# Investigation of pH as Secondary Information in the Estimation of Soil Salinity

L. M. Bloom and U. A. Mueller

*School of Engineering and Mathematics, Edith Cowan University, Perth, Western Australia*
*(l.bloom@ecu.edu.au)*

**Abstract:** Soil salinity is a major problem in many rural areas of Australia. However, by the time the problem is visible at the surface, remediation, even if possible, is lengthy and costly. Soil salinity describes soils that have a high concentration of water-soluble salts and is usually quantified by (electrical) conductivity. In this paper we analyse soil pH and conductivity samples taken from cleared land in the Jimperding Brook catchment area in Western Australia. Parts of this valley have an obvious salinity problem, as does Jimperding Brook itself. In general, the collection and the subsequent analysis of conductivity samples is more expensive than obtaining the corresponding pH measurements, which can be taken in the field. Therefore we investigate the use of pH as a secondary variable in conductivity estimation. We assume that pH can be exhaustively sampled and for our case study we use already sampled pH data to simulate an exhaustive set. We consider three approaches for incorporating exhaustive secondary information. These are simple kriging with varying local means, kriging with an external drift, and collocated cokriging, with its Markov Model variants. These methods differ in the way in which the secondary variable is incorporated. While with cokriging the value of the datum directly influences the prediction, with kriging with an external drift only the trend contributes to it and in the case of simple kriging with varying local means only the residuals are modelled. With each of these methods, the burden of statistical inference increases over that required for ordinary kriging. In the case of kriging with an external drift and simple kriging with varying local means, the inference of the residual covariance is required, while for colocated cokriging the inference of the cross semivariograms becomes necessary. We compare the results obtained from using pH as a secondary variable with the results obtained from ordinary kriging of conductivity alone.

**Keywords:** Soil salinity; Cokriging; Geostatistics; Semivariograms

## 1. INTRODUCTION

Soil salinity refers to soils that have a high concentration of water-soluble salts and is a major problem in many rural areas of Australia. It is usually quantified by electrical conductivity (EC). Here we investigate the use of pH, which is easier and cheaper to sample, as a secondary variable in the estimation of electrical conductivity. We consider the question of whether there is any benefit in replacing results obtained solely from a reasonably large EC sample with those obtained from using a smaller EC sample together with a set of pH values. We apply a number of geostatistical estimation methods and compare the results from those incorporating secondary data with one another and with the results from direct estimation using the larger EC sample alone. Firstly we consider ordinary kriging and (traditional, two constraints) ordinary cokriging, Then, on the assumption that it is in fact possible t to sample pH exhaustively over the required study area, we apply simple kriging with varying local means, kriging with an external drift and ordinary colocated cokriging with its two Markov Model variants [Journel, 1999]. The relevant experimental semivariograms and cross-semivariograms were calculated and modelled. The algorithms used were applied [using GSLIB software; see Deutsch and Journel, 1998] to obtain not only EC estimates over the whole study area but also EC jackknife estimates in order to evaluate the effectiveness of the various methods.

## 2. DATA SET AND TREATMENT

The data used come from measurements taken from a field in the Jimperding Brook catchment area in the south west of Western Australia [Bloom and Kentwell, 1999]. Parts of this valley already have an obvious salinity problem, as does

Jimperding Brook itself. The initial sample data, named as *SECpH140*, were obtained on an 11 × 11 regular grid with 6m grid spacing, with 5 missing values due to the presence of rock, together with a further 24 samples taken at locations other than the grid nodes. The collection and analysis of conductivity samples is more expensive than making corresponding pH measurements, which can be obtained quite easily in the field. Here we have used the pH data in *SECpH140* to simulate pH on a 1 × 1 grid and the resulting dense data set is taken as the exhaustive secondary data set *EpH*. We randomly selected a sample set of 50 observations, named as *S50*, from *SECpH140* to use as our small sample set. From the remaining 90 data we randomly selected a subset of size 50, named as *J50*, to be a jackknife test set. We then combined the remaining 40 observations from *SECpH140* with the S50 data to obtain a large sample of 90 observations, named as *S90*. Summary EC statistics for these data sets are given in Table 1.

**Table 1.** Summary statistics for EC.

| Statistics | *S50* | *S90* | *J50* |
|---|---|---|---|
| Mean | 84.52 | 85.57 | 81.86 |
| Std. Dev. | 29.77 | 32.81 | 24.38 |
| Minimum | 45 | 38 | 40 |
| Median | 75 | 75 | 81 |
| Maximum | 178 | 215 | 140 |
| Skewness | 1.09 | 1.64 | 0.39 |
| pH Correlation | 0.57 | 0.47 | 0.40 |

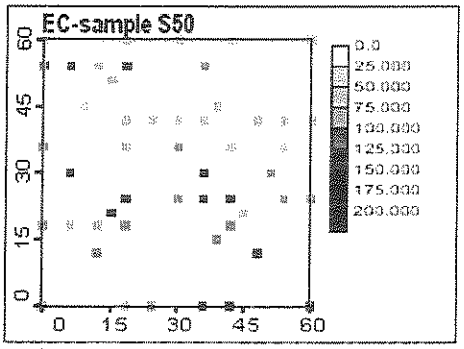

**Figure 1.** Postplot of EC data in S50.

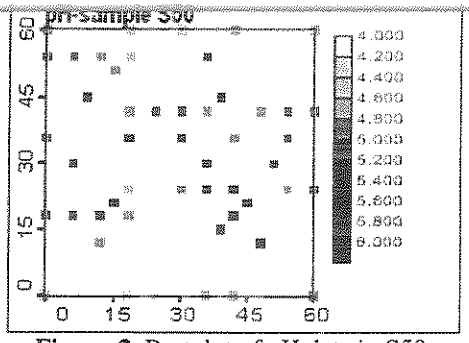

**Figure 2.** Postplot of pH data in S50.

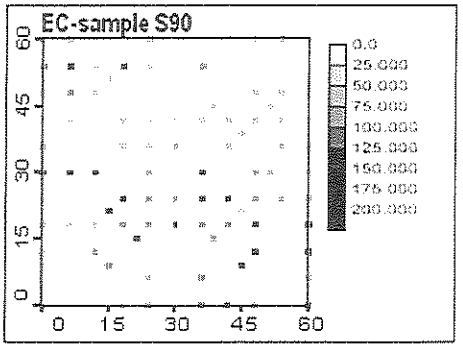

**Figure 3.** Postplot of data set S90.

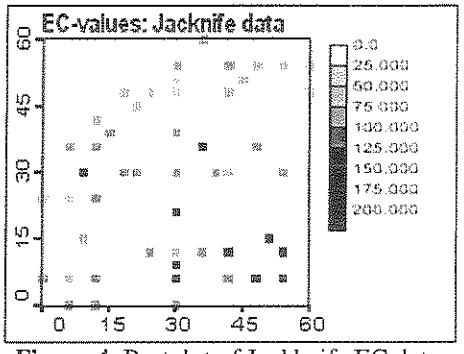

**Figure 4.** Postplot of Jackknife EC data.

Postplots of the EC and the pH data from *S50* are given in Figure 1 and Figure 2 respectively, while postplots of the EC data from *S90* and *J50* are given in Figures 3 and 4 respectively.

## 3. KRIGING METHODS

All methods considered here are variants of the linear regression estimator defined as

$$Z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u})[Z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)] \qquad (1)$$

where $m(\mathbf{u})$ and $m(\mathbf{u}_\alpha)$ denote the expected values of the random variables $Z(\mathbf{u})$ and $Z(\mathbf{u}_\alpha)$ respectively, $n(\mathbf{u})$ denotes the number of data locations near $\mathbf{u}$ and $\lambda_\alpha(\mathbf{u})$ is the weight associated with the datum $z(\mathbf{u}_\alpha)$ interpreted as a realisation of $Z(\mathbf{u}_\alpha)$ [for details see Goovaerts,

For ordinary kriging the mean is assumed to be locally constant but unknown and so the estimator is

$$Z^*_{OK}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) Z(\mathbf{u}_\alpha) \qquad (2)$$

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OK}(\mathbf{u}) = 1 \qquad (3)$$

Equation (3) filters the mean from the estimator.

The Simple Kriging with varying local means estimator assumes the means to be locally constant, but known and is identical to the estimator given in (1). From a procedural point of view, the residuals are modelled and estimated and the local mean is added after performance of the estimation.

For Kriging with External Drift the mean is modelled as a linear function of the secondary attribute

$$m(\mathbf{u}) = a_0(\mathbf{u}) + a_1(\mathbf{u})\, y(\mathbf{u}) \qquad (4)$$

and the estimator is

$$Z^*_{KED}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{KED}(\mathbf{u}) Z(\mathbf{u}_\alpha) \qquad (5)$$

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{KED}(\mathbf{u}) = 1 \text{ and } \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{KED}(\mathbf{u}) y(\mathbf{u}_\alpha) = y(\mathbf{u}) \qquad (6)$$

The traditional ordinary cokriging estimator is

$$Z^{(1)*}_{OCK}(\mathbf{u}) = \sum_{\alpha_1=1}^{n_1(\mathbf{u})} \lambda_{\alpha_1}^{OCK}(\mathbf{u}) Z_1(\mathbf{u}_{\alpha_1}) + \sum_{\alpha_2=1}^{n_2(\mathbf{u})} \lambda_{\alpha_2}^{OCK}(\mathbf{u}) Z_2(\mathbf{u}_{\alpha_2})$$

$$\sum_{\alpha_1=1}^{n_1(\mathbf{u})} \lambda_{\alpha_1}^{OCK}(\mathbf{u}) = 1 \text{ and } \sum_{\alpha_2=1}^{n_2(\mathbf{u})} \lambda_{\alpha_2}^{OCK}(\mathbf{u}) = 0 \qquad (8)$$

Finally, for colocated ordinary cokriging only the secondary information at the location to be estimated is used, and the estimator is

$$Z^{(1)*}_{OCCK}(\mathbf{u}) = \sum_{\alpha_1=1}^{n_1(\mathbf{u})} \lambda_{\alpha_1}^{OCCK}(\mathbf{u}) Z_1(\mathbf{u}_{\alpha_1}) + \lambda_2^{OCCK}(\mathbf{u})[Z_2(\mathbf{u}) - m_2 + m_1] \qquad (9)$$

$$\sum_{\alpha_1=1}^{n_1(\mathbf{u})} \lambda_{\alpha_1}^{OCCK}(\mathbf{u}) + \lambda_2^{OCCK}(\mathbf{u}) = 1 \qquad (10)$$

## 4. ORDINARY KRIGING

Ordinary Kriging (OK90) for EC was carried out using the EC data from *S90*. A zonal anisotropic semivariogram model was used with East West (azimuth angle 90°) as the direction of maximum continuity. The model consisted of a nugget of 110, together with a long range isotropic spherical structure together with an anisotropic short-range spherical structure in the direction of minimum continuity (see Table 2).

**Table 2.** Variogram parameters for OK90.

| | Azimuth | Sill | Range | Anis. |
|---|---|---|---|---|
| Sph. 1 | 90 | 950 | 26 | 1 |
| Sph. 2 | 90 | 320 | 9000 | 0.001 |

Cross-validation was carried out and EC estimates were obtained at the jackknife data locations as well as over the entire study area. The correlation coefficient between sample data and OK-estimates from cross-validation was 0.70, compared to 0.55 for the jackknife data.

## 5. ORDINARY COKRIGING USING SAMPLE pH DATA

Traditional (two constraints) Ordinary Cokriging (TOCK) was carried out using the isotopic EC and pH data from *S50*. In this case isotropic spherical variogram and cross- variogram models were used (see Table 3).

**Table 3.** Cross-variogram model parameters.

| | Nug. | Sill | Range |
|---|---|---|---|
| EC-EC | 0.2 | 0.8 | 11 |
| EC-pH | 0 | 0.6 | 11 |
| pH-pH | 0.2 | 0.8 | 11 |

Again, cross-validation was carried out and EC estimates were obtained at the jackknife data locations as well as over the entire study area.

## 6. CONDUCTIVITY ESTIMATION USING EXHAUSTIVE pH DATA

Electrical Conductivity estimation was then carried out by a number of methods, each of which assumes that the secondary variable can be sampled exhaustively. These are simple kriging with varying local means (SKlm), Kriging with an External Drift (KED) and Ordinary Colocated Cokriging (OCCK) with its Markov Model variants [Journel, 1999], labelled here as MM1OCCK and MM2OCCK. The OCCK and KED methods differ in the way in which the secondary variable is incorporated. With OCCK the value of the colocated datum directly influences the prediction, while with KED only the trend contributes to it. In the case of KED the inference of the residual covariance is required, while for OCCK the inference of one or more of the cross semivariograms becomes necessary. In each case, the pH data in *EpH* (see Figure 5) were used as the exhaustive secondary data.

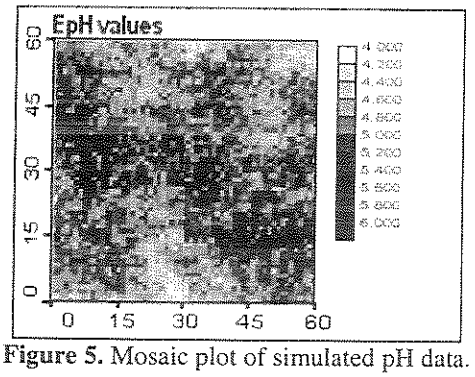


**Figure 5.** Mosaic plot of simulated pH data.

### 6.1 Simple Kriging with Varying Local Means

Since SKlm requires knowledge of the local means at all locations to be estimated, these were estimated via linear regression of EC on pH (see Figure 6 for the mosaic plot of the local means). The EC and pH data from *S50* were used to obtain the regression equation:

$$EC = -204.154 + 59.096\ pH \qquad (11)$$

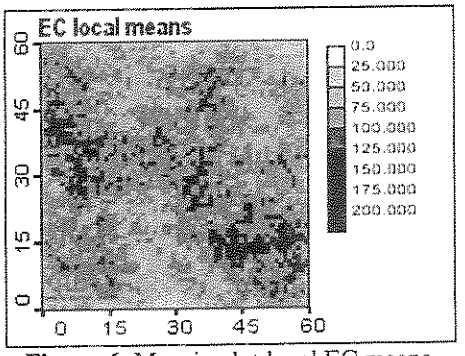


**Figure 6.** Mosaic plot local EC means.

The semivariograms in the directions of maximum and minor continuity were then fitted with a nested model consisting of a nugget of 140, together with a long range isotropic spherical structure together with an anisotropic short-range spherical structure in the direction of minimum continuity (see Table 4).

**Table 4.** Variogram parameters for SKlm.

| | Azimuth | Sill | Range | Anis. |
|---|---|---|---|---|
| Sph. 1 | 90 | 360 | 11 | 1 |
| Sph. 2 | 90 | 150 | 11000 | 0.001 |

### 6.2 Kriging with External Drift

This method is based on local linear regression using the EC and pH sample values from *S50* and the exhaustive pH values in *EpH*. In contrast to SKlm, in KED the local means are not calculated in advance, making the estimation of the residuals semivariogram more difficult. Since the EC data exhibited anisotropy with East-West as the direction of maximum continuity, the semivariogram taken to model the spatial continuity of the residuals was the experimental semivariogram for EC in the North-South direction. The model consisted of a nugget of 100 and a spherical isotropic structure (see Table 5).

**Table 5.** Variogram parameters for KED.

| | Azimuth | Sill | Range | Anis. |
|---|---|---|---|---|
| Sph. | 0 | 775 | 11 | 1 |

### 6.3 Ordinary Colocated Cokriging

In contrast to cokriging, colocated cokriging only uses the secondary information at the location to be estimated. As a consequence the inference of the semivariogram for pH is unnecessary, as only the pH-value at the location of interest is used in the cokriging system. The Markov assumptions are aimed at further simplifying the modelling process. MM1 assumes the cross-semivariogram between the primary and the secondary variable to be proportional to the semivariogram of the primary variable, while MM2 assumes the cross semivariogram to be proportional to the semivariogram of the secondary data. In each case the constant of proportionality is given by the correlation coefficient. The semivariogram for the primary variable is unaffected. For the semivariogram parameters for OCCK see Table 3. The cross-semivariogram models for MM1OCCK and MM2OCCK are given in Table 6. Note that the range for the semivariogram for EC has been adjusted to 15 for MM2OCCK to ensure that the covariance matrix is positive semidefinite.

**Table 6.** Variogram parameters for colocated Cokriging.

| | Nug. | Model | Sill | Range |
|---|---|---|---|---|
| MM1OCCK | 0.11 | Sph. | 0.46 | 11 |
| MM2OCCK | 0.08 | Sph. | 0.43 | 15 |

## 7. SUMMARY OF RESULTS AND CONCLUSIONS

Mosaic plots of the resulting EC estimates are shown in Figures 7, 8 and 9. The estimates from SKlm and KED exhibit greater variability than those obtained from the other methods. This is a direct consequence of the variability in the exhaustive pH data (*EpH*) itself. In each case the correlation coefficient, the mean square error (MSE) and the mean absolute deviation (MAD) were obtained from the true and estimated EC values of *J50*. Table 7 gives the error comparison between the various methods for this jackknife test set.

It can be seen from these values that the use of a small sample, together with pH secondary data has in fact led to an improvement over the use of a larger EC sample alone. This improvement was by far the greatest when exhaustive secondary data and the Kriging with External Drift method

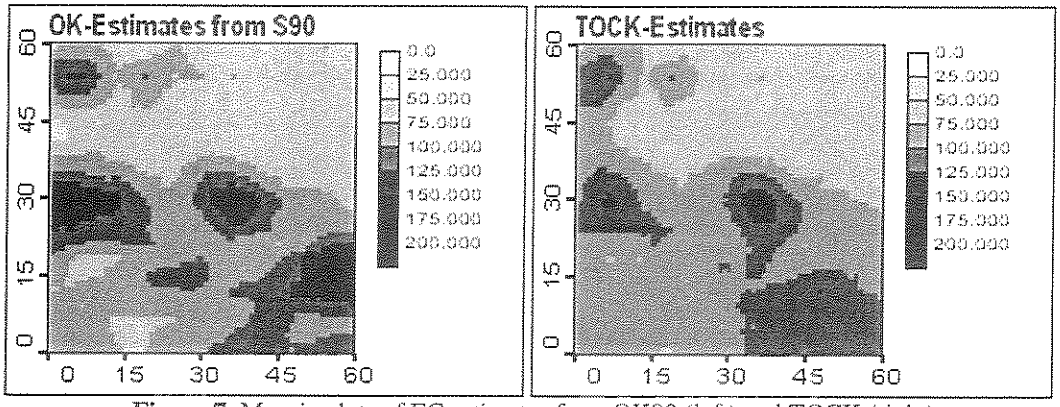


**Figure 7.** Mosaic plots of EC estimates from OK90 (left) and TOCK (right).

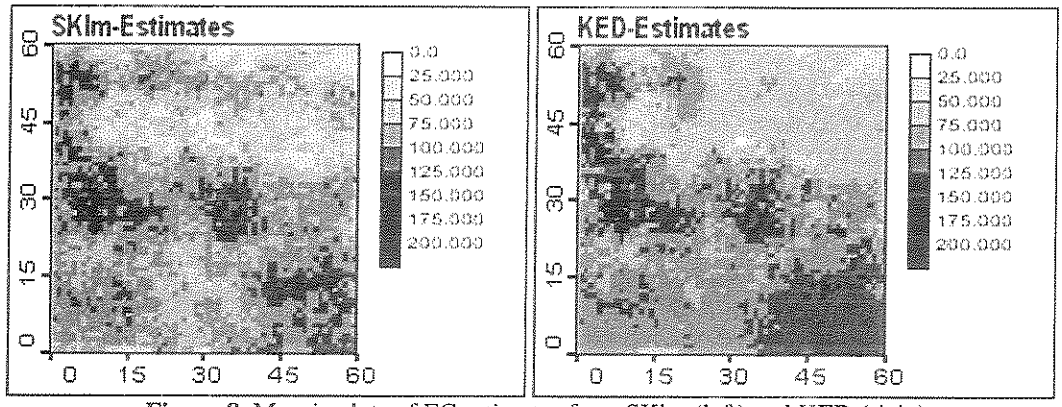


**Figure 8.** Mosaic plots of EC estimates from SKlm (left) and KED (right).

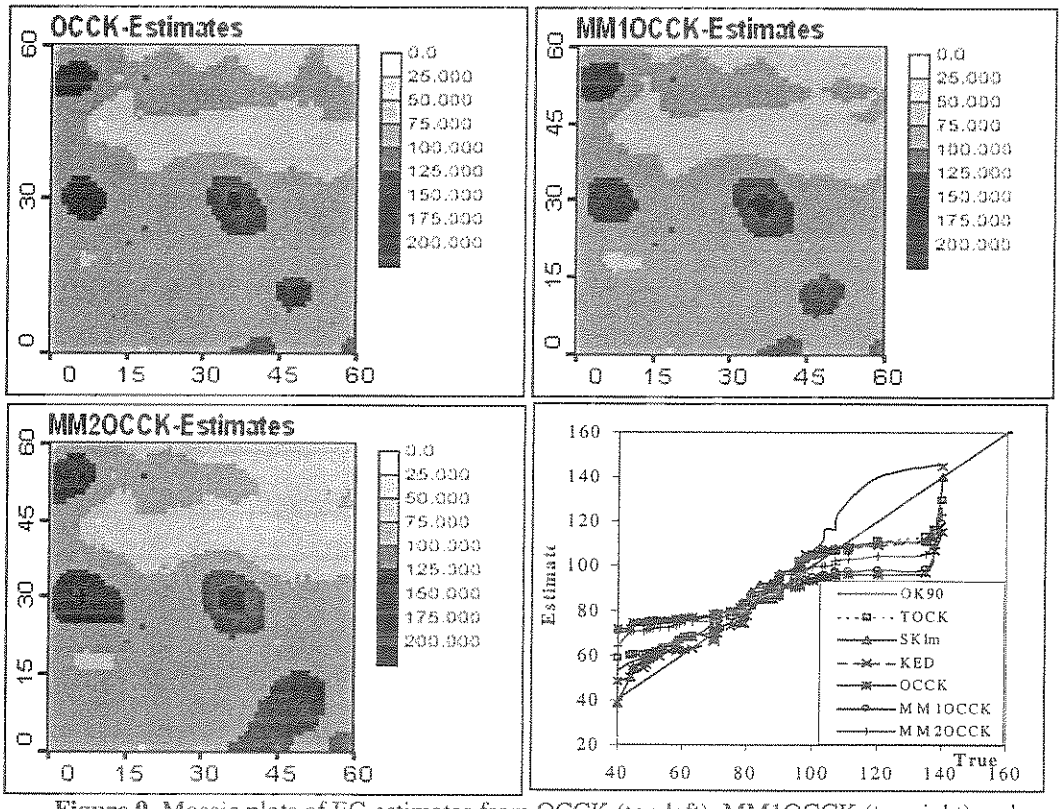


**Figure 9.** Mosaic plots of EC-estimates from OCCK (top left), MM1OCCK (top right) and MM2OCCK (bottom left) and qq-plots of true versus estimated jackknife data (bottom right).

**Table 7.** Correlation, MSE and MAD for *J50*.

| Method | Correlation | MSE | MAD |
|---|---|---|---|
| OK90 | 0.55 | 625.1 | 19.2 |
| TOCK | 0.58 | 417.3 | 16.2 |
| SKlm | 0.56 | 416.6 | 18.3 |
| KED | 0.70 | 315.4 | 14.8 |
| OCCK | 0.52 | 444.4 | 17.1 |
| MM1OCCK | 0.52 | 442.2 | 17.0 |
| MM2OCCK | 0.52 | 441.1 | 17.0 |

was used. It is worth noting that there is virtually no difference here between the results from the three colocated cokriging methods. This is due to the fact that the relevant modelling led to very similar cross-variogram models in this particular case.

The q-q plots of true versus estimated jackknife data in Figure 11 indicate that conditional bias is present for all models, but is less pronounced for SKlm and KED. All methods overestimate low data values and, except for OK90, underestimate high values. These q-q plots indicate that, globally, KED and SKlm (in that order) are the best EC estimators in this case. Although the global performances of all three colocated cokriging methods are similar, the MM2OCCK variation performs best in this context.

Finally, the evidence from this study is that, even though the EC and pH are in fact only moderately correlated (between 0.4 and 0.6 in the various sample sets considered), estimation of EC values using a relatively small EC sample, together with exhaustively sampled pH as a secondary variable, provides an improvement over the use of a larger EC sample on its own.

## 8. REFERENCES


Bloom, L. M. and D. J. Kentwell, A geostatistical analysis of cropped and uncropped soil in the Jimperding brook catchment area of Western Australia, in *geoENVII – geostatistics for environmental applications*, J. Gomez-Hernandez et al (eds), Kluwer Academic Publishers, Dordrecht, 369 – 380, 1999.

Deutsch, C. and A. Journel, *GSLIB: Geostatistical Software Library and User's Guide*, 2nd. Ed.,Oxford University Press, 369 pp, New York, 1998.

Goovaerts, P. *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 483 pp, New York, 1997.

Journel, A. Markov Models for Cross-Covariances, *Mathematical Geology*, 31(8), 955-964, 1999.