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) Consider a geostatistical process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, where D is the domain of interest, with mean function $\mu_Z(\mathbf{s})$ and covariance function $\mathcal{C}_Z(\mathbf{s}, \mathbf{s}+\mathbf{h})$.

(a) Define what it means for $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ to be weakly stationary and isotropic. **[3 MARKS]**

(b) Define the theoretical semi-variogram for a weakly stationary and isotropic geostatistical process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$. Sketch the general shape of the theoretical semi-variogram, making sure you identify (label) the sill, nugget and range. **[4 MARKS]**

(c) Define mathematically the binned empirical semi-variogram, and explain briefly how it can be used to check the assumption of isotropy. **[3 MARKS]**

CONTINUED OVERLEAF/

(ii) Consider the following semi-variogram model for an isotropic geostatistical process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$,

$$\gamma_Z(h) = \begin{cases} 0 & h = 0, \\ h^2 & h > 0, \end{cases}$$

which is accompanied by the mean model $\mu_Z(\mathbf{s}) = 0$ for all $\mathbf{s} \in D$. Is this geostatistical process weakly stationary? Justify your answer. **[3 MARKS]**

(iii) Data were collected on summer-time average pollen levels at 60 sites in Scotland. A researcher modelled these data, and came up with 4 candidate models that she thought represented the data well.

(a) How should she choose the best model if her goal is to explain the spatial variation in the observed data. **[2 MARKS]**

(b) How should she choose the best model if her goal is to predict the pollen levels at unmeasured locations. **[2 MARKS]**

(c) The researcher now wishes to predict pollen levels at an unmeasured location $\mathbf{s}_0$, given measurements at locations $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_{60}))$. She proposes to use the following predictor of $Z(\mathbf{s}_0)$

$$P_{\mathbf{z}}(\mathbf{s}_0) = \frac{\sum_{k=1}^{60} \exp(-d_{k0}) z(\boldsymbol{s}_k)}{\sum_{k=1}^{60} \exp(-d_{k0})},$$

where $d_{k0} = ||\mathbf{s}_k - \mathbf{s}_0||$. How do the weights change with distance d_{ko} from the prediction location $\mathbf{s}_0$, and are these weights sensible in this regard? Additionally, name one downside with this predictor $P_{\mathbf{z}}(\mathbf{s}_0)$. **[3 MARKS]**

2. (i) Consider a vector of areal unit data $\mathbf{Z} = (Z_1, \ldots, Z_n)$ relating to n non-overlapping areal units. Additionally, consider a binary $n \times n$ neighbourhood matrix $\mathbf{W}$, where $w_{kj} = 1$ if areas (k, j) share a common border and $w_{kj} = 0$ otherwise.

(a) Define Geary's C function, and explain which values correspond to spatial autocorrelation and which values correspond to independence. **[3 MARKS]**

(b) Now consider the following model for $\mathbf{Z}$.

$$Z_k | \mathbf{Z}_{-k} \sim \mathrm{N}\left(\frac{\sum_{j=1}^{n} w_{kj} Z_j}{\sum_{j=1}^{n} w_{kj}}, \frac{\tau^2}{\sum_{j=1}^{n} w_{kj}} \right),$$

where in the usual notation $\mathbf{Z}_{-k}$ denotes all the observations except the kth. What type of model is this and give two limitations of it. **[3 MARKS]**

CONTINUED OVERLEAF/

(c) Now suppose that one of the areal units is an island, and hence does not share a common border with any of the other areas. Given the definition of the neighbourhood matrix $\mathbf{W}$ above, is the model described in the previous part a valid model? Justify your answer. If it is not a valid model, how could $\mathbf{W}$ be altered to make it a valid model? **[4 MARKS]**

(d) Now suppose the following alternative model is used for the data $\mathbf{Z}$:

$$\mathbf{Z} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \tau^2[diag(\mathbf{W1}) - \rho\mathbf{W}]^{-1}\right),$$

where $diag(\mathbf{W1})$ denotes a diagonal matrix, whose kth diagonal element is $\sum_{j=1}^{n} w_{kj}$. Additionally, $\mathbf{X}$ is an $n \times p$ matrix of known covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters, and ρ is a scalar dependence parameter. Compute the full conditional distribution $f(Z_k|\mathbf{Z}_{-k})$, where $\mathbf{Z}_{-k}$ denotes all the elements except Z_k. **[6 MARKS]**

Hint - You may use the result that if $\mathbf{Z}$ is partitioned into two components, that is, $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ and its joint distribution is given by

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

then the conditional distribution of $\mathbf{Z}_1|\mathbf{Z}_2$ is given by

$$\mathbf{Z}_1|\mathbf{Z}_2 \sim \mathrm{N}\left(\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\mathbf{Z}_2 - \boldsymbol{\mu}_2)\ ,\ \mathbf{Q}_{11}^{-1}\right)$$

Here $\mathbf{Q}$ is the precision and not the covariance matrix.

(ii) The Poisson log-linear Leroux CAR model is fitted to a data set on coronary heart disease counts in the $n = 271$ intermediate zones that make up the Greater Glasgow and Clyde health board.

(a) The posterior median and 95% credible interval for the spatial dependence parameter in the Leroux CAR model were: ρ 0.921 (0.891, 0.983). What does this tell you about the level of spatial autocorrelation in the data? **[2 MARKS]**

(b) Particulate matter air pollution was included as a covariate in the model for coronary heart disease, and its parameter estimate and 95% credible interval on the linear predictor scale (log-risk scale) are given by: β 0.00234 (0.00167, 0.00297). Compute the relative risk for coronary heart disease for a 1 unit increase in particulate matter concentrations and interpret the result. **[2 MARKS]**

CONTINUED OVERLEAF/

3. (i) Describe briefly what it means for a spatial point process to be: (1) Completely spatially random; (2) a regular process; and (3) a clustered process. **[3 MARKS]**

(ii) Describe briefly the difference between a **marked** and an **unmarked** spatial point process. **[2 MARKS]**

(iii) Consider a spatial point process $Z = \{Z(A) : A \subset D\}$ (where D is the domain of interest) with first-order intensity function $\lambda(\mathbf{s})$, where $\mathbf{s} \in D$.

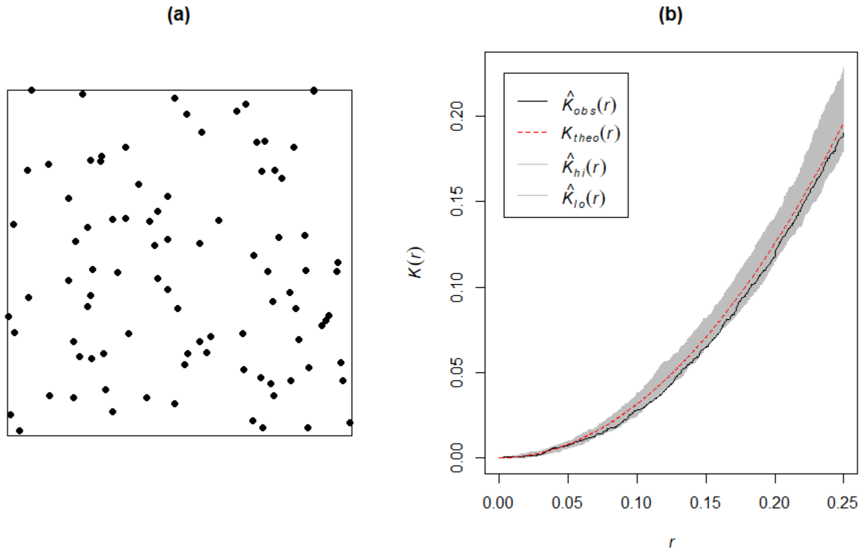
(a) For the process Z it is believed that the first order intensity function is constant in space and that the points occur independently (randomly) in space. Name and write down the mathematical definition for a model for the spatial point process Z that meets these assumptions. **[3 MARKS]**

(b) Now suppose that the first first order intensity function is assumed to vary in space, but that points still occur independently (randomly) in space. Name and write down a model for the spatial point process Z that meets these assumptions. **[3 MARKS]**.

(c) Now suppose it was believed that Z may be a clustered point process. Define Ripley's K function and describe how it could be used to determine if Z was clustered or completely spatially random. **[3 MARKS]**

(iv) A spatial point pattern is visualised below, where the left panel (a) displays the locations of the points, while the right panel (b) displays a plot of distance r vs $K(r)$, where $K(r)$ is Ripley's K function. The dotted line is the estimated $K(r)$ function, the solid line is πr^2, and the grey shading is a Monte Carlo envelope generated under independence.

CONTINUED OVERLEAF/



Is the spatial point pattern completely spatially random? Justify your answer. [**3 MARKS**]

(v) Consider a spatial point process $Z = \{Z(A) : A \subset D\}$, where D is the domain of interest, and $\lambda(\mathbf{s})$ ($\mathbf{s} \in D$) is the first order intensity function. Now consider the general model

$$Z(A) \sim \text{Poisson}\left(\int_{\mathbf{s}\in A} \lambda(\mathbf{s})ds\right).$$

The following two candidate models are considered for modelling the first order intensity function at n observed point locations $(\lambda(\mathbf{s}_1), \ldots, \lambda(\mathbf{s}_n))$.

(a) $\lambda(\mathbf{s}_j) \sim \text{Gamma}(1,1)$ for $j = 1, \ldots, n$.

(b) $\lambda(\mathbf{s}_j) \sim \text{N}(0,1)$ for $j = 1, \ldots, n$.

Which of the above models (a) and (b) are valid models for the first order intensity function and which induce spatial autocorrelation (clustering) into the spatial point process? Justify your answers. [**3 MARKS**]

Total: 60

END OF QUESTION PAPER.