University of Glasgow

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 $C_Z(\mathbf{s}, \mathbf{s}+\mathbf{h})$. Define what it means for the process to be weakly stationary and isotropic. **[3 MARKS]**

(ii) Now suppose the process has mean function $\mu_Z(\boldsymbol{s}) = 0$ and a covariance function given by

$$C_Z(h) = \begin{cases} \sigma^2, & h = 0, \\ d, & h \neq 0, \end{cases}$$

where $0 \leq d \leq \sigma^2$.

(a) Derive the correlation function for $\rho_Z(h)$. **[2 MARKS]**

(b) Define the semi-variogram for a general geostatistical process, and derive it for the process $Z(\mathbf{s})$ defined above. **[3 MARKS]**

CONTINUED OVERLEAF/

(c) Sketch the semi-variogram and covariance functions for the process $Z(\mathbf{s})$ defined above. **[3 MARKS]**

(d) Give 2 reasons why this covariance model is not a sensible one for spatial data. **[3 MARKS]**

(iii) Consider two independent weakly stationary and isotropic geostatistical processes $(X(\mathbf{s}), Y(\mathbf{s}))$, which have means $(0,0)$ and exponential covariance functions. The parameters of the covariance functions are: partial sill - (σ_x^2, σ_y^2); nugget - $(0,0)$; range parameter - ϕ common to both $(X(\mathbf{s}), Y(\mathbf{s}))$. Consider the process $Z(\mathbf{s}) = X(\mathbf{s}) + Y(\mathbf{s})$.

(a) Derive the mean and covariance function for $Z(\mathbf{s})$. What type of covariance model does $Z(\mathbf{s})$ have? **[4 MARKS]**

(b) If data were collected and modelled using the process $Z(\mathbf{s})$, can all the parameters $(\sigma_x^2, \sigma_y^2, \phi)$ be reliably estimated? Explain your answer. **[2 MARKS]**

2. (i) A researcher wishes to compare two different geostatistical models denoted A and B.

(a) Define Akaike's Information Criterion (AIC) and describe what aspect of a model it quantifies. **[2 MARKS]**

(b) Describe how one can use a leave-one-out cross validation approach for assessing predictive accuracy of a model. Ensure you define how leave-one-out cross validation is implemented and how you quantify the quality of the predictions. **[3 MARKS]**

(c) The researcher finds that model A has a lower AIC value but model B does better in terms of predictive accuracy via leave-one-out cross validation. Which model should the researcher use? **[2 MARKS]**

(ii) Data were collected on the depth of peat at 718 locations on Dartmoor in Devon, and exploratory plots of the data are shown in Figure 1.

(a) Should the peat depth data be transformed before being modelled with a Gaussian geostatistical model, and if so why and what transformation would you suggest? **[2 MARKS]**

(b) Describe what problems the sampling design may cause for predicting peat depth at unmeasured locations across the square study region shown in the top left panel in Figure 1. **[2 MARKS]**

(iii) Consider a spatial point process $Z = \{Z(A) : A \subset D\}$.

CONTINUED OVERLEAF/

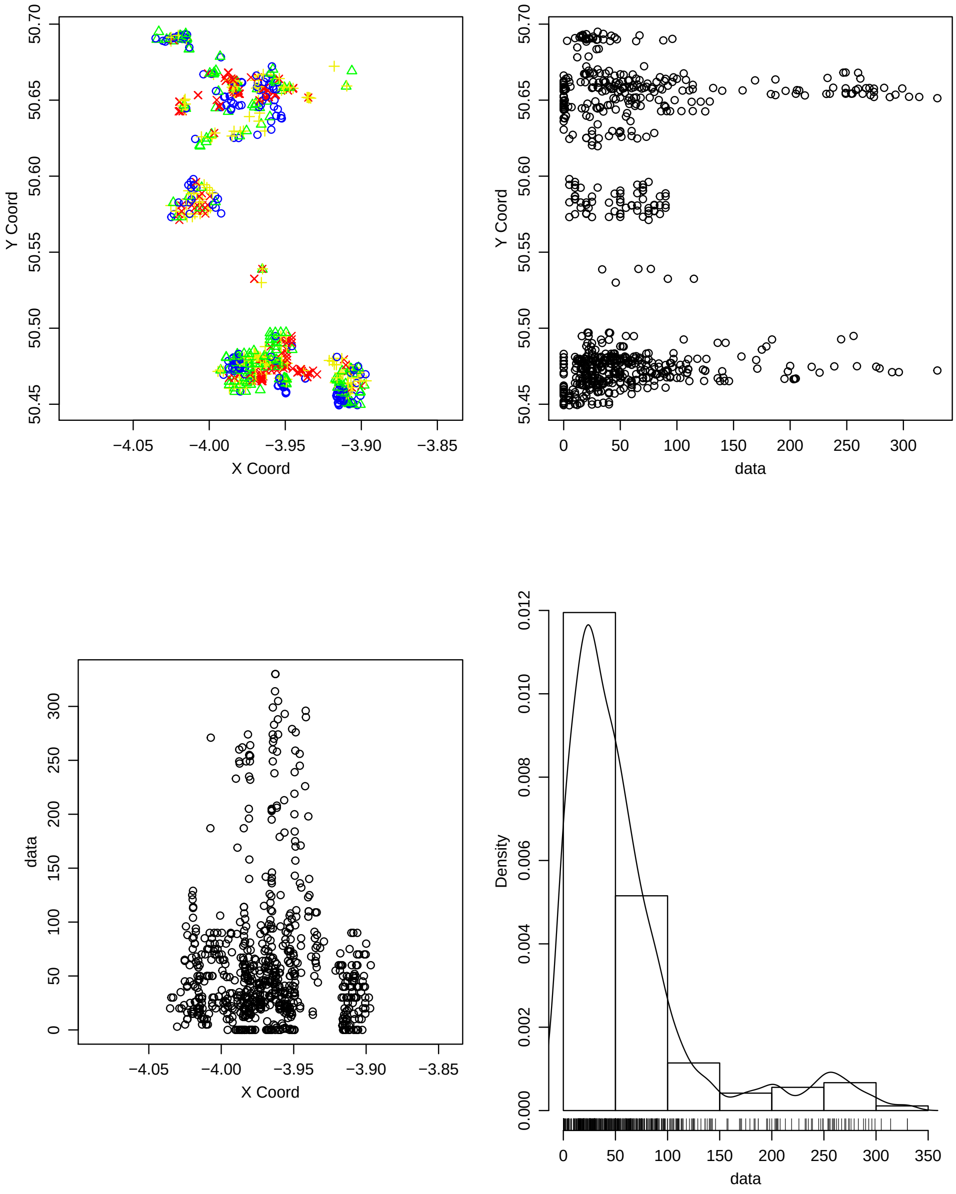

Figure 1: Exploratory plots of the peat depth data

CONTINUED OVERLEAF/

(a) Define a homogeneous Poisson process and briefly describe why it corresponds to complete spatial randomness. **[2 MARKS]**

(b) Define the first and second order intensity functions for a point process. **[2 MARKS]**

(c) The pair correlation function is defined to be

$$\rho(\mathbf{s}, \mathbf{t}) = \frac{\lambda_{2,Z}(\mathbf{s}, \mathbf{t})}{\lambda_Z(\mathbf{s})\lambda_Z(\mathbf{t})},$$

where $(\lambda_Z(\mathbf{s}), \lambda_{2,Z}(\mathbf{s}, \mathbf{t}))$ are respectively defined to be the first and second order intensity functions. Define Ripley's K function in terms of the pair correlation function, and compute Ripley's K function for a homogeneous Poisson process. Ensure you calculate the first and second order intensity functions for the homogeneous Poisson process. **[5 MARKS]**

3. (i) Consider data $\mathbf{z} = (z_1, \ldots, z_n)$ relating to a set of n areal units, and a corresponding $n \times n$ binary neighbourhood matrix $\mathbf{W}$ defined by geographical contiguity. Define Moran's I statistic and describe why it measures spatial autocorrelation. Describe a hypothesis test using Morans I for testing whether a data set contains spatial autocorrelation including a description of the hypotheses, test statistic and how the p-value is calculated. **[4 MARKS]**

(ii) Consider modelling data from a set of n areal units by the process-convolution model given by

$$Z_k = \sum_{j=1}^{n} w_{kj}\theta_j, \qquad \theta_j \sim \mathrm{N}(0, \tau^2),$$

where w_{kj} is the (k, j)th element of the neighbourhood matrix $\mathbf{W}$ and each θ_j is independent of all the other elements θ_k where $k \neq j$.

(a) Briefly describe why this process-convolution model induces spatial autocorrelation into $\mathbf{Z}$ despite $\boldsymbol{\theta}$ being a vector of independent random variables. **[2 MARKS]**

(b) Write down a formula for the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ in terms of $\mathbf{W}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. **[1 MARKS]**

(c) Derive the joint distribution of $\mathbf{Z} = (Z_1, \ldots, Z_n)$. What is the mean and variance of Z_k and the covariance and correlation between (Z_k, Z_j)? **[4 MARKS]**

(iii) Data were obtained for the $n = 624$ electoral wards in Greater London for 2009 on the observed (`y`) and expected (computed by indirect standardisation) (`e`), numbers of

CONTINUED OVERLEAF/

admissions to hospital due to respiratory disease. Also collected were two covariates, the percentage of people defined to be poor (`poor`) in each area, and the average air pollution concentrations (`pollution`) in each area.

(a) An initial simple Poisson generalised linear model was fitted to the hospital admission counts (`y`), with both covariates and the (log) expected numbers of admissions as a known offset term. The residuals were then tested for the presence of spatial autocorrelation, and the results of a Moran's I test are shown below.

```
Monte-Carlo simulation of Moran I

data:  res
weights: W.list
number of simulations + 1: 1001

statistic = 0.39417, observed rank = 1001, p-value = 0.000999
alternative hypothesis: greater
```

What does this test tell you about the presence or absence of residual spatial autocorrelation? Justify your answer. **[2 MARKS]**

(b) A Poisson log-linear CAR model was then fitted to these data, where the linear predictor contained a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ in addition to the covariates. The CAR model used has full conditional distributions given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathrm{N}\left(\frac{\rho\sum_{j=1}^{n} w_{ij}\phi_j}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + (1-\rho)}\right).$$

Output from fitting this model is shown below.

```
############
#### Results
############
Posterior quantities

             Median    2.5%   97.5% n.sample % accept n.effective Geweke.diag
(Intercept) -0.8512 -1.2169 -0.4939     9000     29.9      1434.0         0.3
pollution    0.0074 -0.0104  0.0258     9000     29.9      1644.2        -0.4
poor         0.0267  0.0239  0.0295     9000     29.9      3283.9         0.3
tau2         0.0798  0.0670  0.0945     9000    100.0      9000.0        -1.6
rho          0.8324  0.6825  0.9492     9000     43.9      9000.0        -1.6
```

What does the estimated value of ρ tell you about the level of residual spatial autocorrelation after adjusting for the covariates? Justify your answer **[2 MARKS]**

(c) Calculate separate relative risks and 95% credible intervals for increases in each covariate by 1 unit and interpret the results. **[4 MARKS]**

CONTINUED OVERLEAF/

(d) How could the CAR model given above be simplified to the intrinsic CAR model for strong spatial autocorrelation? **[1 MARK]**

Total: 60

END OF QUESTION PAPER.