SOME EXAM DATE
SOME EXAM TIME
EXAMINATION FOR THE DEGREES OF M.SCi. AND M.Sc.
(SCIENCE)

STATISTICS
Spatial Statistics M

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

NOTE: Candidates should attempt THREE out of the FOUR questions. If more than three questions are attempted please indicate which questions should be marked; otherwise, the first three questions will be graded.

1. (i) Let $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ $(D \subset \mathbb{R}^2)$ be a mean-zero weakly stationary and isotropic geostatistical process with covariance function $C_Z(h)$.

   (a) Write down the covariance function $C_Z(h)$ of a white noise process and describe why it is generally not a suitable process for modelling spatial data. **[2 MARKS]**

   (b) The semi-variogram for a geostatistical process at two locations $(\mathbf{s}, \mathbf{t})$ a distance $h = ||\mathbf{s} - \mathbf{t}||$ apart is given by

$$\gamma_Z(\mathbf{s}, \mathbf{t}) = \frac{1}{2}\text{Var}[Z(\mathbf{s}) - Z(\mathbf{t})].$$

   Using this definition derive the semi-variogram for a white noise process and draw a graph showing the shape of the semivariogram $\gamma_Z(\mathbf{s}, \mathbf{t})$ against distance $h = ||\mathbf{s} - \mathbf{t}||$. **[3 MARKS]**

CONTINUED OVERLEAF/

(c) Describe the two conditions required for a weakly stationary and isotropic covariance function $C_Z(h)$ to be a valid covariance function. **[2 MARKS]**

(d) Using the answer to the previous question prove that the white noise process has a valid covariance function. **[4 MARKS]**

(ii) Consider 2 independent Gaussian Geostatistical processes $(X(\mathbf{s}), Y(\mathbf{s}))$ that have mean-zero and the following weakly stationary and isotropic covariance functions

$$\begin{aligned} C_X(h) &= \begin{cases} \sigma^2 \exp(-h/\phi), & h > 0, \\ \sigma^2, & h = 0, \end{cases} \\ C_Y(h) &= \begin{cases} \nu^2 \exp(-h/\phi), & h > 0, \\ \nu^2 + \sigma^2, & h = 0, \end{cases} \end{aligned}$$

where h represents distance.

(a) What type of covariance models are defined for $(X(\mathbf{s}), Y(\mathbf{s}))$ and what are the differences between them. **[3 MARKS]**.

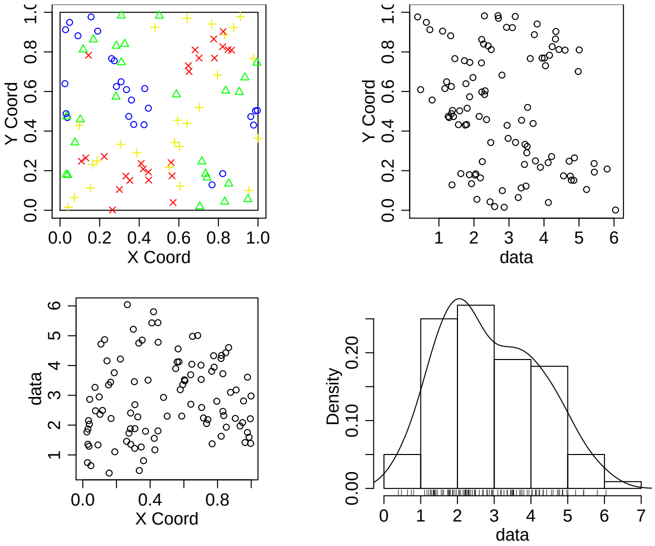
(b) Derive the covariance function for

$$Z(\mathbf{s}) = a + bX(\mathbf{s}) + Y(\mathbf{s})$$

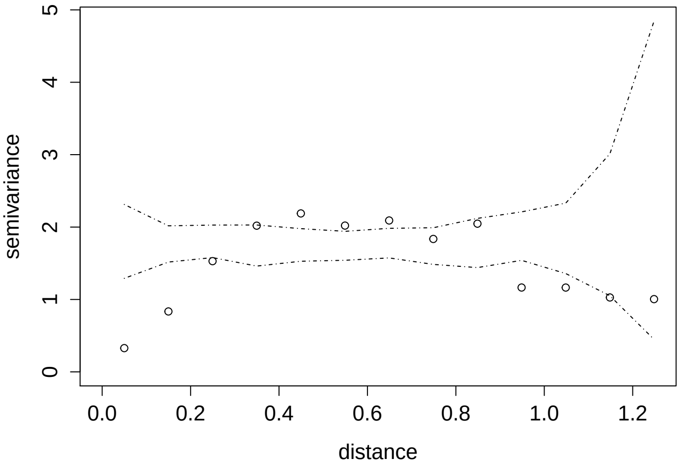
and define the nugget, partial sill and range parameters of this function. What type of covariance model is this? **[6 MARKS]**

CONTINUED OVERLEAF/

2. (i) Samples of magnesium (Mg) levels in a 1 kilometre square in Glasgow were collected, and are shown below as a *geodata* object. Comment on the choice of sampling design and the presence or absence of trends and spatial autocorrelation in these samples. **[3 MARKS]**



(ii) The empirical semi-variogram is shown below for these data together with Monte-Carlo envelopes. Describe briefly how the Monte-Carlo envelopes are constructed and determine whether the magnesium data exhibit any spatial autocorrelation giving a reason for your answer. **[4 MARKS]**



CONTINUED OVERLEAF/

(iii) Two models were fitted to these data using the *geoR* software and maximum likelihood estimation, the first with an exponential covariance function and the second with a spherical covariance function, and model summaries are shown below. From these summaries which of these two models better fits the data and are either better than a model with no spatial autocorrelation (a nugget only model)? **[2 MARKS]**

```
Exponential model
 Maximised Likelihood:
   log.L n.params      AIC      BIC
"-113.9"      "4"  "235.8"  "246.2"

non spatial model:
   log.L n.params      AIC      BIC
"-169.9"      "2"  "343.8"    "349"

Spherical model
Maximised Likelihood:
   log.L n.params      AIC      BIC
"-112.3"      "4"  "232.5"    "243"

non spatial model:
   log.L n.params      AIC      BIC
"-169.9"      "2"  "343.8"    "349"
```

(iv) The full output from fitting the spherical model is shown below. Identify the estimates of the nugget, partial sill and range parameters. What do these values tell you about the signal to noise ratio in these data. **[4 MARKS]**

```
Summary of the parameter estimation
-----------------------------------
Estimation method: maximum likelihood

Parameters of the mean component (trend):
  beta
2.4309

Parameters of the spatial component:
   correlation function: spherical
      (estimated) variance parameter sigmasq =  2.892
      (estimated) cor. fct. parameter phi  =  0.9304
   anisotropy parameters:
      (fixed) anisotropy angle = 0  ( 0 degrees )
      (fixed) anisotropy ratio = 1
```

CONTINUED OVERLEAF/

```
Parameter of the error component:
      (estimated) =  0.0624

Transformation parameter:
      (fixed) Box-Cox parameter = 1 (no transformation)

Practical Range with cor=0.05 for asymptotic range: 0.930408

Maximised Likelihood:
   log.L n.params      AIC      BIC
"-112.3"       "4"  "232.5"    "243"

non spatial model:
   log.L n.params      AIC      BIC
"-169.9"       "2"  "343.8"    "349"
```

(v) Describe briefly how you could compare the predictive accuracy of the spherical and exponential models, that is the accuracy with which the models predict the magnesium level at unmeasured locations rather than their fit to the observed data. **[3 MARKS]**

(vi) Consider predicting the magnesium concentrations at location $\mathbf{s}_0$, and denote the random variable representing this prediction by $Z(\mathbf{s}_0)$. Further, let $\mathbf{Z}$ denote the vector of random variables representing the observed magnesium concentrations. Finally, let $P_Z(\mathbf{s}_0) = t(\mathbf{Z})$ (for some function $t()$) be a point prediction for $Z(\mathbf{s}_0)$ based on the observed random variables $\mathbf{Z}$. Prove that $\mathbb{E}[(P_Z(\mathbf{s}_0) - Z(\mathbf{s}_0))^2]$ takes its minimum value when $P_Z(\mathbf{s}_0) = \mathbb{E}[Z(\mathbf{s}_0)|\mathbf{Z}]$. **[4 MARKS]**

3. (i) In areal unit data the spatial closeness between the set of m areal units is summarised by a non-negative $m \times m$ neighbourhood or proximity matrix W , where the ijth element of this matrix w_{ij} determines the proximity of areal units (i, j) Describe two different ways in which W can be constructed and give a disadvantage of each approach. **[4 MARKS]**

(ii) Consider random variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$ relating to a set of m areal units, whose spatial adjacency is summarised in a binary $m \times m$ proximity matrix W, where the ijth element of this matrix is w_{ij}. An exploratory measure of spatial autocorrelation for these data is Geary's C statistic, which is given by

$$C = \frac{(m-1)\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(Z_i - Z_j)^2}{2(\sum_{j=1}^{m}\sum_{i=1}^{m} w_{ij})\sum_{i=1}^{m}(Z_i - \bar{Z})^2}.$$

State the values of Geary's C statistic that correspond to positive spatial autocorrelation

CONTINUED OVERLEAF/

and independence, and explain briefly why (those values relate to positive autocorrelation and independence). **[4 MARKS]**

(iii) Consider random variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$ coming from a Gaussian Markov Random Field with $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, Q^{-1})$, where Q is the precision matrix given by $Q = D^{-1}(I - B)$. Here D is a diagonal $m \times m$ matrix whose ith diagonal element is d_{ii}, while B is an $m \times m$ matrix with ijth element b_{ij}. Suppose the vector $\mathbf{Z}$ is partitioned into two components $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. Then partitioning the mean and variance of $\mathbf{Z}$ similarly as

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}^{-1} \right),$$

it can be shown that

$$\mathbf{Z}_1 | \mathbf{Z}_2 \sim \mathrm{N}\left(-Q_{11}^{-1} Q_{12} \mathbf{Z}_2 \,,\; Q_{11}^{-1} \right).$$

(a) Derive the conditions on the elements of (D, B) so that Q is symmetric. **[3 MARKS]**

(b) Suppose W is a symmetric $m \times m$ proximity matrix with zero elements on the leading diagonal (i.e. $w_{ii} = 0$), and let $d_{ii} = \tau^2 / \sum_{j=1}^{m} w_{ij}$ and $b_{ij} = \rho w_{ij} / \sum_{j=1}^{m} w_{ij}$. Show that this specification meets the conditions on the symmetry of Q obtained in the previous part of this question. **[2 MARKS]**.

(c) Using the result above and the specification of (D, B) given in (b), derive the full conditional distribution $f(Z_i | \mathbf{Z}_{-i})$, where $\mathbf{Z}_{-i}$ denotes all observations except the ith. **[6 MARKS]**

(d) Give one disadvantage of the model you have derived in (c) in terms of its ability to model spatial autocorrelation. **[1 MARK]**

4. (i) Briefly describe the differences between Geostatistical data, areal unit data and point process data, and give an example of each type of data. **[6 MARKS]**

(ii) For a spatial domain $D \subset \mathbb{R}^2$, let $A \subset D$ and $Z(A)$ denote the random variable representing the observed number of points in a sub-region of the domain A.

(a) Define the conditions required for Z to be a *homogeneous Poisson process.* **[2 MARKS]**

(b) Write down the general formula for the first order intensity function $\lambda_Z(\boldsymbol{s})$ of a point process at location $\boldsymbol{s}$, and the simplification that arises if the point process is a *homogeneous Poisson process.* **[3 MARKS]**.

CONTINUED OVERLEAF/

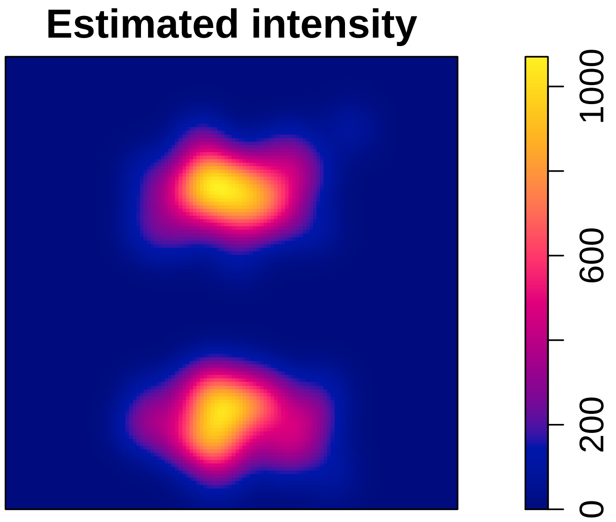
(c) Ripley's K function is given by

$$K_Z(t) \;=\; \frac{\mathbb{E}[Z(\boldsymbol{s}_0, t)]}{\lambda_Z}$$

write down this expression for a *homogeneous Poisson process.* **[1 MARK]**.

(iii) Describe what it means to *thin* a point process, and the difference between *independent* and *dependent* thinning. Suppose you thin a *homogeneous Poisson process.* Does the resulting thinned process remain a *homogeneous Poisson process* in the cases that the thinning is: (a) *independent*, and (b) *dependent*? **[5 MARKS]**

(iv) The first order intensity function was estimated for a spatial point process, and is shown in the figure below. Describe the main features of the spatial point process from the figure. **[3 MARKS]**



Total: 80

END OF QUESTION PAPER.