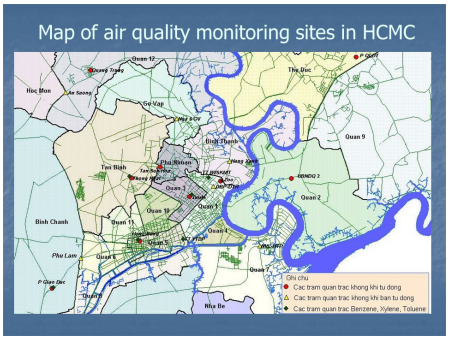

Figure 2: Air quality monitoring sites in HCMC

According to the *Center for Monitoring and Analyzing Environment* (HCMC Department of Resources and Environment), the city's network of air quality monitoring (Figure 2) had 9 automatic observing stations and 6 semi-automatic (i.e. combining manpower and equipment in sampling and analysis) monitoring stations from 2003.

Having played a key role in continuously updating of data of environmental monitoring system in HCMC, these stations were built thanking to financial supports of the Danish and Norwegian governments around 2000. However, this system had been severely degraded, no longer usable since 2009. Costly installation and difficult preservation of stations in humid tropical weather in Vietnam lead to a highly demand in using statistical methods and models to analyze data sets already collected at good stations, then predict air pollution level at some malfunctioning stations, or any place in the city.

1.3 Realistic dataset collected in HCMC metropolis

Figure 2 shows the geographical locations of air monitoring stations, in which the Universal Transverse Mercator (UTM) coordinate system is used. Data obtained from nine automatic monitoring stations, including 4 roadside stations and 5 residential area stations. Daily measurements (24/24 hours) cover at least parameters PM_{10}, SO_2, NO_2, CO, O_3, TSP... (measured in $\mu g/m^3$).

Table 1: Air pollution data

Station	X(m)	Y(m)	PM_{10}	benzen
Thong Nhat- TN	680690	1193530	65.99	31
Binh Chanh- BC	674500	1183000	17.48	29
Zoo	686420	1193370	73.14	31
Doste	684430	1192220	123.50	34
Hong Bang- HB	681620	1189460	NA	NA
District 2- D2	691160	1193510	73.31	39
Quang Trung- QT	677940	1200080	NA	NA
Thu Duc- TD	693640	1199790	NA	NA
Tan Son Hoa- TSH	682830	1193930	67.33	33

The contribution of particulate matter concentration (as PM10- particulate matter with an aerodynamic diameter of at most 10 μm) to air pollution and the effects of high levels of these pollutants to human health have been documented extensively in the literature. Exposure to high concentrations of PM, to a large extent, has been associated with increased rates of morbidity and mortality, caused primarily by cardiovascular, respiratory diseases (Anderson [3]).

Table 2: Air pollution PM-10 data in January - March, 2003

Date	**TN**	**BC**	**ZOO**	**DOSTE**	**TSH**	**D2**	**TD**	**QT**	**HB**
1/1/2003 1:00	404	5.73	87	91.9	149.2	95	NA	NA	NA
1/1/2003 2:00	188	3.82	66	165.7	129.1	100	NA	NA	NA
1/1/2003 3:00	91	1.91	54	149.3	61.1	71	NA	NA	NA
1/1/2003 4:00	73	5.73	45	100.6	4.4	63	NA	NA	NA
...	...		...	...	...	...	...	...	...
4/30/2003 19:00	42	21.01	51	83.1	45.1	47	NA	NA	NA
4/30/2003 20:00	67	9.55	46	48.5	49.2	57	NA	NA	NA
3/31/2003 21:00	57	13.37	32	149.3	69.7	59	NA	NA	NA
3/31/2003 22:00	47	13.37	33	68.4	4.4	66	NA	NA	NA
3/31/2003 23:00	50	13.37	23	140.7	69.7	50	NA	NA	NA

We only use PM_{10} pollution data measured at 9 stations in 3 months of January - March, 2003 and benzen as secondary parameter. Their average values in 3 months, 2013 are listed in Table 1, and a portion (January - March, 2003) of full realistic data set is shown in Table 2, where NA is not available.

Table 3 summarizes the statistics. The data were transformed to common logarithms ($\log_e$) to stabilize the variance in order to better normalize the variate's distribution prior to geostatistical analysis.

Table 3: Summary statistics of PM_{10} and benzen

	$PM_{10}/\mu gm^{-3}$	$\log_e(PM_{10})$	$benzen/\mu gm^{-3}$	$\log_e(benzen)$
Number of data	6	6	6	6
Minimum	17.48	2.86	29	3.37
Maximum	123.50	4.82	39	3.66
Mean	70.125	4.111	32.833	3.487
Std deviation	33.659	0.655	3.488	0.103
Variance	1132.906	0.4284	12.167	0.01055
Skewness	0.04	-1.24	0.86	0.72

Research motivations

Our study is motivated by, firstly the *cost-benefit analysis* in sustainable urban management, the *public health* concern, and lastly, the *statistical- computational interest*. HCMC is the fastest developing city in Vietnam with a lot of problems in urban management, as deciding which geographic locations should be planned for industry ('gray' or 'high-tech'), for building up new infrastructure or creating green living condition for dwellers throughout the time scale. Our study may provide hints to the city's administration when considering trade-offs between sustainable developing and environmentally friendly inhabiting, perhaps by firstly exploiting knowledge extracted from valuable air/groundwater pollution datasets collected with expensive budget? The last motivation of the work is purely statistical, how can we rightly analyze the monitored data sets if so many monitoring sites are malfunctioning?

The paper's contributions and structure

The paper's contributions include, firstly formulating co-kriging models and computing their optimal estimators for the cases of two and three pollutants of air pollution with observed multivariate datasets; and secondly presenting a brief temporal analysis of the pollution process during 2003-2004 in HCMC.

The paper is structured in six parts and an appendix. We first recall background of *spatial random processes* in Section 2, the *kriging* method in Section 3. In Section 4 we employ the *cokriging* method, a key approach of geostatistics, to predict pollutant values. We then propose a computational mechanism - (*progressively co-kriging imputation*) - to deal with incomplete or missing data matter at malfunctioning monitoring sites in Section 5, finally conclusion and new looks follow in Section 6.

2 Background

2.1 Stationarity of spatial random processes

We are going to use geostatistical structures (as *variograms*) to predict key at most three major fatal pollutants of air pollution, say PM_{10}, benzen and SO_2 concentrations at unobserved areas surrounding the observed stations, located in a spatial region D covering HCMC metropolis.

In general let $Z(\boldsymbol{x}) = \{z(\boldsymbol{x}) : \boldsymbol{x} \in D \subset \mathbb{R}^n\}$ be *spatial random function (or random process)*, being used as a model for (or a collection of) the regionalized variables

$$\{z(\boldsymbol{x}) : \boldsymbol{x} \in D \subset \mathbb{R}^n\}$$

representing geological or environmental reality. For simplicity we make no notational distinction between the (uppercase) parent random function $Z(\boldsymbol{x})$ and its particular (lowercase) realization or random variable $z(\boldsymbol{x})$. Denote by S the set of points where $Z(\boldsymbol{x})$ has been sampled. In most cases, S is finite and consists of n data points, denoted and called as locations or **sampling places** $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$.

For each spatial random variable z, however, we have only a single realization. Consequently, we cannot compute statistics for the realization or draw inferences from this spatial random process (obtained by measuring the random variable z at many places in spaces). To overcome this troublesome we must assume that the spatial process is **stationary**. Being stationary means its spatial statistics or laws are invariant under translation in $\mathbb{R}^n$.

To be precise, a spatial random process satisfies second-order stationary if
i) $\mathbf{E}[Z(\boldsymbol{s})]$ exists and does not depend on $\boldsymbol{s}$, and furthermore
ii) $\mathbf{E}[Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})] = 0$, the expected differences are zero.

Definition 1 (Covariance of a spatial random process).

We represent the process by the model

$$Z(\boldsymbol{s}) = \mu + \varepsilon(\boldsymbol{s}), \tag{2.1}$$

where $\mu = \mathbf{E}[Z]$ is the process mean and $\varepsilon(\boldsymbol{s})$ is a random quantity with a mean of zero and a covariance, $\mathbf{C}(\boldsymbol{h}) = \mathbf{Cov}[Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{h})]$, given by

$$\mathbf{C}(\boldsymbol{h}) = \mathbf{E}[(Z(\boldsymbol{s}) - \mu)(Z(\boldsymbol{s}+\boldsymbol{h}) - \mu)] = \mathbf{E}[\varepsilon(\boldsymbol{s})\ \varepsilon(\boldsymbol{s}+\boldsymbol{h})]. \tag{2.2}$$

In these equations the lag $\boldsymbol{h}$ is the separation between samples in both distance and direction; $Z(\boldsymbol{s})$ and $Z(\boldsymbol{s}+\boldsymbol{h})$ are the values of Z at places $\boldsymbol{s}$ and $\boldsymbol{s}+\boldsymbol{h}$, and $\mathbf{E}$ denotes the expectation. Under the second-order stationarity, we replace the covariance $\mathbf{C}(\boldsymbol{h})$ by half the variance of the differences, the semivariance:

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathbf{E}\big[\{Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})\}^2\big]. \tag{2.3}$$

2.2 Variograms

Therefore, under the second-order stationary conditions (Webster [13]), one obtains $\mathbf{E}[Z(\boldsymbol{s})] = \mu$ and the covariance $\mathbf{C}(\boldsymbol{h})$, given by:

$$\mathbf{C}(\boldsymbol{h}) = \mathbf{E}[(Z(\boldsymbol{s}) - \mu)(Z(\boldsymbol{s}+\boldsymbol{h}) - \mu)] = \mathbf{E}[Z(\boldsymbol{s})Z(\boldsymbol{s}+\boldsymbol{h}) - \mu^2]. \tag{2.4}$$

Then $\mathbf{Var}[Z(\boldsymbol{s})] = \mathbf{C}(0) = \mathbf{E}[Z(\boldsymbol{s}) - \mu]^2$, and for second-order stationary processes the covariance function (2.4) and the semivariance function

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathbf{E}\big[\{Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})\}^2\big] = \mathbf{C}(0) - \mathbf{C}(\boldsymbol{h}), \tag{2.5}$$

are equivalent, where $\mathbf{C}(\mathbf{0}) = \mathbf{V}[Z]$ is the variance of the random process.

Definition 2 (The variograms).

- This semivariance $\gamma(\boldsymbol{h})$ given in (2.5) is said to be the **theoretic variogram**, depending on $\boldsymbol{h}$ and only on $\boldsymbol{h}$. This variogram expresses the *spatial correlation* between neighboring observations, expressible in terms of the (auto)covariance function.
- With observed data we practically use the *sample variogram* $\gamma(\boldsymbol{h})$ (Oliver [7]), being one estimated from data $z(\boldsymbol{s}_i),\ i = 1, 2, \ldots$, defined as one-half of the variance of the difference between the attribute values at all points separated by a distance $\boldsymbol{h}$, given by

$$\gamma(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} \big\{z(\boldsymbol{s}_i) - z(\boldsymbol{s}_i + \boldsymbol{h})\big\}^2 \tag{2.6}$$

  where $N(\boldsymbol{h})$ is the total number of pairs of attributes that are separated by a lag $\boldsymbol{h}$. We also call $\gamma(\boldsymbol{h})$ the **experimental variogram**.

We next recall key concepts and results of kriging method, which is helpful for the subsequent developing story.

3 The classic kriging method

Kriging technique employs an exact interpolation estimator, aimed to find the best linear unbiased estimate (BLUE, having a minimum variance of the error of estimation), named *kriging variance or estimator*. We start with **ordinary kriging** (**OK**, for spatial data following an intrinsically stationary process) for the subsequent spatial and temporal analysis. **OK** method is mainly applied for datasets without a trend of the *unknown mean* for medium size areas, in which assuming constant unknown mean $a_0 = \mathbf{E}[\boldsymbol{x}]$ when predicting $Z(\boldsymbol{s})$ at unsampled places is still reasonable. If the study area is considerably large, we can employ the **universal kriging** (**UK**, see [6] for more), assuming a trend in the mean over the spatial region D in our prediction.

3.1 Kriging estimator

Kriging in the simplest case is a problem of point estimation. Select any place $\boldsymbol{s}_0 \in D$, our sampling space of interest. We want to estimate $Z_0 = Z(\boldsymbol{s}_0)$ from n observations $Z(\boldsymbol{s}_i)$ using the general affine (kriging) estimator, given by the following equation

$$Z^* = \hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i \; Z(\boldsymbol{s}_i) + \lambda_0 \tag{3.1}$$

where $Z^* = \hat{Z}(\boldsymbol{s}_0)$ is the kriged (estimated) value at place $\boldsymbol{s}_0$, $Z(\boldsymbol{s}_i)$ is the observed value at place $\boldsymbol{s}_i$, $w_i \geq 0$ is the non-negative weight associated with that observation, and λ_0 is said to be uncertainty (nuisance, uncontrollable) constant from the global environment. The constant λ_0 and w_i are selected so as to minimize the expected **m**ean **s**quare **e**rror (mse) $\mathbf{E}[(Z^* - Z_0)^2]$.

In the stationary case, the variance of a linear combination $\sum\limits_{i=1}^{n} w_i \; Z(\boldsymbol{s}_i)$ can be expressed via the variograms by

$$\mathbf{V}\left[\sum_{i=1}^{n} w_i \; Z(\boldsymbol{s}_i)\right] = -\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(\boldsymbol{s}_j - \boldsymbol{s}_i). \tag{3.2}$$

Look back to the affine estimator Z^* given in Equation (3.1), its mse can be written as

$$\mathbf{E}[(Z^* - Z_0)^2] = \mathbf{V}[Z^* - Z_0] + \text{Bias}(Z^*)^2 \tag{3.3}$$

where

$$\text{Bias}(Z^*) = \mathbf{E}[Z^* - Z_0] = \lambda_0 + \left(\sum_i w_i - 1\right) a_0.$$

To achieve unbiased estimations in **OK**, in which $\text{Bias}(Z^*) = 0$, for whatever the unknown mean a_0 is, we have to set $\lambda_0 = 0$, additionally require the convex condition $\sum\limits_{i=1}^{n} w_i = 1$. Then, geometrically $\boldsymbol{s}_0$ is in the convex hull $\mathbf{CH}(S)$ of S, given as

$$\mathbf{CH}(S) = \left\{\boldsymbol{x} \in \mathbb{R}^n : \; \boldsymbol{x} = \sum_i \; w_i \; \boldsymbol{s}_i, \text{ where } \; \sum_i \; w_i = 1, \; w_i \geq 0, \; \boldsymbol{s}_i \in S\right\}. \tag{3.4}$$

3.2 Unbiased ordinary kriging estimator

Our problem can now be reformulated as follows: Find n weights w_i summing to 1 and minimizing

$$\mathbf{V}[Z^* - Z_0] = \sum_i \sum_j w_i w_j \; \sigma_{ij} - 2\sum_{i=1}^{n} \; w_i \sigma_{i0} + \sigma_{00}, \tag{3.5}$$

where $\sigma_{ij} = \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$. This is solved by the method of Lagrange multipliers, with η the Lagrange multiplier; employing the function

$$Q = \mathbf{V}[Z^* - Z_0] + 2\eta \left(\sum_{i=1}^{n} w_i - 1 \right)$$

from which we solve the system $\dfrac{\partial Q}{\partial w_1} = \cdots = \dfrac{\partial Q}{\partial w_n} = \dfrac{\partial Q}{\partial \eta} = 0$ to get the OK variance

$$\sigma^2_{OK} = \mathbf{E}[(Z^* - Z_0)^2] = \sigma_{00} - \sum_{i=1}^{n} w_i \sigma_{i0} \ - \eta.$$

Its variance can be calculated with the variograms, by Equation 3.2 and with $\sigma_{ij} := \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ we get the following system of *kriging equations*:

$$\begin{cases} \sum\limits_{j=1}^{n} w_j \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) - \eta = \gamma(\boldsymbol{s}_i, \boldsymbol{s}_0), \ i = 1, \ldots, n; \\ \sum\limits_{i=1}^{n} w_i = 1, \ w_i \geq 0, \end{cases} \tag{3.6}$$

where $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_0)$ is the value of the variogram between the ith sampling data point $\boldsymbol{s}_i$ and the target point $\boldsymbol{s}_0$, and $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the value of variogram between the points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. The kriging variance for unbiased OK is

$$\sigma^2_{OK} = \sum_{i=1}^{n} w_i \gamma(\boldsymbol{s}_i, \boldsymbol{s}_0) - \eta.$$

The linear system (3.6) has a unique solution if and only if the covariance matrix $\Sigma = [\sigma_{ij}]$ is strictly positive definite, which is the case if we use a strictly positive definite covariance function model and if all data points are distinct (Jean Paul [6, Chapter 3]).

4 Cokriging approach

4.1 Co-kriging predictor and the MSE of prediction

Since our spatial-temporal air pollution data is multivariate by nature, see Section 1.3, we might use multiple parameters to exploit their relationships. We can estimate certain parameters, and use relevant information of other parameters, employ ingredients recalled in Appendix A. Cokriging is an extension of ordinary kriging, [of a single variable given in Eqn. (3.1)], in which it takes into account additional correlated information in the subsidiary variables.

In cokriging technique, suppose that at each spatial location $\boldsymbol{s}_i$ we observe $k > 1$ variables Z_j, summarized in a data matrix M:

$$\begin{array}{l} Z_1(\boldsymbol{s}_1)\ Z_1(\boldsymbol{s}_2)\cdots Z_1(\boldsymbol{s}_i)\cdots Z_1(\boldsymbol{s}_n), \\ \cdots \quad \cdots \quad \cdots \\ Z_j(\boldsymbol{s}_1)\ Z_j(\boldsymbol{s}_2)\cdots Z_j(\boldsymbol{s}_i)\cdots Z_j(\boldsymbol{s}_n), \\ \cdots \quad \cdots \quad \cdots \\ Z_k(\boldsymbol{s}_1)\ Z_k(\boldsymbol{s}_2)\cdots Z_k(\boldsymbol{s}_i)\cdots Z_k(\boldsymbol{s}_n), \end{array} \tag{4.1}$$

for $j = 1, 2, \cdots, k$, and $i = 1, 2, \cdots, n$.

We want to predict $Z_1(\boldsymbol{s}_0)$, the value of variable Z_1 at unobserved location $\boldsymbol{s}_0 \in \mathbf{CH}(S)$. Given the fact that the variable under consideration (*the target variable* Z_1) occurs with other variables (called *co-located variables*), we explore the possibility of improving the prediction of variable Z_1 by taking into account the correlation of Z_1 with the other variables.

Definition 3 (The cokriging predictor).

The cokriging predictor takes the form

$$\begin{aligned} \hat{Z}_1(\boldsymbol{s}_0) = \sum_{j=1}^{k}\sum_{i=1}^{n} w_{ji} Z_j(\boldsymbol{s}_i) = w_{11}\ Z_1(\boldsymbol{s}_1) + \cdots + w_{1n}\ Z_1(\boldsymbol{s}_n) + \cdots \\ + w_{k1}\ Z_k(\boldsymbol{s}_1) + \cdots + w_{kn}\ Z_k(\boldsymbol{s}_n) \end{aligned} \tag{4.2}$$

We see that there are weights associated with variable Z_1 but also with each one of the other variables. For instance, to our observed multivariate data, with chemical factors of PM_{10}, SO_2, NO_2, CO, O_3, TSP and benzen we could set max $k = 7$. Assume that the whole sampling area is rather homogeneous, i.e. distinct sampling points $\boldsymbol{s}_i$ have different values $Z_j(\boldsymbol{s}_i)$ but their expectation are the same, we denote $\mu_j = \mathbf{E}[Z_j(\boldsymbol{s}_i)] = \mathbf{E}[Z_j(\boldsymbol{s})]$, for each $j = 1, \cdots, k$; for all $i = 1, \cdots, n$, and for any sampling point $\boldsymbol{s} \in D$.

We will examine *ordinary co-kriging* (the extension of ordinary kriging of a single variable to two or more variables). The expectation vector of k variables Z_j then is

$$\mathbf{E}[\mathbf{Z}(\boldsymbol{s})] = \begin{pmatrix} \mathbf{E}[Z_1(\boldsymbol{s})] \\ \mathbf{E}[Z_2(\boldsymbol{s})] \\ \vdots \\ \mathbf{E}[Z_k(\boldsymbol{s})] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \mu$$

We want the predictor $\widehat{Z}_1(\boldsymbol{s}_0)$ to be unbiased, that is $\mathbf{E}[\hat{Z}_1(\boldsymbol{s}_0)] = \mu_1$, where

$$\begin{aligned}\mathbf{E}[\hat{Z}_1(\boldsymbol{s}_0)] &= \sum_{j=1}^{k}\sum_{i=1}^{n} w_{ji}\mathbf{E}[Z_j(\boldsymbol{s}_i)] = \sum_{i=1}^{n} w_{1i}\ \mu_1 + \cdots + \sum_{i=1}^{n} w_{ki}\ \mu_k \\ &= w_{11}\mathbf{E}[Z_1(\boldsymbol{s}_1)] + \cdots + w_{1n}\mathbf{E}[Z_1(\boldsymbol{s}_n)] + \ \cdots \\ &+ w_{k1}\mathbf{E}[Z_k(\boldsymbol{s}_1)] + \cdots + w_{kn}\mathbf{E}[Z_k(\boldsymbol{s}_n)].\end{aligned} \tag{4.3}$$

Definition 4. *The mean squared error (MSE) of prediction of Z_1 is given by*

$$\mathbf{E}\big[\{Z_1(\boldsymbol{s}_0) - \hat{Z}_1(\boldsymbol{s}_0)\}^2\big] = \sigma_e^2. \tag{4.4}$$

Therefore, to get $\mathbf{E}[\hat{Z}_1(\boldsymbol{s}_0)] = \mu_1$ we must have the followings

$$\sum_{i=1}^{n} w_{1i} = 1,\ \sum_{i=1}^{n} w_{2i} = 0,\ \cdots,\ \sum_{i=1}^{n} w_{ki} = 0 \tag{4.5}$$

As with the other forms of kriging, co-kriging minimizes the MSE, with certain conditions:

$$\begin{aligned}&min\ \sigma_e^2 = \mathbf{E}\big[\{Z_1(\boldsymbol{s}_0) - \hat{Z}_1(\boldsymbol{s}_0)\}^2\big] = \mathbf{E}\big[\{Z_1(\boldsymbol{s}_0) - \sum_{j=1}^{k}\sum_{i=1}^{n} w_{ji}Z_j(\boldsymbol{s}_i)\}^2\big], \\ &\text{subject to } \sum_{i=1}^{n} w_{1i} = 1,\ \sum_{i=1}^{n} w_{2i} = 0,\ \cdots,\ \sum_{i=1}^{n} w_{ki} = 0\end{aligned}$$

4.2 Co-kriging: the case of two pollutants

Let's assume $k = 2$, in other words, we observe variables Z_1 and Z_2 (e.g. PM_{10} and benzen in our sample data) and we want to predict $Z = Z_1$.

Lemma 4.1. *The variance $\sigma_e^2 = \mathbf{E}\big[\{Z_1(\boldsymbol{s}_0) - \hat{Z}_1(\boldsymbol{s}_0)\}^2\big]$ has explicit expansion*

$$\begin{aligned}\sigma_e^2 = \mathbf{E}[\{Z(\boldsymbol{s}_0) - \mu_1\}^2] &- 2\sum_{i=1}^{n} w_{1i}\mathbf{E}[Z_1(\boldsymbol{s}_0) - \mu_1][Z_1(\boldsymbol{s}_i) - \mu_1] \\ &- 2\sum_{i=1}^{n} w_{2i}\mathbf{E}[Z_1(\boldsymbol{s}_0) - \mu_1][Z_2(\boldsymbol{s}_i) - \mu_2] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{1j}\mathbf{E}[Z_1(\boldsymbol{s}_i) - \mu_1][Z_1(\boldsymbol{s}_j) - \mu_1] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} w_{2i}w_{2j}\mathbf{E}[Z_2(\boldsymbol{s}_i) - \mu_2][Z_2(\boldsymbol{s}_j) - \mu_2] \\ &+ 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{2j}\mathbf{E}[Z_1(\boldsymbol{s}_i) - \mu_1][Z_2(\boldsymbol{s}_j) - \mu_2]\end{aligned} \tag{4.6}$$

Proof. From the constraints (4.5), we have $0 = \sum_{i=1}^{n} w_{2i} = \sum_{i=1}^{n} w_{2i}\mu_2$. We rewrite

$$\sigma_e^2 = \mathbf{E}\big[\{Z(\boldsymbol{s}_0) - \sum_{i=1}^{n} w_{1i}Z_1(\boldsymbol{s}_i) - \sum_{i=1}^{n} w_{2i}Z_2(\boldsymbol{s}_i)\}^2\big].$$

Let's add the quantity of $(-\mu_1 + \mu_1 + \sum_{i=1}^{n} w_{2i}\ \mu_2)$ into the variance, where $\mu_1 = \mathbf{E}[\hat{Z}_1(\boldsymbol{s}_0)]$ and $\mu_2 = \ \mathbf{E}[Z_2(\boldsymbol{s})]$, we see that

$$\begin{aligned}\sigma_e^2 &= \mathbf{E}\big[\{Z(\boldsymbol{s}_0) - \sum_{i=1}^{n} w_{1i}Z_1(\boldsymbol{s}_i) - \sum_{i=1}^{n} w_{2i}Z_2(\boldsymbol{s}_i) - \ \mu_1 + \mu_1 + \sum_{i=1}^{n} w_{2i}\mu_2\}^2\big] \\ &= \mathbf{E}\big[\{(Z(\boldsymbol{s}_0) - \mu_1) - \sum_{i=1}^{n} w_{1i}[Z_1(\boldsymbol{s}_i) - \mu_1] - \ \sum_{i=1}^{n} w_{2i}[Z_2(\boldsymbol{s}_i) - \mu_2]\}^2\big]\end{aligned}$$

Expanding the core term of the expectation above we get:

$$\begin{aligned}&[Z(\boldsymbol{s}_0) - \mu_1]^2 - 2\sum_{i=1}^{n} w_{1i}[Z_1(\boldsymbol{s}_0) - \mu_1][Z_1(\boldsymbol{s}_i) - \mu_1] - 2\sum_{i=1}^{n} w_{2i}[Z_1(\boldsymbol{s}_0) - \mu_1][Z_2(\boldsymbol{s}_i) - \mu_2] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{1j}[Z_1(\boldsymbol{s}_i) - \mu_1][Z_1(\boldsymbol{s}_j) - \mu_1] + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{2i}w_{2j}[Z_2(\boldsymbol{s}_i) - \mu_2][Z_2(\boldsymbol{s}_j) - \mu_2] \\ &+ 2\Big(\sum_{i=1}^{n} w_{1i}[Z_1(\boldsymbol{s}_i) - \mu_1]\Big)\Big(\sum_{i=1}^{n} w_{2i}[Z_2(\boldsymbol{s}_i) - \mu_2]\Big)\end{aligned} \tag{4.7}$$

where the last term is reduced to $2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{2j}[Z_1(\boldsymbol{s}_i) - \mu_1]\ [Z_2(\boldsymbol{s}_j) - \mu_2]$. □

We next consider the following optimization model

$$min\ \sigma_e^2 = \mathbf{E}\big[\{Z(\boldsymbol{s}_0) - \sum_{i=1}^{n} w_{1i}Z_1(\boldsymbol{s}_i) - \sum_{i=1}^{n} w_{2i}Z_2(\boldsymbol{s}_i)\}^2\big]$$

Theorem 4.2. *The above optimization model can be transformed to the following cokriging system of linear equations*

$$\boldsymbol{G}\,\boldsymbol{w} = \boldsymbol{c} \tag{4.8}$$

where the vector $\boldsymbol{w}, \boldsymbol{c}$ *have dimensions* $(2n+2) \times 1$ *and the matrix* $\boldsymbol{G}$ *has dimensions* $(2n+2) \times (2n+2)$. *The optimal weights will be obtained as*

$$\boldsymbol{w} = \boldsymbol{G}^{-1}\,\boldsymbol{c}.$$

Proof. We now minimize $\sigma_e^2 = \mathbf{E}\big[\{Z(\boldsymbol{s}_0) - \sum_{i=1}^{n} w_{1i} Z_1(\boldsymbol{s}_i) - \sum_{i=1}^{n} w_{2i} Z_2(\boldsymbol{s}_i)\}^2\big]$ as

$$
\begin{aligned}
&min\ \mathbf{E}[\{Z(\boldsymbol{s}_0) - \mu_1\}^2] - 2\sum_{i=1}^{n} w_{1i}\mathbf{E}[Z_1(\boldsymbol{s}_0) - \mu_1][Z_1(\boldsymbol{s}_i) - \mu_1] \\
&- 2\sum_{i=1}^{n} w_{2i}\mathbf{E}[Z_1(\boldsymbol{s}_0) - \mu_1][Z_2(\boldsymbol{s}_i) - \mu_2] + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{1j}\mathbf{E}[Z_1(\boldsymbol{s}_i) - \mu_1][Z_1(\boldsymbol{s}_j) - \mu_1] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} w_{2i}w_{2j}\mathbf{E}[Z_2(\boldsymbol{s}_i) - \mu_2][Z_2(\boldsymbol{s}_j) - \mu_2] + 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{2j}\mathbf{E}[Z_1(\boldsymbol{s}_i) - \mu_1][Z_2(\boldsymbol{s}_j) - \mu_2]
\end{aligned}
\tag{4.9}
$$

Denote the covariances involving variable Z_i by $\mathbf{C}_{ii} = \sigma_i^2$, the cross-covariance between variables Z_i, Z_j by $\mathbf{C}_{ij}$. The covariances and the cross-covariances are

$$
\begin{aligned}
&\mathbf{C}[Z_1(\boldsymbol{s}_0), Z_1(\boldsymbol{s}_0)] = \mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_0) = \mathbf{C}_{11}(0) = \sigma_1^2 \\
&\mathbf{C}[Z_1(\boldsymbol{s}_0), Z_1(\boldsymbol{s}_i)] = \mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_i),\ \mathbf{C}[Z_1(\boldsymbol{s}_i), Z_1(\boldsymbol{s}_j)] = \mathbf{C}_{11}(\boldsymbol{s}_i, \boldsymbol{s}_j) \\
&\mathbf{C}[Z_1(\boldsymbol{s}_i), Z_2(\boldsymbol{s}_j)] = \mathbf{C}_{12}(\boldsymbol{s}_i, \boldsymbol{s}_j),\ \mathbf{C}[Z_1(\boldsymbol{s}_0), Z_2(\boldsymbol{s}_j)] = \mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_j) \\
&\mathbf{C}[Z_2(\boldsymbol{s}_i), Z_1(\boldsymbol{s}_j)] = \mathbf{C}_{21}(\boldsymbol{s}_i, \boldsymbol{s}_j),\ \mathbf{C}[Z_2(\boldsymbol{s}_i), Z_2(\boldsymbol{s}_j)] = \mathbf{C}_{22}(\boldsymbol{s}_i, \boldsymbol{s}_j).
\end{aligned}
\tag{4.10}
$$

Finally, with the Lagrange multipliers we get:

$$
\begin{aligned}
&min\ \sigma_1^2 - 2\sum_{i=1}^{n} w_{1i}\mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_i) - 2\sum_{i=1}^{n} w_{2i}\mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_i) \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{1j}\mathbf{C}_{11}(\boldsymbol{s}_i, \boldsymbol{s}_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{2i}w_{2j}\mathbf{C}_{22}(\boldsymbol{s}_i, \boldsymbol{s}_j) \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{2j}\mathbf{C}_{12}(\boldsymbol{s}_i, \boldsymbol{s}_j) - 2\lambda_1\left[\sum_{i=1}^{n} w_{1i} - 1\right] - 2\lambda_2\left[\sum_{i=1}^{n} w_{2i} - 0\right].
\end{aligned}
\tag{4.11}
$$

The unknowns are $w_{11}, ..., w_{1n}$; $w_{21}, ..., w_{2n}$; and the two Lagrange multipliers λ_1 and λ_2. Take derivatives with respect to these unknowns and set them equal to zero. For every $i = 1, \ldots, n$, we have $\sum_{j=1}^{n} w_{1j} = 1$, $\sum_{j=1}^{n} w_{2j} = 0$ and

$$
-2\mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_i) + 2\sum_{j=1}^{n} w_{1j}\mathbf{C}_{11}(\boldsymbol{s}_i, \boldsymbol{s}_j) + 2\sum_{j=1}^{n} w_{2j}\mathbf{C}_{12}(\boldsymbol{s}_i, \boldsymbol{s}_j) - 2\lambda_1 = 0, \tag{4.12}
$$

$$
-2\mathbf{C}_{12}(s_0, s_i) + 2\sum_{j=1}^{n} w_{2j}\mathbf{C}_{22}(\boldsymbol{s}_i, \boldsymbol{s}_j) + 2\sum_{j=1}^{n} w_{1j}\mathbf{C}_{21}(\boldsymbol{s}_i, \boldsymbol{s}_j) - 2\lambda_2 = 0. \tag{4.13}
$$

Recall that $\boldsymbol{s}_i$ $(i = 1, 2, \ldots, n)$ are sampling places, denote square matrices

$$[\mathbf{C}_{11}] = \begin{pmatrix} \mathbf{C}_{11}(\boldsymbol{s}_1, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{11}(\boldsymbol{s}_1, \boldsymbol{s}_n) \\ \vdots & \vdots & \vdots \\ \mathbf{C}_{11}(\boldsymbol{s}_n, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{11}(\boldsymbol{s}_n, \boldsymbol{s}_n) \end{pmatrix}; \; [\mathbf{C}_{12}] = \begin{pmatrix} \mathbf{C}_{12}(\boldsymbol{s}_1, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{12}(\boldsymbol{s}_1, \boldsymbol{s}_n) \\ \vdots & \vdots & \vdots \\ \mathbf{C}_{12}(\boldsymbol{s}_n, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{12}(\boldsymbol{s}_n, \boldsymbol{s}_n) \end{pmatrix}$$

$$[\mathbf{C}_{21}] = \begin{pmatrix} \mathbf{C}_{21}(\boldsymbol{s}_1, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{21}(\boldsymbol{s}_1, \boldsymbol{s}_n) \\ \vdots & \vdots & \vdots \\ \mathbf{C}_{21}(\boldsymbol{s}_n, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{21}(\boldsymbol{s}_n, \boldsymbol{s}_n) \end{pmatrix}; \; [\mathbf{C}_{22}] = \begin{pmatrix} \mathbf{C}_{22}(\boldsymbol{s}_1, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{22}(\boldsymbol{s}_1, \boldsymbol{s}_n) \\ \vdots & \vdots & \vdots \\ \mathbf{C}_{22}(\boldsymbol{s}_n, \boldsymbol{s}_1) & \cdots & \mathbf{C}_{22}(\boldsymbol{s}_n, \boldsymbol{s}_n) \end{pmatrix}$$

and vectors of $2n + 2$ entries

$$[\mathbf{1}] = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}; \; [\mathbf{0}] = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; \; W_1 = \begin{pmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1n} \end{pmatrix}; \; W_2 = \begin{pmatrix} w_{21} \\ w_{22} \\ \vdots \\ w_{2n} \end{pmatrix};$$

$$[\mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_i)] = \begin{pmatrix} \mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_1) \\ \vdots \\ \mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_n) \end{pmatrix}; \; [\mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_i)] = \begin{pmatrix} \mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_1) \\ \vdots \\ \mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_n) \end{pmatrix}.$$

The co-kriging system in matrix form is

$$\begin{pmatrix} [\mathbf{C}_{11}] & [\mathbf{C}_{12}] & [\mathbf{1}] & [\mathbf{0}] \\ [\mathbf{C}_{21}] & [\mathbf{C}_{22}] & [\mathbf{0}] & [\mathbf{1}] \\ [\mathbf{1}]' & [\mathbf{0}]' & 0 & 0 \\ [\mathbf{0}]' & [\mathbf{1}]' & 0 & 0 \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ -\lambda_1 \\ -\lambda_2 \end{pmatrix} = \begin{pmatrix} [\mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_i)] \\ [\mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_i)] \\ 1 \\ 0 \end{pmatrix}$$

Put

$$\mathbf{G} = \begin{pmatrix} [\mathbf{C}_{11}] & [\mathbf{C}_{12}] & [\mathbf{1}] & [\mathbf{0}] \\ [\mathbf{C}_{21}] & [\mathbf{C}_{22}] & [\mathbf{0}] & [\mathbf{1}] \\ [\mathbf{1}]' & [\mathbf{0}]' & 0 & 0 \\ [\mathbf{0}]' & [\mathbf{1}]' & 0 & 0 \end{pmatrix}; \mathbf{w} = \begin{pmatrix} W_1 \\ W_2 \\ -\lambda_1 \\ -\lambda_2 \end{pmatrix}; \quad \mathbf{c} = \begin{pmatrix} [\mathbf{C}_{11}(\boldsymbol{s}_0, \boldsymbol{s}_i)] \\ [\mathbf{C}_{12}(\boldsymbol{s}_0, \boldsymbol{s}_i)] \\ 1 \\ 0 \end{pmatrix}$$

we have

$$\mathbf{G}\mathbf{w} = \mathbf{c} \tag{4.14}$$

here $\mathbf{C}_{12}(\boldsymbol{h})$ may not be the same as $\mathbf{C}_{21}(\boldsymbol{h})$, $|h| = |\boldsymbol{s}_i - \boldsymbol{s}_j|$ for $i, j = 1, 2, ..., n$; the vector $\mathbf{w}, \mathbf{c}$ have dimensions $(2n + 2) \times 1$ and the matrix $\mathbf{G}$ has dimensions $(2n + 2) \times (2n + 2)$. This cokriging system gives us the optimal weights

$$\mathbf{w} = \mathbf{G}^{-1} \, \mathbf{c}.$$

If G is not invertible, use its generalized inverse. □

5 Algorithm for coping with missing data points

5.1 What if stations give incomplete or missed data?

However, we got only $m = 6$ good data points, more than 33% of monitoring sities (3/9: stations of Hong Bang, Quang Trung and Thu Duc) do not provide

any numerical information for model fitting. Aiming to increase the accuracy of prediction, it is useful to discuss few ways for handling this matter. The main idea of the linear kriging in (3.1) is that the predicted value

$$\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{6} w_i \; Z(\boldsymbol{s}_i) \tag{5.1}$$

at certain unknown point s_0 infact is just a convex combiation of six observed value $Z(s_i)$ at $m = 6$ monitoring points, where the weights w_i fulfill $\sum\limits_{i=1}^{m} w_i = 1$ and $w_i \geq 0$.

Basic techniques like *imputation* type algorithms [12] can be employed to fill in numerical values of PM_{10} and benzen to three defective stations before exploiting the estimator

$$\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{9} w_i \; Z(\boldsymbol{s}_i); \; \sum_{i=1}^{9} w_i = 1 \text{ and } w_i \geq 0. \tag{5.2}$$

With sampling mechanism for our data sets, the monitored data belongs to the class of *missing completely at random* (MCAR), i.e. the probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained or the set of observed responses. The essential feature of MCAR is that the observed data can be viewed as a random sample of a complete data. However, due to the disticntive nature of our dataset (already possessing MCAR), the well known imputation methods seem to be not suitable.

We suggest a data imputation mechanism, named **progressively co-kriging imputation**, to predict values at a single unobservable station (from data of good stations), then recursively extend (enlarge) prediction at the remaining unobservable stations, since more available info the more precise outcomes we get. First, we need to decode the cokriging system, as given in Equation 4.8. If use m good stations ($m \leq n$, the total number of originally designed stations) to predict value at unmeasured station in our case study, the original cokriging system (4.8) could be more explicitly written as

$$\mathbf{G}\,\mathbf{w} = \mathbf{c}, \tag{5.3}$$

where the vectors **w** and **c** currently have dimensions $(2m + 2) \times 1$ and the matrix **G** has dimensions $(2m + 2) \times (2m + 2)$. The explicit forms of **G**, **w** and **c** in the next Algorithm 1 are given by

$$\mathbf{G} = \begin{pmatrix} \mathbf{C}_{11}(s_1,s_1)\cdots\mathbf{C}_{11}(s_1,s_m)\ \mathbf{C}_{12}(s_1,s_1)\cdots\mathbf{C}_{12}(s_1,s_m)\ 1\ 0 \\ \vdots \qquad\qquad \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad \vdots\vdots \\ \mathbf{C}_{11}(s_m,s_1)...\mathbf{C}_{11}(s_m,s_m)\ \mathbf{C}_{12}(s_m,s_1)...\mathbf{C}_{12}(s_m,s_m)\ 1\ 0 \\ \mathbf{C}_{21}(s_1,s_1)\cdots\mathbf{C}_{21}(s_1,s_m)\ \mathbf{C}_{22}(s_1,s_1)\cdots\mathbf{C}_{22}(s_1,s_m)\ 0\ 1 \\ \vdots \qquad\qquad \vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad \vdots\vdots \\ \mathbf{C}_{21}(s_m,s_1)...\mathbf{C}_{21}(s_m,s_m)\ \mathbf{C}_{22}(s_m,s_1)...\mathbf{C}_{22}(s_m,s_m)\ 0\ 1 \\ 1 \quad \cdots \quad 1 \quad 0 \quad \cdots \quad 0 \quad 0\ 0 \\ 0 \quad \cdots \quad 0 \quad 1 \quad \cdots \quad 1 \quad 0\ 0 \end{pmatrix},$$

$$\mathbf{w} = \begin{pmatrix} w_{11} \\ \vdots \\ w_{1m} \\ w_{21} \\ \vdots \\ w_{2m} \\ -\lambda_1 \\ -\lambda_2 \end{pmatrix}; \text{ and } \mathbf{c} = \begin{pmatrix} \mathbf{C}_{11}(s_0,s_1) \\ \vdots \\ \mathbf{C}_{11}(s_0,s_m) \\ \mathbf{C}_{12}(s_0,s_1) \\ \vdots \\ \mathbf{C}_{12}(s_0,s_m) \\ 1 \\ 0 \end{pmatrix}.$$

Our algorithm of progressively imputation allowing us to get possibly accurate predicted observation value at arbitrary sites in the area of study is given below.

Algorithm 1 Progressively Co-kriging Imputation

INPUT: a finite set V_0 of all n designed monitoring sites,
$V = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_m\}$ of observable sites ($m \le n$), $V \subseteq V_0$,
a $k \times m$ data matrix M of k observable factors (monitored at m sites), described in matrix (4.1)
OUTPUT: the fully updated set V of n known sites with available data, from which observation $Z(\boldsymbol{s})$ at any location $\boldsymbol{s} \in \mathbf{CH}(V)$ can be estimated by Equation (5.1) and the system (5.3); [$\mathbf{CH}(V)$ is defined by Equation (3.4)].

If $m = n$ stop, else proceed to next step.
while $m < n$ **do**
1. Select a a site $\boldsymbol{s}_0 \in V_0 \setminus V$ (an unmeasured site)
2. Set up the system (4.14) from data M and $\boldsymbol{s}_0$
3. Compute the weight vector $\mathbf{w} = \mathbf{G}^{-1}\ \mathbf{c}$
4. Calculate the predicted value at the site $\boldsymbol{s}_0$
5. Update $V := V \cup \{\boldsymbol{s}_0\}$; update $m = |V|$; update data matrix M;

end while
return The full network V of all observable sites.

5.2 Criteria for selecting the best-fit model

We limited the scope of this study to focus on analyzing these four models (spherical, exponential, linear, and Gaussian), described by Equations (6.1), (6.2), (6.3), and (6.4).

A critical question we need to answer now is "what model is most suitable for modeling our sample data"? To answer this, we decide to construct potential variogram models based on different parameters and compare them, and then select the best-ft one for our analysis. The model that uses interpolation is called *optimal* if it has the *lowest error forecast*. However, there are other criteria for assessing whether the forecasting model is good or not. Few following statistics (integrated in GS+) can be used to explain the output of the model.

- First, the **residual sum of squares** (RSS) is the sum of the squares of deviations predicted from empirical values. A small RSS indicates a tight fit of the model to the data.

- Second, the **coefficient of determination**, r^2, is the proportion of the variance in the dependent variable that is predictable from the independent variables. This value is not a strong criterion for fitting the model as RSS, but used to look at the impact of change in the model parameters.

- Third, the **proportion C/(Co+C)** statistic provides a measure of the proportion of sample variance (C_0+C) that is explained by spatially structured variance C. This value will be 1.0 for a variogram with no nugget variance (where the curve passes through the origin).

  Conversely, it will be 0 where there is no spatially dependent variation at the range specified, i.e. where there is a pure nugget effect.

5.3 Choosing suitable variograms for specific stations

The variogram values are presented in Table 4, where the Gaussian model returns the highest $r^2 = 0.866, Residual\ sum\ of\ squares(RSS) = 0.007092$. Thus, from the stations $s_1, s_2, s_3, s_4, s_5, s_6$ we find the best interpolation model as shown in Table 4, based on RSS, r^2 and $C/(C_0 + C)$. Using the models found, we will forecast for three stations Hong Bang, Quang Trung and Thu Duc, where missing data occurred.

At each of 3 missed data stations, we check assumptions (nudget and sill) of the four mentioned models to see their applicability, then estimate the Z value correspondingly with the selected variogram. Next we enlarge the interpolation, paying attention to the RSS, the r^2 and the $C/(C_0 + C)$ in Table 4. Apply specifically the suggested cokriging predictor [Formula 4.2] for our dataset with PM_{10} and benzen ($k = 2$), we predict PM_{10} at an unobservable station, say **Hong Bang**, next use current information of $m := m + 1 = 7$ stations,

repeat procedure until $m = n = 9$ (progressively fill in values for Quang Trung, Thu Duc), then obtain estimated value for the last unmeasured station.

Table 4: Isotropic variograms values of PM_{10}, benzen and two parameters

	Estimates	of	parameters			
Data set	**Nugget**	**Sill**	**Range**	RSS^a	r^2	$\frac{C}{C_0+C)}$
and model	(m) C_0	(m) C_0+C	(m) A			
PM_{10} $(n=6)$						
Linear	0	0	485	3.96	0.613	-5.508
Gaussian	**0**	**1.487**	**27107**	**0.846**	**0.650**	**0.999**
Spherical	0	1.56	41100	1.07	0.615	0.999
Exponential	0	2	91260	1.14	0.606	1
Benzen $(n=6)$						
Linear	0	0	466	0.00100	0.869	-33.498
Gaussian	**0**	**0.075**	**29341**	$\mathbf{3.71}\times10^{-5}$	**0.871**	**0.972**
Spherical	0	0.046	31100	5.724×10^{-5}	0.865	0.998
Exponential	0	0.0765	93300	6.412×10^{-5}	0.861	0.999
PM_{10} and benzen $n=6$						
Linear	0	0	104225	0.060100	0.736	-3.281
Gaussian	**0**	**0.86**	**71187**	**0.007092**	**0.866**	**0.999**
Spherical	0	0.189	41100	0.018000	0.708	0.999
Exponential	0	0.316	123300	0.019300	0.687	1

RSS^a is the sum of squares of the residuals from the fitted function.

5.4 Find the estimated prediction variance at sampling points

We multiply Eq. (4.12) by w_{1i}; Eq. (4.13) by w_{2i} and sum over $i = 1..n$, to get:

$$-\sum_{i=1}^{n} w_{1i}\mathbf{C}_{11}(s_0,s_i)+\sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{1j}\mathbf{C}_{11}(s_i,s_j)+\sum_{i=1}^{n}\sum_{j=1}^{n} w_{1i}w_{2j}\mathbf{C}_{12}(s_i,s_j)-\sum_{i=1}^{n} w_{1i}\lambda_1=0 \tag{5.4}$$

$$-\sum_{i=1}^{n} w_{2i}\mathbf{C}_{12}(s_0,s_i)+\sum_{i=1}^{n}\sum_{j=1}^{n} w_{2i}w_{2j}\mathbf{C}_{22}(s_i,s_j)+\sum_{i=1}^{n}\sum_{j=1}^{n} w_{2i}w_{1j}\mathbf{C}_{21}(s_i,s_j)-\sum_{i=1}^{n} w_{2i}\lambda_2=0 \tag{5.5}$$

To simplify the expression for the variance of the predicted value we substitute (5.4) and (5.5) into Eqn. (4.11) (with $k = 2$):

$$\widehat{\sigma}_1^2=\mathbf{C}_{11}(0)-\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ji}\mathbf{C}_{11}(s_0,s_i)+\lambda_1 \tag{5.6}$$

Due to Table 4, the Gaussian variogram (6.4) is best suited, hence used for prediction at all three missing data stations. Table 5 then shows comparison of PM_{10} outcomes when using the geo-statistical software **GS+** (by default) and the general statistical software **R**, implemented by our co-kriging imputation algorithm. More precisely, the predicted value of PM_{10} for Hong Bang station

at location $s_7(693640, 1199790)$ is $Y = 138.83\ \mu g/m^3$ (by Eqn. 4.2), as a result, the standard deviation $SD(Y) = \widehat{\sigma}_1 = 0.514$; predicted values at Quang Trung - $s_8(677940, 1200080)$ and Thu Duc - $s_9(693640, 1199790)$ are also shown.

Table 5: Comparison of PM_{10} outcomes when using **GS+** and package **R**.

Station	Station name	Realistic value	Predicted value by GS+		Predicted value by **R**	
			$Y = PM_{10}$	$SD(Y)$	Y	$SD(Y)$
s_1	Thong Nhat (TN)	65.99				
s_2	Binh Chanh (BC)	17.48				
s_3	ZOO	73.14				
s_4	Doste (DOSTE)	123.5				
s_5	District 2 (D2)	73.31				
s_6	Tan Son Hoa (TSH)	67.33				
s_7	**Hong Bang (HB)**	NA	137.14	1.98	138.83	0.514
s_8	Quang Trung (QT)	NA	131.43	1.81	133.83	0.096
s_9	Thu Duc (TD)	NA	NA	NA	140.17	0.807

6 Conclusion and future work

For this specific data, we applied Kriging-based interpolation methods to predict only key pollutant parameters of air pollution, dropping off other important parameters such as temperature, humidity, wind, cloud cover, height of site ... which are key factors for Gaussian class of dispersion models.

In summary, our approach has solved the missing data at monitoring stations, and clearly produces a more precise prediction than using the default approach via the popular soft GS+, see [8]. The work's contributions, in more details, include exploiting cokriging models in air pollution study in HCMC. The cokriging approach, elucidated in Section 4.1, captures correlation of pollutants in a powerful identity (4.8), allowing us to figure out value at any location in the convex hull of observable locations. Secondly, our algorithm - formulated in Section 5 - can handle of missing data in large scale, which haven't been considered in other studies, up to the time of preparing this paper, for instance, see Pham and Doan's works [1, 17].

Our future work would possibly be conducting the spatial analysis to show interaction effects of pollutants on prediction, and possibly investigating cokriging models with $k > 2$ covariates, say PM_{10}, SO_2, NO_2 and benzen, since interactions of more than two factors could also potentially impact on predicted values of cokriging models. This is meaningful since the software GS+ is currently not able to obtain optimal co-kriging models and visualize their variograms of more than 2 pollutant covariates.

Acknowledgment

The authors would like to thank Dr. Dung Q. Ta, Faculty of Geology and Petroleum Engineering, VNUHCM, Vietnam. The first author appreciates valuable supports of *Center of Excellency in Mathematics* (CEM), Ministry of Education, Thailand, and Department of Mathematics, *Faculty of Science*, *Mahidol University*, Thailand. He thanks Faculty of Environment & Natural Resources, *University of Technology*, VNUHCM for support during 2015-2017.

Furthermore, we sincerely appreciate the anonymous reviewer whose valuable and helpful comments led to significant improvements from the original to the final version of the article.

Appendix A: Geostatistical terms and variograms

We employed the following specific geostatistical concepts (Dung [14]).

- *Range*: as the separation lag $\boldsymbol{h}$ between pairs increases, the corresponding variogram value will generally increase. The distance at which the variogram reaches the max averaged squared difference between pairs of values (named *plateau*) is **the range**.

- *Sill*: a concept describes i) the variance of the data (1.0 if the data are normal), and ii) the plateau that the variogram reaches at the range.

- *Nutget*: In geostatistical practice, a "nugget effect" refers to a discontinuity at the origin in the variogram; its magnitude (of the discontinuity) is called the **nugget**.

- *Anisotropy*: a concept is applied both to a random function and to it's variogram when the values of the variogram depend on both the distance and the direction, i.e depend on the lag $\boldsymbol{h} \in \mathbb{R}^2$ or $\mathbb{R}^3$.

Geostatistical modeling is generally useful for few activities, such as putting geological and/or environmental problems into quantitative models, as a result, estimating important parameters of the processes of interest, quantifying uncertainty, sample designing, simulation and risk analysis...

In particular, a variogram is a geostatistical model used to examine the spatial continuity of a regionalized variable and how this continuity changes as a function of a lag $\boldsymbol{h}$ (including distance and direction).

Computation of a variogram involves plotting the relationship between the variogram $(\gamma(\boldsymbol{h}))$ and the lag $(\boldsymbol{h})$. We write $h = |\boldsymbol{h}|$, the norm of $\boldsymbol{h}$ in the following most commonly used variogram models: spherical, exponential, linear, and Gaussian.

Spherical model. The spherical function is one of the most frequently used models in geostatistics (Webster [13]). The spherical model is a good choice when the nugget variance is important but not too large, and when there is also a clear range and sill (Burrough [5]):

$$\gamma(h) = \begin{cases} 0, & h = 0, \\ C_0 + C_1(\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3), & 0 < h < a, \\ C_0 + C_1, & h \geq a, \end{cases} \tag{6.1}$$

where

- $\gamma(h)$ is the semivariance,
- a is the range,
- C_0 is nugget variance, and
- $C_0 + C_1$ is the sill.

Exponential model. The exponential model is a good choice when there is a clear nugget and sill but only a gradual approach to the range:

$$\gamma(h) = C_0 + C_1(1 - exp(-\frac{h}{a})) \tag{6.2}$$

Linear model. This is a nontransitive variogram as there is no sill within the area sampled and typical attributes vary at all scales:

$$\gamma(h) = C_0 + bh \tag{6.3}$$

where b is the slope of the line.

Gaussian model. If the variance is very smooth and the nugget variance is very small compared to the spatially dependent random variation, then the ariogram can often be best fitted with the Gaussian model ([5]):

$$\gamma(h) = C_0 + C_1(1 - exp(-\frac{h^2}{a^2})) \tag{6.4}$$

References

[1] *Application of airborne pollutant emission models in assessing the current state of the air environment in Hanoi area caused by industrial sources*, Anh TV. Pham, **Ha Noi National University Publisher**, Vietnam no. 1, vol. 1, year = 2001, pp. 8–17

[2] *Geostatistical analysis of Spatial and Temporal Variations of groundwater level*, S. H. Ahmadi and A. Sedghamiz, **Environmental Monitoring and Assessment**, DOI: 10.1137/140951758, no. 129, vol. 35, pp. 277–294, 2007

[3] *Clearing the air: a review of the effects of particulate matter air pollution on human health*. Anderson J.O., Thundiyil, J.G., Stolbach, A., **J. Med. Toxicol**. 8 (2), pp. 166–175, 2012

[4] *Airliner World Magazine*, March 2018, United Kingdom, 2018, link `https://www.airlinerworld.com`

[5] *Principles of Geographical Information Systems*, P. Burrough and R. McDonnell, **Oxford University Press**, USA, 1998

[6] *Geostatistics- Modeling Spatial Uncertainty*, 2nd edition, Jean-Paul Chiles et al., **Wiley**, 2011

[7] *A tutorial guide to geostatistics: Computing and modeling variograms and kriging*, M. A. Oliver and R. Webster, Vol. 113, pp 56–69, **Catena, Elservier**, 2014

[8] *Geostatistics for the Environmental Science*, Version 5.1.1 **Gamma Design Software Corp.**, `https://geostatistics.com`, 2015

[9] *Geostatistics for Natural Resources Evaluation*, P. Goovaerts, **Oxford University Press**, 1997

[10] *Annual Report on National State of Environment*, **Ministry of Natural Resources and Environment**, 2010

[11] *On the Air of Towns*, Robert A. Smith, **Journal of the Chemical Society**, no. 9, vol. 1, 1859, pp. 196–235

[12] *Statistical Analysis with Missing Data*, Roderick JA. Little and Donald B. Rubin, **Wiley**, 1987

[13] *Geostatistics for Environmental Scientists*, R. Webster and M. A. Oliver, **Wiley**, 2007

[14] *Exploratory Data Analysis and Spatial Modeling*, Ta Quoc Dung, Geostatistics Course in **Workshop on Reliability Engineering**, HCMUT, Vietnam, 2016

[15] *Air-quality dispersion modeling alternative models*, **The United States Environmental Protection Agency**, Accessed date 2018/03/29, link `https://www.epa.gov/scram/air-quality-dispersion-modeling-alternative-models`,

[16] *Relevant potential impacts and methodologies for environmental impacts assessment related to solid waste management in Asian developing countries*, J. Van Buuren and J. Potting, Project report, deliverable 3.1, **Integrated Sustainable Solid Waste Management in Asia project**, Bremerhaven, Germany, 2011

[17] *Applying the Meti-lis model to calculate the emission of air pollutants from traffic and industrial activities in Thai Nguyen city, orienting to 2020*, Yen Doan, **Journal of Science and Technology, VAST**, Vietnam no. 6, vol. 106, 2013