

APPENDIX A. GEOSTATISTICAL METHODS

A.1 Chicken or the Egg?

The geostatistical modeling approach requires both structural analysis and specification of a positive definite covariance model (or a conditionally negative definite variogram model) to represent the theoretical spatial structure. Structural analysis consists of analyzing available concentration data to determine the correlational structure. The analysis uses historical contaminant data to discretely approximate the *theoretical variogram* $\gamma(\bar{h})$ given in equation (A.1) where \bar{h} is a distance separation vector (or increment) for two locations \bar{u}_{ψ} and $\bar{u}_{\psi'}$. $C(\bar{u}_{\psi})$ is a regionalized variable representing concentration that can be decomposed into its mean, $m(\bar{u}_{\psi})$, and zero mean random fluctuation, $r(\bar{u}_{\psi})$, components.

$$\begin{aligned} E[C(\bar{u}_{\psi}) - C(\bar{u}_{\psi'})]^2 &= \text{var}[C(\bar{u}_{\psi}) - C(\bar{u}_{\psi'})] + [m(\bar{u}_{\psi}) - m(\bar{u}_{\psi'})]^2 \\ &= 2\gamma(\bar{h}) + [m(\bar{u}_{\psi}) - m(\bar{u}_{\psi'})]^2 \end{aligned} \quad (\text{A.1})$$

The theoretical variogram $\gamma(\bar{h})$ represents the average dissimilarity between two concentrations separated by a distance \bar{h} . In the ordinary kriging approach, the mean component is assumed to be constant (or locally stationary) in which case the trend (or drift) term, $[m(\bar{u}_{\psi}) - m(\bar{u}_{\psi'})]^2$, in equation (A.1) is equal to zero and the theoretical variogram can be deduced directly from concentration samples. For nonstationary phenomena, the "chicken-and-egg" (Armstrong 1984) conundrum exists because the trend term is nonzero and a proper model for the theoretical variogram cannot be directly identified using concentration data solely without making an assumption on the functional form of the trend [i.e., specifying a mathematical function to model $[m(\bar{u}_{\psi}) - m(\bar{u}_{\psi'})]^2$ in equation (A.1)]. The next sections of this appendix give a detailed

description of the five kriging systems important to this thesis. Note that the kriging systems are presented in terms of covariance $K(\bar{h})$, which can be derived from the theoretical variance $K(0)$ and the variogram $\gamma(\bar{h})$ using equation (A.2).

$$\gamma(\bar{h}) = K(0) - K(\bar{h}) \quad (\text{A.2})$$

A.1.1 Ordinary Kriging (OK) in Local Neighborhoods

The OK estimator presented in equation (A.3) is formulated to provide the Best Linear Unbiased Estimate (or BLUE) for contaminant concentration at an unsampled location \bar{u}_j .

$$c_{est}^{OK}(\bar{u}_j) = \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{OK}(\bar{u}_j) c(\bar{u}_{\psi}) \text{ where } \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{OK}(\bar{u}_j) = 1, \forall j \quad (\text{A.3})$$

The ordinary kriging weights are found using equation (A.4), which can be derived from equation (A.3) by minimizing the expected estimation errors using the Lagrange multiplier method (for a detailed derivation see *Kitanidis 1997, Goovaerts 1997*). The first equation of (A.4) represents the unbiasedness constraint for the OK system and $\mu_{OK}(\bar{u}_j)$ is the Lagrange multiplier corresponding to this constraint.

$$\begin{aligned} \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{OK}(\bar{u}_j) &= 1, \forall j \\ \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{OK}(\bar{u}_j) K(\bar{h}_{\psi}) + \mu_{OK}(\bar{u}_j) &= K(\bar{u}_{\psi} - \bar{u}_j), \forall j \end{aligned} \quad (\text{A.4})$$

Solving the linear system represented by equation (A.4) yields the kriging weights used to compute each OK estimate and its corresponding estimation variance using equations (A.4) and (A.5), respectively.

$$\sigma_{OK}^2(\bar{u}_j) = K(0) - \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{OK}(\bar{u}_j) K(\bar{h}_{\psi}) - \mu_{OK}(\bar{u}_j), \forall j \quad (\text{A.5})$$

Note as stated previously, the estimation variance is only a function of the theoretical model used to represent the variogram and the geometrical configuration of the sampling data (i.e. a function of the kriging weights only). The OK approach invokes the *intrinsic hypothesis*, which assumes that mean or trend terms given in equation (A.1) are locally constant but unknown within neighborhoods surrounding the current unsampled location (i.e., $[m(\bar{u}_\psi) - m(\bar{u}_{\psi'})]^2 \approx 0$ in local neighborhoods). This assumption is equivalent to modeling the local trend using a zeroth order polynomial.

A.1.2 Multigaussian Kriging (MGK) in Local Neighborhoods

The MGK approach to estimation utilizes the normal score transform of the sample data and models the contaminant concentrations as a multivariate Gaussian random function, $T(\bar{u})$ (for details on the transform see *Goovaerts* 1997). The normal score transform maps the asymmetrical discrete sample distribution to a zero-one normal distribution. Estimation is performed using equations (A.3) thru (A.5) on the transformed data while again assuming a zeroth order polynomial trend.

A.1.3 Quantile Kriging (QK) in Local Neighborhoods

The distributional dependence of OK and MGK, as well as their sensitivity to highly skewed sample data, led to the development of non-parametric or distribution-free kriging approaches using indicator (*Journel* 1983) or rank-order (*Journel & Deutsch* 1997) transformations of the data. Indicator kriging uses a binary transformation of sample data where the data value is set to either 1 or zero depending on whether it is greater than or equal to a user specified threshold value. The indicator approach requires n_q threshold indicator transforms each of which requires its own variogram and kriging system. Correctly quantifying n_q variograms is generally not possible in plume interpolation applications due to data limitations

(see *Chilès & Delfiner* 1999). Moreover, solving equations (A.4) and (A.5) for each of the n_q cutoffs for every point in the interpolation represents a severe computational limitation of the method.

These limitations motivated the development of quantile kriging (*Journal & Deutsch* 1997), which transforms concentrations into standardized ranks (or quantiles) using equation (A.6) and computes estimates in quantile space using OK (see *Journal & Deutsch* 1997 and *Juang et al.* 2001).

$$c(\bar{u}) = \frac{\text{rank of sample } i}{N + 1} \quad (\text{A.6})$$

The transform consists of ranking the concentration data in ascending order and dividing these ranks by one plus the total number of sample data N . Figure (5.2) graphically illustrates quantile or standardized ranks transform. The i^{th} sample is assigned the probability that $c(\bar{x})$ is less than or equal to its concentration value [$F(c)$ in Figure (5.2)]. Estimates and estimation variances are computed in quantile space using equations (A.4) and (A.5). The empirical cdf of the sample data is also used to back transform estimates from quantile space to concentration space (for details see *Juang et al.* 2001) Since quantiles are known to have a uniform distribution, this fact can be used in conjunction with the estimation variances to compute non-parametric local uncertainty estimates at unsampled locations in the interpolation domain (see Chapter 5).

A.1.4 Kriging with a Trend (KT))

The KT approach (also termed “universal kriging”) models concentrations as a regionalized variable, $c(\bar{u})$, that is composed of the sum of a smoothly varying mean component $m(\bar{u})$ termed a trend and a second order stationary residual fluctuation $r(\bar{u})$. The KT estimator for contaminant concentration at an unsampled location \bar{u}_j is shown in equation (A.7) below.

$$c_{est}^{KT}(\bar{u}_j) = \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{KT}(\bar{u}_j) c(\bar{u}_{\psi}) \text{ where } \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{KT} p_s(\bar{u}_{\psi}) = p_s(\bar{u}_j) \quad \forall j \text{ and } s = 0, \dots, S \quad (\text{A.7})$$

The KT estimator requires the specification of a linear combination of known functions of spatial coordinates, $p_s(\bar{u})$, to model the trend component of the concentration. Typically the functions represent the sum of s -order monomials composing the overall S^{th} order polynomial used to model the trend. The kriging weights λ_{ψ}^{KT} are computed using the linear system shown in equation (A.8), which is derived by minimizing the expected estimation error for equation (A.7).

$$\begin{aligned} \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{KT}(\bar{u}_j) &= 1, \quad \forall j \\ \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{KT}(\bar{u}_j) p_s(\bar{u}_{\psi}) &= p_s(\bar{u}_j), \quad \forall j \text{ and } s = 0, \dots, S \\ \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{KT}(\bar{u}_j) K_r(\bar{h}) + \sum_{s=0}^S \mu_s^{KT}(\bar{u}_j) p_s(\bar{u}_{\psi}) &= K_r(\bar{u}_{\psi} - \bar{u}_j), \quad \forall j \end{aligned} \quad (\text{A.8})$$

Note that this system has $S+1$ constraints to enforce the unbiasedness condition and the stipulation that the monomials of the second equation of (A.8) are filtered from the system. After the kriging weights are computed using (A.8), equation (A.9) is used to compute the estimation variance at each unsampled location \bar{u}_j .

$$\sigma_{KT}^2(\bar{u}_j) = K_r(0) - \sum_{\psi=1}^{nsamp} \lambda_{\psi}^{KT}(\bar{u}_j) K_r(\bar{h}) - \sum_{s=0}^S \mu_s^{KT}(\bar{u}_j) p_s(\bar{u}_j), \quad \forall j \quad (\text{A.9})$$

The KT approach requires the practitioner to explicitly specify a functional form for the trend term $[m(\bar{u}_{\psi}) - m(\bar{u}_{\psi'})]^2$ in equation (A2.1) and subsequent structural analysis of the residual to attain its covariance function $K_r(\bar{h})$. Inference of the proper variogram is difficult

because the concentration data are not directly reflective of the residual function $r(\bar{u})$ thereby causing the “chicken-and-egg” (Armstrong 1984) conundrum. The proper model for the theoretical variogram cannot be directly identified using concentration data solely without knowing the functional form of the trend, which itself cannot be deduced without knowing the proper model for the theoretical variogram. Note the OK system in equation (A.3) can be directly derived from the KT system by modeling the trend as a zeroth order polynomial ($s = 0$) set equal to 1 in equation (A.7). Goovaerts (1997) shows that the only difference between the OK and KT systems results from the practitioner’s arbitrary decision to explicitly model the local trend as either a constant or a polynomial of order S .

A.1.5 Intrinsic Kriging (ItK)

The KT and variations of OK discussed above all require the practitioner to accept the dichotomy of the regionalized variable representing contaminant concentrations into a smoothly varying mean component and a stochastic fluctuation. Several studies have shown that this assumption is often an arbitrary choice that can significantly bias both estimates and their respective estimation variances (Volpi & Gambolati 1978, Hughes & Lettenmaier 1981, Russo & Jury 1987, Journel & Rossi 1989, Crawford & Hergert 1997). Intrinsic random function of order k (IRF- k) theory avoids the above dichotomy by not requiring the explicit specification of a model for the mean trend. Instead, the ItK approach defines allowable linear combinations of the sample data, termed *generalized increments*, that are implicitly capable of filtering trends from the data, facilitating a more direct measure of the underlying spatial structure of contaminant concentrations. The approaches discussed in previous sections model spatial structure using pairs of sample data to deduce the variogram shown in equation (A.1). Alternatively, IRF- k theory defines *generalized increments of order k* to be $(k+1)$ point

increments that are capable of filtering polynomial trends of degree less than or equal to k (*Chilès & Delfiner* 1999). Generalized increments, as the name implies, generalize the inference of spatial structure from a 2-point to a $(k+1)$ point measure of spatial correlation termed the *generalized covariance* $G(\bar{h})$.

In the ItK approach, concentration is modeled as an IRF-k, which for the proper generalized increment (or allowable linear combination of sample data) $C(\bar{u})$ is a stationary random function with a zero mean and a covariance equal to $G(\bar{h})$. It has been shown that the ItK system is identical to the UK system presented in equations (A.7) thru (A.9) except for the substitution of the generalized covariance function $G(\bar{h})$ for the 2-point residual covariance function $K_r(\bar{h})$ (for details see *Chilès & Delfiner* 1999). The ItK method was formulated with the goal of avoid the estimation biases caused by the explicit specification of a functional trend model.

