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 all questions.

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

   (a) Define what it means for the process to be weakly stationary. **[2 MARKS]**

   (b) Define what it means for the process to additionally be isotropic. **[2 MARKS]**

   (c) Define the semi-variogram for a general geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, and show how the definition can be simplified if the process is weakly stationary and isotropic. **[3 MARKS]**

   (ii) Consider the weakly stationary and isotropic covariance function given by:

$$C_Z(h) \quad = \quad \begin{cases} \sigma^2 \frac{\sin(h/\phi)}{h/\phi}, & h \neq 0, \\ \sigma^2 + \tau^2, & h = 0, \end{cases}$$

where h represents distance.

CONTINUED OVERLEAF/

(a) Define what it means for a function $f(.)$ to be an even function, and using the fact that $\sin(-x) = -\sin(x)$ show that the above covariance function $C_Z(h)$ is an even function. **[4 MARKS]**

(b) Compute the semi-variogram for the covariance function $C_Z(h)$ **[3 MARKS]**

(c) Sketch the covariance function $C_Z(h)$ against distance h and say what is unusual about this function compared to most other covariance functions. **[3 MARKS]**

(iii) Consider geostatistical data $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_m))$ and a prediction location $\mathbf{s}_0$. Define the inverse-distance weighted predictor of $Z(\mathbf{s}_0)$ and state two disadvantages of this predictor. **[3 MARKS]**

2. (i) Consider geostatistical data $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_m))$, which are assumed to be multivariate Gaussian with the following mean and covariance matrix:

$$\mathbb{E}[\mathbf{Z}] = \mathbf{X}\boldsymbol{\beta} \quad \mathrm{Var}[\mathbf{Z}] = \Sigma(\boldsymbol{\theta}).$$

Here the covariance matrix is defined by the spherical covariance function, namely:

$$C_Z(h) = \begin{cases} \sigma^2[1 - \frac{3}{2}(h/\phi) + \frac{1}{2}(h/\phi)^3], & 0 < h \leq \phi \\ \tau^2 + \sigma^2, & h = 0 \end{cases},$$

where for simplicty we assume $h < \phi$.

(a) Consider the $m \times m$ distance matrix $\mathbf{D}$, where $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$. Define $\Sigma(\boldsymbol{\theta})$ in terms of $\mathbf{D}$. **[2 MARKS]**

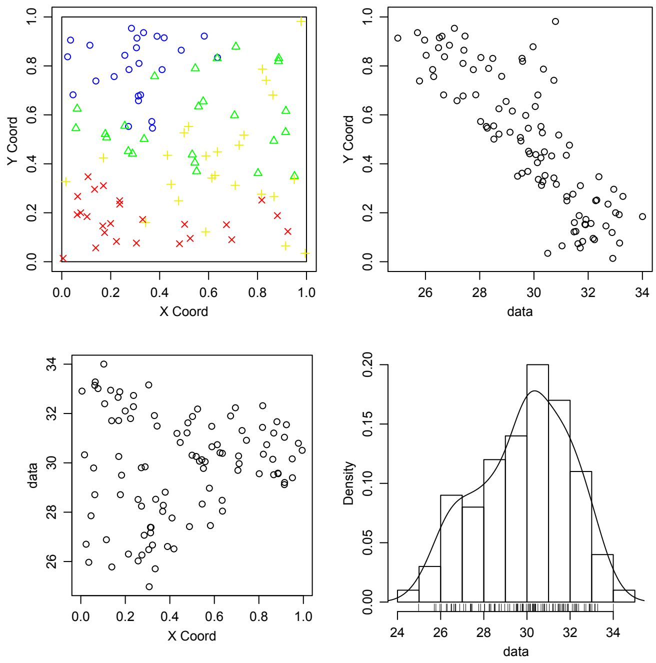
(b) Write down the log-likelihood function for $\mathbf{z}$. **[2 MARKS]**

(c) Derive the maximum likelihood estimator for the regression parameters $\boldsymbol{\beta}$ from this model. **[4 MARKS]**.

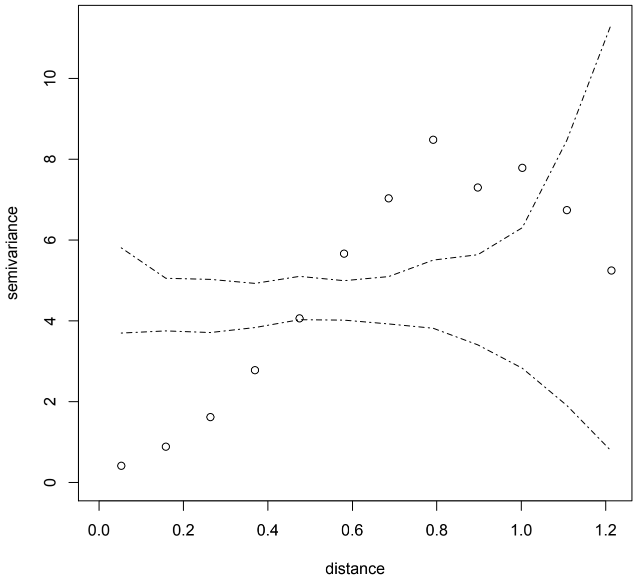
(d) Derive the maximum likelihood estimator for the spatial variance σ^2 from this model. **[4 MARKS]**.

(ii) Nitrogen dioxide concentrations in a 1 km square in Glasgow city were collected for the month of January 2015, and the data are shown on the next page as a `geodata` object.

CONTINUED OVERLEAF/



Below is the estimated empirical semi-variogram for these data, together with 95% Monte Carlo envelopes generated under independence.



(a) From the two plots describe the main features of the data, commenting on the

CONTINUED OVERLEAF/

spatial trend, sampling design and presence of spatial autocorrelation, justify your comments. **[4 MARKS]**

(b) Exponential and Spherical covariance models were fitted to these data, and the resulting output from the `geoR` package is given below:

Exponential

```
Parameters of the spatial component:
   correlation function: exponential
      (estimated) variance parameter sigmasq (partial sill) =  9.176
      (estimated) cor. fct. parameter phi (range parameter)  =  2
   anisotropy parameters:
      (fixed) anisotropy angle = 0  ( 0 degrees )
      (fixed) anisotropy ratio = 1

Parameter of the error component:
      (estimated) nugget =  0.0862

Practical Range with cor=0.05 for asymptotic range: 5.990634

Maximised Likelihood:
   log.L n.params      AIC      BIC
  "-115"      "4"    "238"  "248.4"
```

Spherical

```
Parameters of the spatial component:
   correlation function: spherical
      (estimated) variance parameter sigmasq (partial sill) =  3.07
      (estimated) cor. fct. parameter phi (range parameter)  =  1.044
   anisotropy parameters:
      (fixed) anisotropy angle = 0  ( 0 degrees )
      (fixed) anisotropy ratio = 1

Parameter of the error component:
      (estimated) nugget =  0.0898

Practical Range with cor=0.05 for asymptotic range: 1.043483

Maximised Likelihood:
   log.L n.params      AIC      BIC
"-113.4"      "4"  "234.8"  "245.2"
```

Which of the two models is best for the data and why? **[2 MARKS]**

(c) Does any of the above output summarise which of the two models will provide the best predictions of nitrogen dioxide concentrations at unmeasured locations? If not how could predictive performance be compared? **[2 MARKS]**

3. (i) Given a set of areal unit data $\mathbf{z} = (z_1, \ldots, z_m)$ relating to a set of m areal units, define what is meant by the term *neighbourhood matrix* and why it is needed. Further, define two approaches for defining it and give a drawback of each. **[4 MARKS]**.

(ii) Given a set of areal unit data $\mathbf{z} = (z_1, \ldots, z_m)$ with neighbourhood matrix $\mathbf{W}$, describe a hypothesis test for determining whether the data exhibit spatial autocorrelation. Define the null and alternative hypotheses, the test statistic and briefly how the p-value of the test is computed. **[4 MARKS]**

(iii) Consider random variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$ coming from a Gaussian model $\mathbf{Z} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Q}^{-1})$, where $\mathbf{Q}$ is the precision matrix and $\boldsymbol{\Sigma} = \mathbf{Q}^{-1}$ is the variance matrix. 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

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mathbf{X}_1\boldsymbol{\beta} \\ \mathbf{X}_2\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$$

it can be shown that the conditional distribution of $\mathbf{Z}_1|\mathbf{Z}_2$ is given by

$$\mathbf{Z}_1|\mathbf{Z}_2 \sim \text{N}\left(\mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Z}_2 - \mathbf{X}_2\boldsymbol{\beta}) \ , \ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

Finally, let the precision matrix $\mathbf{Q}$ be partitioned similarly as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}$$

and the relationships between the elements can be written as:

1. $\boldsymbol{\Sigma}_{11} = \mathbf{Q}_{11}^{-1}(\mathbf{I} - \mathbf{Q}_{12}\boldsymbol{\Sigma}_{21})$.
2. $\boldsymbol{\Sigma}_{12} = -\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}\boldsymbol{\Sigma}_{22}$.

(a) Derive formulae for the conditional mean and variance of $\mathbf{Z}_1|\mathbf{Z}_2$ so that they do not depend on any elements of the covariance matrix $\boldsymbol{\Sigma}$ and instead depend only on elements of the precision matrix $\mathbf{Q}$. **[4 MARKS]**

(b) Computationally, why might one wish to use the formulation involving $\mathbf{Q}$ only rather than the formulation stated in the question involving $\boldsymbol{\Sigma}$? **[1 MARK]**

CONTINUED OVERLEAF/

(c) Suppose $\mathbf{Q} = \text{diag}(\mathbf{W1}) - \mathbf{W}$, where $\mathbf{W}$ is an $m \times m$ neighbourhood matrix and $\mathbf{1}$ is an $m \times 1$ vector of ones. Using the result from (a), derive the conditional distribution of $Z_i|\mathbf{Z}_{-i}$, where $\mathbf{Z}_{-i}$ denotes all data points except the ith. **[4 MARKS]**

(d) Give two downsides to the model defined in (c) and briefly describe how it could be extended to a more general model that could encompass different types of spatial autocorrelation. **[3 MARKS]**

Total: 60

END OF QUESTION PAPER.