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) Suppose $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ ($D \subset \mathbb{R}^2$) is a geostatistical process with covariance function $\mathcal{C}_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) = \mathrm{Cov}(Z(\boldsymbol{s}), Z(\boldsymbol{s}+\boldsymbol{h}))$.

   (a) Define the conditions under which the geostatistical process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is weakly stationary and isotropic. **[4 MARKS]**

   (b) Define what it means for a function $f(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h})$ to be *nonnegative definite* and prove that the covariance function $\mathcal{C}_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h})$ is a *nonnegative definite* function. **[5 MARKS]**

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

$$C_Z(h) = \begin{cases} \sigma^2 \exp(-h/\phi), & h > 0, \\ \tau^2 + \sigma^2, & h = 0, \end{cases}$$

where h represents distance.

CONTINUED OVERLEAF/

(a) Define the nugget, partial sill, sill and range parameters in this model. **[3 MARKS]**

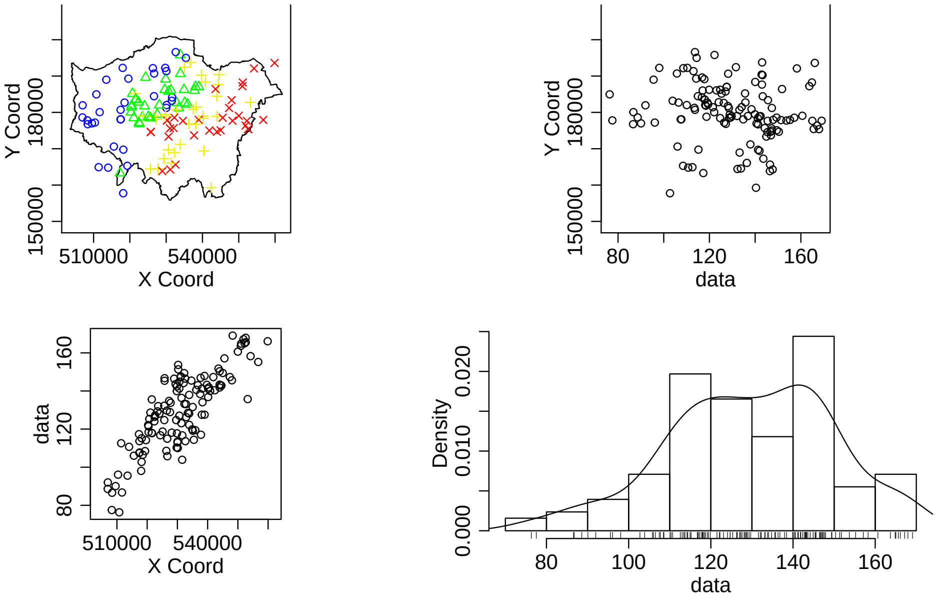
(b) Write down the formula for the variogram of a weakly stationary and isotropic geostatistical process in terms of the covariance function. Use this formula to derive the variogram for the exponential covariance model. **[3 MARKS]**

(iii) Consider a weakly stationary and isotropic process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ with constant mean μ_Z and covariance function $\mathcal{C}_Z(\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}) = \sigma^2 C(||\boldsymbol{h}||)$ for some correlation function $C(.)$ satisfying $C(0) = 1$. Now consider a second process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, which is a white noise process with mean 0 and variance 1. In addition, $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ and $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ are independent of each other. Derive the mean, variance and covariance function for the new process

$$X(\boldsymbol{s}) = a + bZ(\boldsymbol{s}) + cY(\boldsymbol{s})$$

for scalars (a, b, c). Is this process second order stationary and isotropic? **[5 MARKS]**

2. (i) Measurements were collected at 127 locations in and around Greater London, and are plotted as a geoR object below. The X co-ordinate represents easting in metres and the Y co-ordinate represents northing in metres. Describe the main features of these data, and comment on whether the assumption that they are weakly stationary is reasonable? **[2 MARKS]**



(ii) Constant, linear and quadratic trend models in easting (X Coord) and northing (Y

CONTINUED OVERLEAF/

Coord) were fitted to these data using the linear model framework, and the results of each are shown below. From the output shown which is the most appropriate trend model for these data? Justify your answer. **[2 MARKS]**

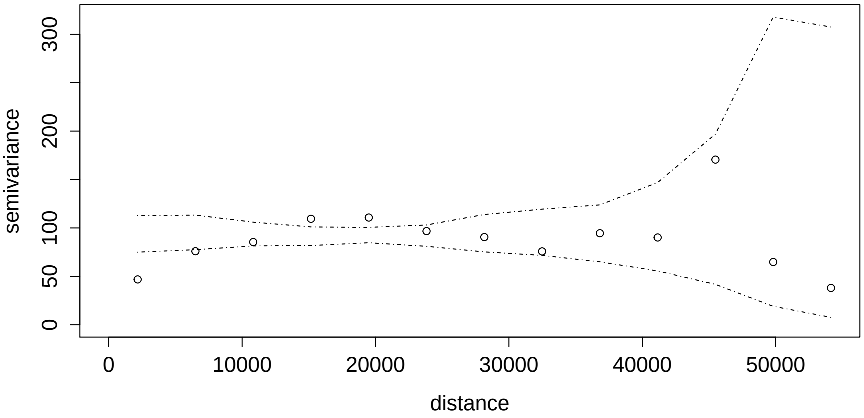
```
Model 1
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  130.975      1.081   121.2   <2e-16 ***
Residual standard error: 12.18 on 126 degrees of freedom


Model 2
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.002e+02  1.352e+01 -29.598   <2e-16 ***
easting      1.003e+00  2.462e-02  40.746   <2e-16 ***
northing    -8.616e-03  3.603e-02  -0.239    0.811
Residual standard error: 3.191 on 124 degrees of freedom
Multiple R-squared:  0.9324,Adjusted R-squared:  0.9313


Model 3
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.565e+02  4.853e+02  -0.941    0.349
easting      9.029e-01  1.757e+00   0.514    0.608
northing     9.218e-01  1.155e+00   0.798    0.427
easting.sq   9.443e-05  1.652e-03   0.057    0.954
northing.sq -2.607e-03  3.236e-03  -0.806    0.422
Residual standard error: 3.209 on 122 degrees of freedom
Multiple R-squared:  0.9328,Adjusted R-squared:  0.9306
```

(iii) The residuals from the most appropriate trend model were computed and the empirical semi-variogram was then plotted together with 95% Monte Carlo envelopes. This plot is shown below.

CONTINUED OVERLEAF/



(a) Write down the formula for the empirical semi-variogram and describe two disadvantages with its estimation **[3 MARKS]**

(b) From the empirical semi-variogram plot of the data approximately estimate the nugget, sill and range of the process. **[3 MARKS]**

(c) Describe briefly how the Monte Carlo envelopes are created for this semi-variogram, and how one should use them to interpret the presence of spatial autocorrelation. What do you conclude about the presence of spatial autocorrelation in the residuals? **[4 MARKS]**

(iv) What does the term Kriging mean in a geostatistical context, and in what sense is it an optimal spatial predictor? Describe the difference between ordinary and universal Kriging. **[3 MARKS]**

(v) Kriging can be undertaken using either likelihood or Bayesian inference. Describe the main difference between these two approaches to Kriging and give an advantage of each method **[3 MARKS]**

3. (i) 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 $m \times m$ neighbourhood matrix W. An exploratory measure of spatial autocorrelation for these data is Moran's I statistic, which is given by

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

CONTINUED OVERLEAF/

(a) Describe briefly how the Moran's I statistic measures spatial autocorrelation, and state the range of values it can take. **[3 MARKS]**

(b) Describe how Moran's I statistic can be used to test for the presence of spatial autocorrelation in areal unit data and write down the null and alternative hypotheses. **[3 MARKS]**

(c) Describe how Moran's I statistic can be transformed to a Local Indicator of Spatial Association (LISA) and describe what such an indicator measures. **[2 MARKS]**

(ii) Consider the vector of random variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$, which are assigned a Gaussian Markov Random Field (GMRF) model, with mean $\mathbf{m}$ and precision matrix τQ. 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

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

it can be shown that

$$\mathbf{Z}_1 | \mathbf{Z}_2 \sim \mathrm{N}\left(\mathbf{m}_1 - Q_{11}^{-1} Q_{12} (\mathbf{Z}_2 - \mathbf{m}_2)\ ,\ [\tau Q_{11}]^{-1} \right)$$

Consider a GMRF model with $\mathbf{m} = \mu \mathbf{1}$ and $Q = \rho[\mathrm{diag}(W\mathbf{1}) - W] + (1 - \rho)I$, where $\mathbf{1}$ is an $m \times 1$ vector of ones and W is an $m \times m$ neighbourhood matrix.

(a) Using the above result derive the full conditional distribution $f(Z_i | \mathbf{Z}_{-i})$, where $\mathbf{Z}_{-i}$ denotes all observations except the ith. **[4 MARKS]**

(b) Derive the partial correlation $\mathrm{Corr}(Z_i, Z_j | \mathbf{Z}_{-ij})$, where $\mathbf{Z}_{-ij}$ denotes all observations except the ith and jth. **[4 MARKS]**

(c) The GMRF model described above was fitted to data on respiratory hospital admissions in Glasgow in 2007, and the covariates included a measure of socio-economic deprivation and a measure of particulate matter air pollution. The model was fitted in two forms, first with the spatial dependence parameter ρ being estimated from the data and again with ρ fixed to equal one. The models were fitted using the CARBayes software and the output from fitting both these models is shown below.

```
Model 1 - Intrinsic CAR

Posterior quantiles and DIC
              Median    2.5%   97.5% n.sample % accept
(Intercept) -0.9655 -1.2992 -0.6368    10000     60.8
pollution    0.0223 -0.0043  0.0490    10000     60.8
```

CONTINUED OVERLEAF/

```
deprivation  0.0946  0.0856  0.1056    10000     60.8
tau2         0.1045  0.0792  0.1382    10000    100.0
DIC =  2135.702       p.d =  172.4182


Model 2  - Leroux CAR

Posterior quantiles and DIC
             Median    2.5%   97.5% n.sample % accept
(Intercept) -0.9499 -1.1928 -0.6741    10000     61.2
pollution    0.0197 -0.0026  0.0391    10000     61.2
deprivation  0.0989  0.0872  0.1106    10000     61.2
tau2         0.0644  0.0430  0.0970    10000    100.0
rho          0.3800  0.1500  0.7000    10000     58.5
DIC =  2134.543       p.d =  181.6448
```

Which of the two models appears to best fit the data and why, and do either of the covariates exhibit any evidence of a relationship with the risk of respiratory hospital admission? **[4 MARKS]**

4. 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.

(i) Describe how spatial point process data differs from geostatistical data, and define what is meant by the terms *marked point process* and *unmarked point process.* **[3 MARKS]**

(ii) Write down the formula for the first order intensity function $\lambda_Z(\boldsymbol{s})$ of the point process at location $\boldsymbol{s}$, and then define the expectation $\mathbb{E}[Z(A)]$ in terms of it. **[2 MARKS]**

(iii) Describe in words what it means for a point process to exhibit *Complete Spatial Randomness (CSR)* and write down the specification for a *homogeneous Poisson process*, which is a commonly used model for CSR. **[3 MARKS]**

(iv) Show that the first order intensity function $\lambda_Z(\boldsymbol{s})$ for a *homogeneous Poisson process* is constant in space. **[2 MARKS]**

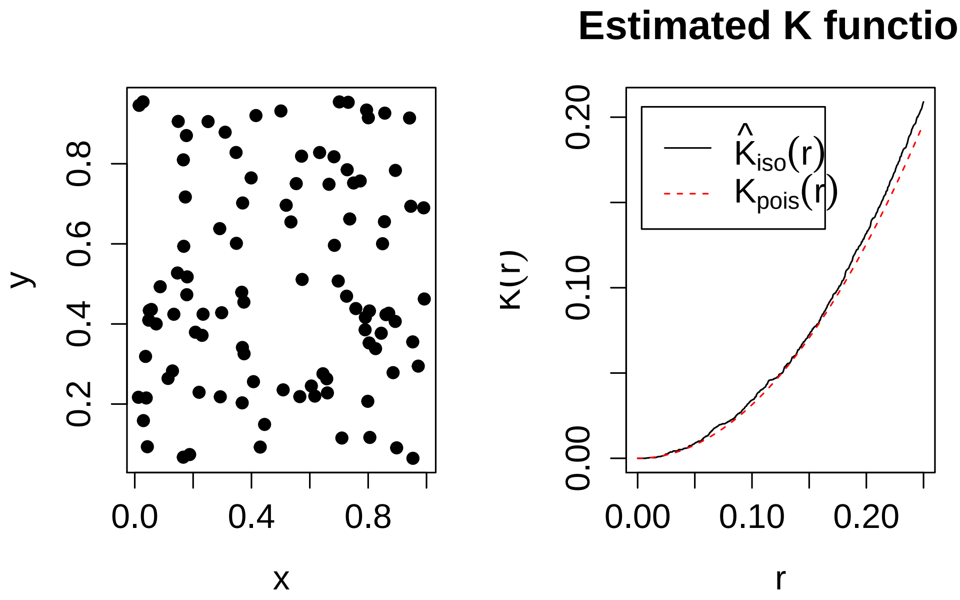
(v) For an arbitrary spatial point process, define Ripley's K function in terms of the pair correlation function $\rho(t)$. Consider a spatial point process with $\rho(t) = 1$ for all

CONTINUED OVERLEAF/

t, derive Ripley's K function for this process. What process does this correspond to? **[4 MARKS]**

(vi) Describe how Ripley's K function can be used to determine whether a given point process is a completely spatially random, clustered or regular process. **[3 MARKS]**

(vii) The data below are the locations of trees in a 1km square grid on Exmoor in Somerset. From looking at both plots below is there any evidence that the trees exhibit spatial dependence? Justify your answer. **[3 MARKS]**



Total: 80

END OF QUESTION PAPER.