Statistics and Data Analysis in Physical Geography (GEO4-4412)

Second Exam

8 November 2012
09.00 – 12.00 h

Name: :.....

Registration number :.....

Study programme :.....

- Answers may be given in Dutch or English
- Read all questions first
- Start answering the easier questions first
- Use the empty space between the questions to write down your answer
- Formulate your answers briefly and to the point
- Write neatly
- Think twice before you write down your answer (avoid making later corrections)
- Good luck!

Question 1 Mark:
Question 2 Mark:
Question 3 Mark:
Question 4 Mark:
Question 5 Mark:
Question 6 Mark:
Question 7 Mark:
Question 8 Mark:

Final score:

1a. Multiple linear regression is often considered ‘multivariate’, but it is basically a ‘univariate’ technique, why?

1b. Show how the general multiple regression model can be derived from the polynomial regression model in which variable *y* is predicted from one independent variable *x*.

1c. A linear regression model explains variable *y* as a function of three variables x_1, x_2, x_3. The obtained regression coefficients are: $b_0 = 10.3$, $b_1 = 0.68$, $b_2 = 0.12$, $b_3 = 0.46$, The variances of the variables are: var(y) = 27.8, var(x_1) = 3.12, var(x_2) = 76.9, var(x_3) = 17.3 Which independent variable explains most of the variance in *y*?

2. Given is the variance-covariance matrix of two variables x_1 and x_2: $\begin{pmatrix} 18 & 10 \\ 10 & 15 \end{pmatrix}$

The eigen values of the matrix are: $\lambda_1 = 26.61$ en $\lambda_2 = 6.39$, and the corresponding eigen vectors are $eig_1 = \begin{pmatrix} 0.76 \\ 0.65 \end{pmatrix}$ and $eig_2 = \begin{pmatrix} 0.65 \\ -0.76 \end{pmatrix}$

a. Make a plot of the ellipse that is spanned by the matrix and its eigen vectors.

b. What is the correlation between variables x_1 and x_2?

c. What is the correlation of the new variables when x_1 and x_2 are projected on the two principle components?

3a. Give three examples of a regionalized variable?
3b. What is the support of a regionalized variable?
3c. What is indicated by the semi-variance of a regionalized variable?
3d. Theoretically, the semi-variance is zero at a lag distance of zero. But, in practice the semi-variance is often non-zero at very small lag distances, why?

4. Given are the following data of elevation (Y in m) at 11 spatial positions (X in m)

x_i (m)	0	1	2	3	4	5	6	7	8	9	10
y_i (m)	3	2	1	1.5	3.5	5	7	8	6	4	5

4a. Calculate the semi-variogram for lag distances 1, 2, 3, 4, 5 and 6 m

4b. Draw the semi-variogram and estimate its range and sill values.

4c. What is the meaning of the range and sill of a semivariogram?

5a. We have sampled a regionalized variable in a certain area and it appears that there are relatively high values (= higher than the mean) in the north and south of the area, while in the central part there are relatively low values. Explain all required steps when we want to apply ordinary kriging on this regionalized variable to create an accurate spatial map.
5b. What is the main difference if we would apply universal kriging instead of ordinary kriging on the same regionalized variable as in question 5a?

6a. Principle Component Analysis (PCA) can be done by using a variance-covariance matrix or by using a correlation matrix of a number of variables x_i. What determines the choice of the type of matrix to be used?

6b. In case the correlation matrix is used, what happens to the relative importance of a certain variable x_1 when its variance is small compared to the variances of the other x variables?

6c. Assume we have a large dataset consisting of 10 variables and 100000 observations for each variable. We want to reduce the size of the dataset by exactly 60% in such a way that the main properties are retained. Describe how we can do this without losing much information.

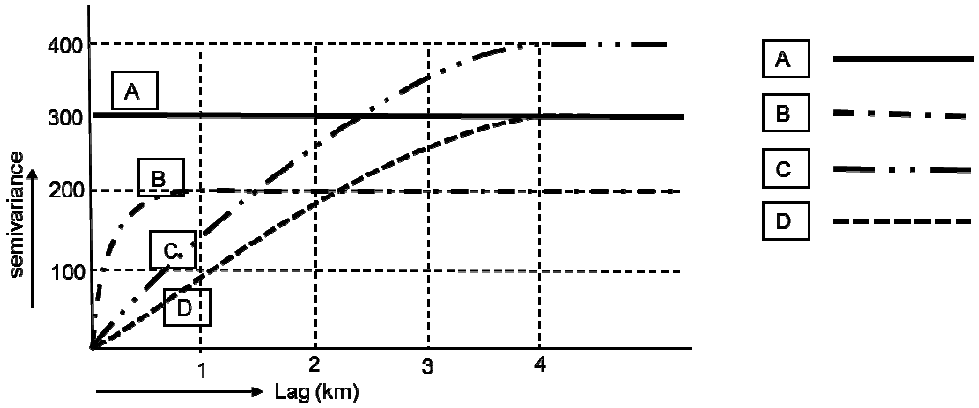

Figure 1. Semivariogram models of soil property Z in four different areas (A, B, C and D).

7. Monitoring Spatial average value

7a. The semivariogram models of soil property *Z* in four different areas, are plotted is in figure 1. (σ_x^2 is the variance of soil property *Z* in area *x*.) Which statement is true for the variance of soil property *Z*:

a. $\sigma_A^2 > \sigma_B^2 > \sigma_C^2$
b. $\sigma_C^2 > \sigma_B^2 > \sigma_A^2$
c. $\sigma_A^2 = \sigma_D^2 > \sigma_B^2$
d. $\sigma_C^2 = \sigma_D^2 > \sigma_B^2$

7b. The standard deviation of the spatial average of soil property *Z* in area *x* is $\sigma_{\bar{x}}$. If there are *N* random observations, the standard deviation of the spatial average surface elevation is calculated with:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} .$$

Let the standard deviation of soil property *Z* in area C be $\sigma_C = 20$. The area $C = 400$ km^2. What is the minimum number of random observations to meet the condition that the standard deviation of the spatial average soil property *Z* does not exceed 5.

7c. The standard deviation of the soil property Z in the areas *B* and *D* are respectively: σ_B = 14, σ_D = 17. In order to obtain the same standard deviation of the spatial average soil property in both areas $(\sigma_{\bar{B}} = \sigma_{\bar{D}})$. The number of random observations in area B should be:

a. larger than the number of observations in area *D*
b. smaller than the number of observations in area *D*
c. equal to the number of observations in area *D*

7d. Area *E* is 1600 km^2. The soil property *Z* in area *E* has the same variogram model as the area *C*. The number of random observations is area $C = N_C$. To obtain the same standard deviation of the average soil property in both areas $(\sigma_{\bar{C}} = \sigma_{\bar{E}})$, the number of random observations in area *E* should be:

a. 4* N_C
b. ¼ N_C
c. N_C

8. Monitoring Spatial interpolation.

Consider the areas given in the previous question. The monitoring objective is a spatial interpolation of soil property *Z*. For all areas the relationship between the monitoring effort and the uncertainty of the interpolation is derived using block Kriging with a block size 100*100m.

8a. Which statement is true for the spatial correlation (range) of the semivariogram models in Figure 1.

a. The range of semivariogram model *B* is larger than the range of semivariogram model *C*
b. The range of semivariogram model *A* is equal to the range of semivariogram model *B*
c. The range of semivariogram model *C* is larger than the range of semivariogram model *D*
d. The range of semivariogram model *C* is equal to the range of semivariogram model *D*

8b. Assume that the observation locations are on a regular square grid and de distance between the observation locations is 500 m. The maximum value of the Kriging standard deviation in area *x* is $\sigma_{max,x}$. Which of the following statements is true?

a. $\sigma_{max,B} > \sigma_{max,C} > \sigma_{max,D}$
b. $\sigma_{max,C} > \sigma_{max,B} > \sigma_{max,D}$
c. $\sigma_{max,B} > \sigma_{max,D} > \sigma_{max,C}$
d. $\sigma_{max,C} > \sigma_{max,D} > \sigma_{max,B}$

8c. We compare two alternative monitoring networks for area *B* with observation locations at a regular square grid. The distance between the observation locations in alternative 1 is 4 km, and in alternative 2 is 10 km. The maximum Kriging standard deviation in alternative 1 is:

a. Larger than in alternative 2,
b. Smaller than in alternative 2
c. Equal to alternative 2.

8d. The variogram of soil property *Z* in area *E* is the same as in area *C*. Area *C* is 400 km^2 and area *E* is 1600 km^2. The monitoring network (observations at a square regular grid) for both areas result in the same maximum Kriging standard deviation ($\sigma_{max,C} = \sigma_{max,E}$) . The total number of observation locations is area *C* is:

a. Larger than in area *E*
b. Smaller than in area *E*
c. Equal to area *E*.

8e. Three network density graphs (the maximum Kriging standard deviation as function of the distance between the observation locations) for area *C* are given in figure 2. The graph Block size I is the network density for the block size 100*100 m. Which of the following statement is true:

a. Block size I is larger than block size II,
b. Block size I is larger than block size III,
c. We can't decide from figure 2 whether block size I is smaller or larger than block sizes II and III.
d.

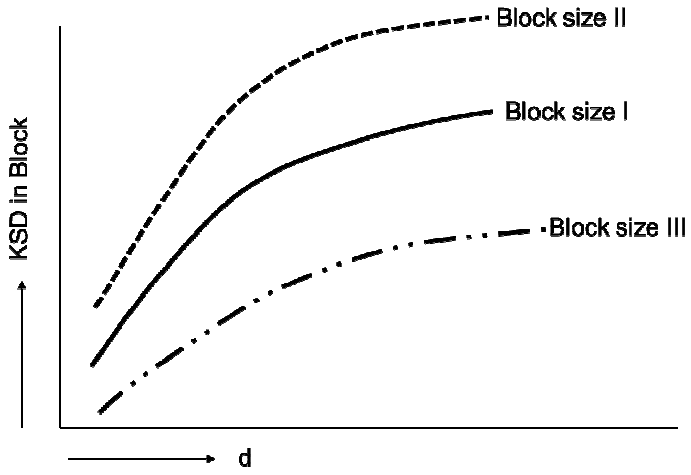

Figure 2. Network density graphs for area C.