Local regression kriging approach for analysing high density data

Wei Sun[A,B,C], Budiman Minasny[B] and Alex McBratney[B]

[A]China Agricultural University, East Campus, Research Center for Precision Agriculture. Qing Hua Dong Lu 17, Beijing, 100083, P.R. China.
[B]Australian Centre for Precision Agriculture. Faculty of Agriculture, Food and Natural Resources, The University of Sydney, NSW 2006, Australia.
[C]Corresponding author. Email sunwei1028@126.com

Abstract
This paper demonstrates the development and an application of a local regression-kriging (RK) program. It illustrates the performance of the local RK method for prediction of soil properties using high density crop yield data in time series. We developed a software named RKGuider to carry out a serial steps of RK, especially local RK automatically and using the program, we produce a high resolution soil electrical conductivity map from a coarse survey. The result demonstrated there is spatial correlation between yield data in time series and soil electrical conductivity. High density yield data can be taken as auxiliary variables to predict soil EC using RK.

Key Words
Local regression kriging, variogram model, linear model, yield, electrical conductivity.

Introduction
Regression kriging is a spatial prediction technique which adds the regression value of exhaustive variables and the kriging value of residuals together .The local RK algorithm was developed especially to take into account the local correlation between environmental variables and the unsatisfactory goodness of fit of the spatial variance model for the entire data set. Beside that, local RK is developed to deal with rich data, using the data in full capacity or to improve prediction from a global model. Local RK result is the sum of a local regression of auxiliary variables and local kriging of the regression residuals. This is called RK in a moving window and was considered as the "next step" for RK development (Hengl *et al.* 2007).This algorithm will play more and more important role in geostatistics because lots of covariates are available dramatically now by advancement technology (Pebesma 2006). However there is a serious constraint to wider use of RK, especially local RK, which is that a user must carry out various steps in a variety of software packages (Hengl *et al.* 2007). Furthermore, there is no software that can do local RK efficiently. Therefore it is necessary to develop a framework for spatial interpolation based on local RK.

Methods
Algorithm
Local RK involves the following steps:
1) Searching for the closest neighbourhood for each prediction site,
2) Fitting a linear regression model predicteing the attribute from the covariates from the neighbourhood data,
3) Calculating a residual of the regression model for each neighbourhood point,
4) Estimating the variogram from the residuals,
5) Fitting a variogram model to them
6) Predicting the residual value and its uncertainty for each prediction site with kriging,
7) Summing up the regression value and the kriging of residuals together, and calculating the uncertainty.

In order to make the local regression-kriging application simpler, a software named RKguider was developed to carry out these steps automatically. It uses automatic linear and nonlinear regression routines, in which a search algorithm was applied for increasing speed. The program also incorporates the maximum likelihood method (Mardia and Marshall 1984) which refines iteratively the prediction of the variogram parameters and linear model regression. The program is an advancement of the Vesper program which only performs kriging with local variogram (Minasny *et al.* 1999). RKguider was developed using Microsoft Visual C++ 6.0 development system. The program offers a friendly interface with a range of options to users to deal with dataset, and provide the flexibility to calculate global RK and local RK. Weighting options could be changed before fitting semivariogram modeling. During processing of local RK, it can provide a real-time graphical display of the local regression and semivariogram modeling and the searching neighbours could be set as analyst wants. The output is an ASCII text including the prediction point coordinates, the predicted value, the regression variance, the kriging variance, the total variance and the regression parameters. One limit of the program is that the covariates would better be continuous. For binary variables, it can sometimes fail when the local neighborhood only has a single value.

Application and testing

The program was applied to high density crop yield data gathered from yield monitors. The objective is to produce a high resolution soil electrical conductivity map from a coarse survey. Continuous four years yield data are available from 2003 to 2006 in 4 meters resolution, which are taken as the target variables used in linear regression, because high spatial relationship between soil EC and yield has been reported.

Results

Digital maps at a resolution of 4m x 4m for soil ECa were predicted using local kriging and local RK. Linear regression map and residuals map were also presented in Figure 2. Data-out validation are used to compare local kring with 100 neighbours, local RK with 100, 200 and 300 neighbours separately and linear regression. The data set was divided into interpolation (about 7500) and validation set (about 2500), then R2 and Standardized squared deviation $\theta(x)$ were taken as index to measure the prediction efficiency and the goodness of theoretical estimates (Table 1)

Table 1. Statistics results.

Prediction Method	Median $\theta(x)$	Mean $\theta(x)$	RMSE	R^2
Linear regression	-	-	9.67	0.251
Local RK with 100 neighbors	0.280	1.097	3.23	0.916
Local RK with 200 neighbors	0.230	0.999	3.13	0.921
Local RK with 300 neighbors	0.226	0.937	3.16	0.920
Local kriging with 100 neibghours	0.700	3.328	2.77	0.939

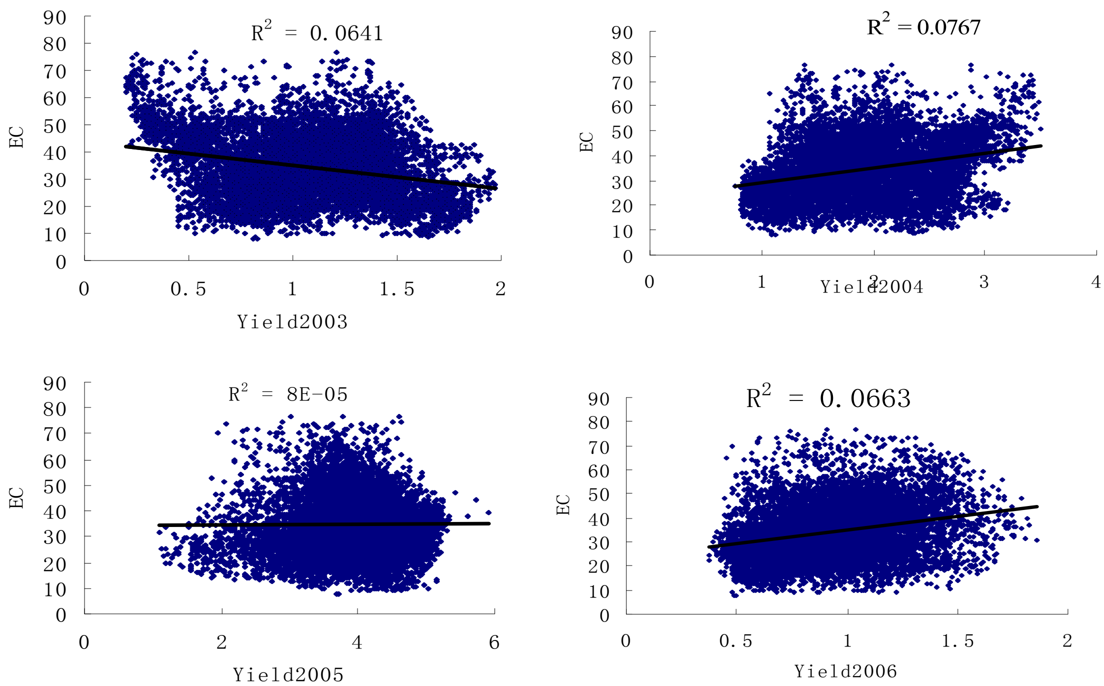

Figure 1. Scatter plot of EC and yield data.

Although the results show that the local RK method does not present much better results than local kriging; however we are able to understand the pattern of soil ECa and its relationship with crop yield. General linear regression (Figure 1) shows that there are variable response between ECa and yield, however the spatial pattern reveal areas where yield for each year has a correlation with soil ECa.

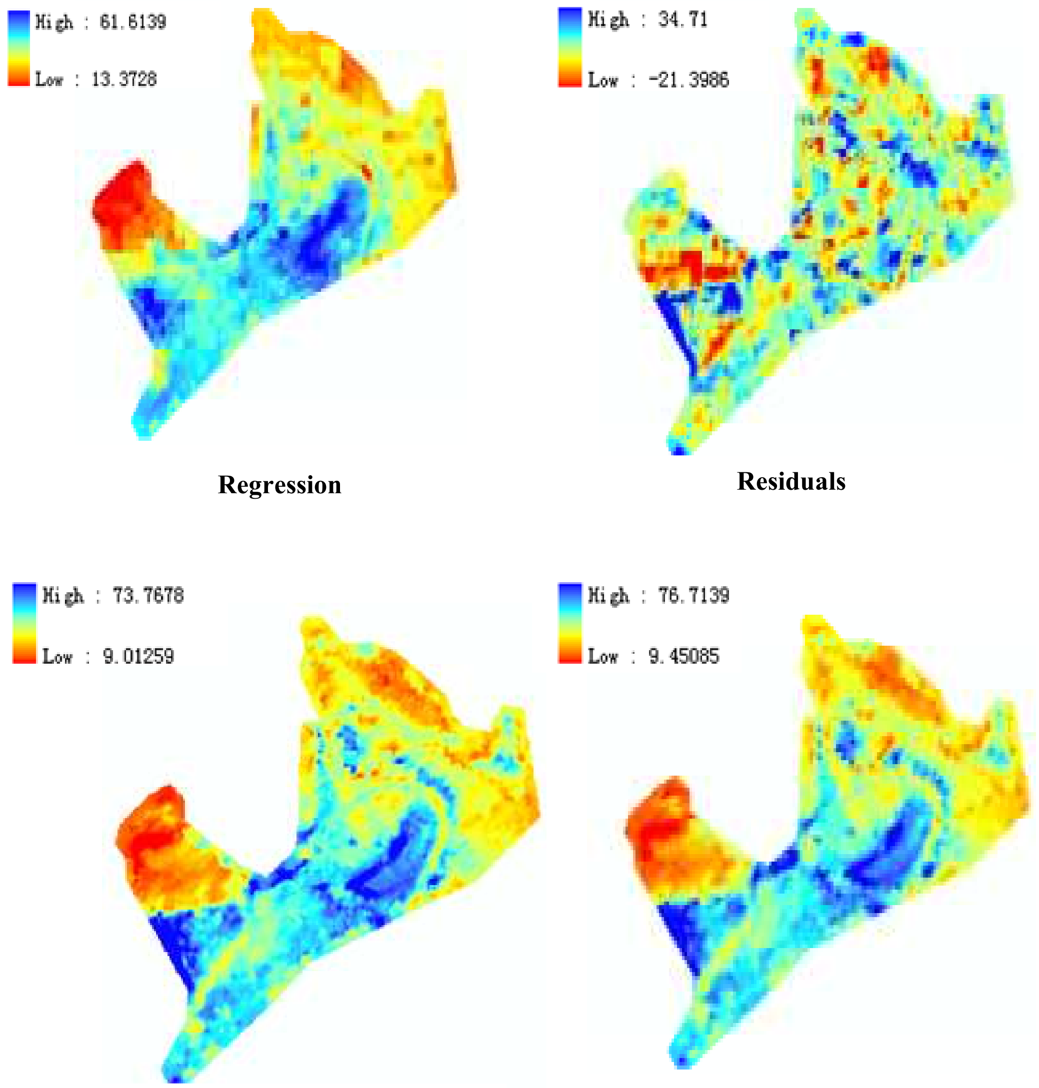

Figure 2. Digital maps at a resolution of 4 m x 4 m for soil ECa obtained using local regression kriging and local kriging.

Conclusion

We have developed a program for conducting local regression kriging. There is correlation between yield data in time series and soil electrical conductivity. High density yield data can be taken as auxiliary variables to predict soil EC using RK.

References

Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Minasny B, McBratney AB, Whelan BM (2002) VESPER, Variogram Estimation and Spatial Prediction plus Error. Australian Center for Precision Agriculture, The University Sydney, NSW 2006, Australia.

Hengl T, Heuvelink GBM, Rossiter D (2007) About regression-kriging: From equations to case studies. *Computers and Geosciences* **33**, 1301-1315

Pebesma E (2006) The Role of External Variables and GIS Databases in Geostatistical Analysis. *Transactions in GIS* **10**, 615-632.