Scott Hauswirth
STOR890-Env. Statistics
Paper Summary
April 28, 2009

Paper Presented: Kennedy, P.L. and Allan D. Woodbury (2002) "Geostatistics and Bayesian Updating for Transmissivity Estimation in a Multiaquifer System in Manitoba, Canada." *Ground Water* 40 (3): 273-283.

The majority of this course has focused on large scale climate and human health applications. The paper presented here, however, is an example of a smaller scale application of statistics to a geo-environmental issue. Researchers intend to model groundwater in an approximately 200 kilometer (km) by 300 km area around the city of Winnipeg, Manitoba, Canada. The study area is located on the eastern edge of a region containing the greatest number of saline lakes in the world, as well as significant salt concentrations in the ground water. Two aquifers, separated by an impermeable layer, are present, an upper fractured carbonate aquifer and an underlying sandstone aquifer.The concern is that certain patterns of future exploitation of ground water resources could result in the intrusion of saline water into the study area.

The model is deterministic and uses a finite-element approach. Although the exact formulation of the model is not specified, a simplified version of the groundwater flow equation containing the elements discussed in the paper is shown below:

$$T\nabla^2 h = S\frac{\partial h}{\partial t} \tag{1}$$

where T is the transmissivity (hydraulic conductivity times aquifer depth), S is the storativity, and h, the head, is the variable to be solved for. From this equation it can be seen that T and S are parameters which must be assigned at each of the model discretization points. Generally (and in this case), data for these parameters is available at only a few isolated points within the domain. Frequently, these parameters are assigned (e.g. by averaging available data) to large spatial blocks, so that there are only a few distinct zones, each of which is considered homogeneous. This technique can introduce major errors and produce results which do not closely resemble the actual system. To better estimate these parameters over the whole domain, interpolation methods may be used.

The more traditional of these is kriging, which uses weighted values of surrounding data

points to interpolate values at unmeasured points. The basic equation is:

$$Z^* - m(\mathbf{u}) = \sum_{i=1}^{N} \lambda_i [Z(\mathbf{u}_i) - m(\mathbf{u}_i)] \tag{2}$$

where λ are the weights (to be determined) such that

$$\sigma^2 = Var\{Z(\mathbf{u})^* - Z(\mathbf{u})\} \tag{3}$$

is minimized, under the constraint that $\sum_{i=1}^{N} \lambda_i = 1$. This method will return exactly the measured value at a measured location, ignoring error in the data. The authors present the results of kriging interpolation for comparison purposes.

The method that is the focus of this paper is Bayesian updating, which is based on Bayes' theorem:

$$p[\mathbf{m}|\mathbf{d^*}, I] = \frac{p[\mathbf{d^*}|\mathbf{m}, I] \cdot p[\mathbf{m}|I]}{\int p[\mathbf{d^*}|\mathbf{m}, I] \cdot p[\mathbf{m}|I] d\mathbf{m}} \tag{4}$$

where $\mathbf{d^*}$ is the data vector (size n), $\mathbf{m}$ is the vector of interpolated values (size m), and I is any other information available. The posterior ($p[\mathbf{m}|\mathbf{d^*}, I]$) is thereby related to a prior ($p[\mathbf{m}|I]$) and a likelihood function($p[\mathbf{d^*}|\mathbf{m}, I]$]. The idea is to interpolate m-n additional points where the data and interpolated points are related by:

$$\mathbf{d^*} = \mathbf{G}\mathbf{m} + \nu \tag{5}$$

where $\mathbf{G}$ is a transformation kernel and ν is the noise in the data. To obtain $\mathbf{m}$, the model which best fits a prior estimate is used.

The authors assume a Gaussian distribution for the measurement errors and prior distribution and applying Bayes' theorem:

$$p[\mathbf{d^*}|\mathbf{m}, I] = \frac{1}{\sqrt{(2\pi)^n |\mathbf{C}_d|}} \exp\left[-\frac{1}{2}(\mathbf{d^*} - \mathbf{G}\mathbf{m})^T \mathbf{C}_d^{-1} (\mathbf{d^*} - \mathbf{G}\mathbf{m})\right] \tag{6}$$

$$p[\mathbf{m}|I] = \frac{1}{\sqrt{(2\pi)^m |\mathbf{C}_m|}} \exp\left[-\frac{1}{2}(\mathbf{m} - \mathbf{s})^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{s})\right] \tag{7}$$

$$p[\mathbf{m}|\mathbf{d^*}, I] = \frac{1}{\sqrt{(2\pi)^m |\mathbf{C}_q|}} \exp\left[-\frac{1}{2}(\mathbf{m} - \langle\mathbf{m}\rangle)^T \mathbf{C}_q^{-1} (\mathbf{m} - \langle\mathbf{m}\rangle)\right] \tag{8}$$

where $\mathbf{C}$ are covariance matrices for the data (d), model (m), and posterior (q) and $\langle\mathbf{m}\rangle$ is the vector of conditional expected values. The covariance matrices are of the form $\mathbf{C}_d = E[\boldsymbol{\nu}\boldsymbol{\nu}^T]$

and $\mathbf{C}_m = E[(\mathbf{m}-\mathbf{s})(\mathbf{m}-\mathbf{s})^T]$. Determining the unconditional moments and marginalizing "hyperparameters," $\mathbf{u}$ (unknown, "nuisance" parameters, e.g. prior unconditional mean $\mathbf{s}$, σ_d^2), then manipulating the equations, the following equations are determined:

$$\langle \mathbf{m} \rangle = \int_u \left(\mathbf{s} + \mathbf{C}_m \mathbf{G}^T (\mathbf{G}\mathbf{C}_m\mathbf{G}^T + \sigma_d^2\mathbf{I})^{-1}(\mathbf{d}^* - \mathbf{G}\mathbf{s}) \right) p(\mathbf{u}) d\mathbf{u} \tag{9}$$

$$\mathbf{C}_q = \int_u \left(\mathbf{C}_m - \mathbf{C}_m \mathbf{G}^T (\mathbf{G}\mathbf{C}_m\mathbf{G}^T + \sigma_d^2\mathbf{I})^{-1}\mathbf{G}\mathbf{C}_m \right) p(\mathbf{u}) d\mathbf{u} \tag{10}$$

Data was collected through the use of single- and multiple well pump tests and specific capacity tests, with a total of 2708 transmissivity measurements for the carbonate aquifer and 78 for the sandstone aquifer. The data showed a log normal distribution for both aquifers, with a slightly rougher histogram for the sandstone aquifer due to the lower number of data points. Variograms were plotted using both the traditional method and with the moving window estimator,

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{m} \sum_{j \in D_{i,h}}^{n} (Z(x_i) - Z(x_j))^2 \right] \tag{11}$$

where $D_{i,h}$ is the set of values within a moving window, h, of x_i. The latter was much smoother than the former, and was used for the analysis. An exponential model fit better than other models tested, however, significant deviations were observed at small lag distances for the carbonate aquifer, so a nested model, consisting of a nugget plus a spherical model plus an exponential model. A best fit exponential model was used for the sandstone aquifer.

The hyperparameters were $\mathbf{s}$ and σ_d^2 for the carbonate aquifer and $\mathbf{s}$, λ, σ_Y^2, and σ_0^2 for the sandstone aquifer. The additional hyperparameters for the sandstone aquifer were due to allowing the correlation structure to vary from the best fit model (due to low quality variogram). The mean was assumed to have a Gaussian distribution. The priors for the other hyperparameters were determined using minimum relative entropy (MRE), through which a prior is determined based on upper and lower bounds. The bounds can be reasonably estimated based on available hydrogeologic data. For comparison to kriging, an additional run was conducted assuming the data was exact. For hyperparameters with known mean and standard deviation the application of this method results in a Gaussian distribution.

A sampling method such as Monte Carlo, or in the case of this paper, Latin hypercube sampling, can be used to obtain a realization of the parameters. This realization can be applied to equations (9) and (10), to obtain realizations of $\langle \mathbf{m} \rangle$ and $\mathbf{C}_q$. The process is repeated until some criterion is met. The results are averaged over all iterations to obtain the final values. Latin hypercube sampling is similar to Monte Carlo, but instead of completely random sampling, the range of each parameter is divided into k sections of equal

probability, where k is the number of samplings. A value for a parameter is chosen randomly from each partition, and paired with a value from a partition for the other parameters. This method has been shown to be more precise than standard Monte Carlo sampling, with fewer iterations.

The resulting log-transmissivity fields show roughly the same features for the carbonate aquifer for all methods (kriging, Bayesian with exact measurements, and Bayesian with noise in the data): a low transmissivity "valley" running north-northeast to south-southwest, a relatively homogeneous area with higher values to the west, and a heterogeneous area on the east. The difference in homogeneity is likely due to the number of samples in these areas. The kriging method results in relatively sharp transitions where predicted values at the measured points return exactly the measured value. Assuming an exact measurement for the Bayesian method will also return the measured value at measured points, but produced a smoother surface. The smoothness of the field was further improved for the case where noise in the data was allowed. Comparison of standard deviations indicate a much higher values for the kriging method, ranging from 1.4 to 1.75 versus 0 to 1.6 for the Bayesian method. As would be expected, the areas of highest standard deviation are those with the fewest data points. For the sandstone aquifer, many fewer data points were available, with all points being on the east side of the study area. The Bayesian-produced field shows variation along the east side of the study area, where data are available, but assigns the mean of the underlying distribution to the rest of the field. Kriging, on the other hand, produced east-west oriented troughs of low transmissivity despite no data in the western portion of the study area. The maximum standard deviation over much of the field (i.e. the data deficient portion) is higher for the kriging method (2) than for Bayesian method (1.7), but for the eastern (data rich) portion, Kriging produces similar or lower standard deviations.