University of Glasgow

SOME EXAM DATE
SOME EXAM TIME

EXAMINATION FOR THE DEGREES OF XXXX

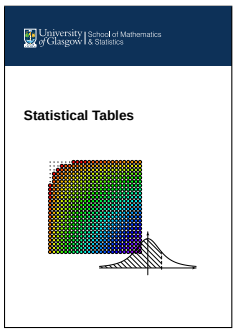
STATISTICS
Spatial Statistics 5M

This paper consists of 5 pages and contains 4 questions.
Candidates should attempt all questions.

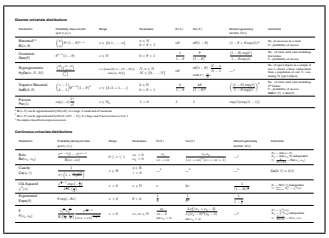
Question 1	20 marks
Question 2	20 marks
Question 3	20 marks
Question 4	20 marks
Total	80 marks

The following material is made available to you:

Statistical tables*



Probability formula sheet



"An electronic calculator may be used provided that it is allowed under the School of Mathematics and Statistics Calculator Policy. A copy of this policy has been distributed to the class prior to the exam and is also available via the invigilator."

CONTINUED OVERLEAF/

"An electronic calculator may be used provided that it is allowed under the School of Mathematics and Statistics Calculator Policy. A copy of this policy has been distributed to the class prior to the exam and is also available via the invigilator."

NOTE: Candidates should attempt all questions.

1. (a) An environmental scientist is studying yearly average nitrogen dioxide (NO_2) concentrations, which is a non-negative measure of air pollution. The scientist has made measurements at 100 locations $\mathbf{z} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_{100}))$ across Glasgow, and their only goal in analysing these data is to produce a map of NO_2 predictions at 1 kilometre intervals across the city.

   i. The scientist is worried that as the concentrations are non-negative and skewed to the right that a Gaussian geostatistical model would be inappropriate. What advice would you give her to overcome this problem? **[2 MARKS]**

   ii. Describe briefly how the scientist would assess the data for the presence of residual spatial autocorrelation? **[2 MARKS]**

   iii. The scientist is considering two different geostatistical models for her data, and decides to choose the one that minimises the Bayesian Information Criterion (BIC). Why is this not a good criteria to use to select the best model given the goal of her analysis? **[2 MARKS]**

   iv. Briefly describe an alternative approach to how she should choose the best model given the goal of her analysis. Briefly describe how this approach works, and define 3 criteria she could use to assess the predictive fit of each model considered. **[5 MARKS]**

   (b) Consider the zero-mean geostatistical process $\{Z(\mathbf{s})|\mathbf{s} \in \mathcal{D}\}$ with a weakly stationary and isotropic covariance function given by

$$C(h) = \begin{cases} \xi^2(1+\rho h)\exp(-\rho h), & h > 0 \\ \nu^2 + \xi^2, & \text{h=0.} \end{cases}$$

   i. Compute the semi-variogram for the geostatistical process $\{Z(\mathbf{s})|\mathbf{s} \in \mathcal{D}\}$. **[3 MARKS]**

   ii. What are the nugget, sill and partial sill for this covariance model? Justify your answer. **[3 MARKS]**

   iii. Would the slightly altered covariance function defined below be a good model for spatial data for $\phi > 0$? Justify your answer. **[3 MARKS]**

CONTINUED OVERLEAF/

$$C(h) \begin{cases} \xi^2(1+\rho h)\exp(-\rho h) + \phi, & h > 0 \\ \nu^2 + \xi^2 + \phi, & \text{h=0.} \end{cases}$$

2. (a) Consider an areal unit process $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$ relating to n unevenly sized areal units with centroids (central points) $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$. One could model these data as a geostatistical process, where the central points $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ represent the spatial locations of the areal units. This would allow a geostatistical covariance function to be specified for this process, such as the exponential model given by $\text{Covariance}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = \sigma^2 \exp(-||\mathbf{s}_i - \mathbf{s}_j||/\phi)$ where $||.||$ denotes Euclidean distance. Is this likely to be a good representation of spatial correlation for the areal unit process described above? Justify your answer. **[3 MARKS]**

   (b) Suppose the areal unit process $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$ follows a zero-mean multivariate Gaussian distribution. A geostatistical style model would be parameterised via the covariance $\mathbf{\Sigma} = \text{Covariance}(\mathbf{Z})$, while an areal unit style model would be parameterised via the precision (inverse of the covariance) $\boldsymbol{Q} = \text{Precision}(\mathbf{Z})$. Which representation is faster computationally for evaluating the multivariate Gaussian data likelihood? Justify your answer. **[2 MARKS]**

   (c) Consider a simple 1 dimensional areal unit process with 4 regions ordered as $[A|B|C|D]$, with a corresponding neighbourhood matrix

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

   so that the only neighbour pairs are (A, B), (B, C) and (C, D). Then suppose that $Z(A) = 6$, $Z(B) = 5$, $Z(C) = 4$ and $Z(D) = 3$. Compute Geary's C statistic and describe what it tells you about the presence/absence of spatial correlation in these data. **[4 MARKS]**

   (d) Consider an areal unit process $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$ with a corresponding binary $n \times n$ neighbourhood matrix $\mathbf{W}$, where $w_{ij} = 1$ if areas (i, j) are spatial neighbours (share a common border) and $w_{ij} = 0$ otherwise. A conditional autoregressive (CAR) model for $\mathbf{Z}$ has the general form $\mathbf{Z} \sim \text{N}(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W})^{-1})$, where τ^2 is a variance parameter and $\mathbf{Q}(\mathbf{W})$ is a precision matrix based on the neighbourhood matrix $\mathbf{W}$.

      i. Define $\mathbf{Q}(\mathbf{W})$ mathematically for the *intrinsic CAR (ICAR)* model and write down formulae for: (i) the diagonal element Q_{ii}; and (ii) the off-diagonal element Q_{ij} where $i \neq j$. **[3 MARKS]**

CONTINUED OVERLEAF/

ii. From part i. above what does this tell you about the partial (conditional) correlations between $(Z(\mathbf{s}_i), Z(\mathbf{s}_j))$ imposed by the model if: (a) $w_{ij} = 1$; and (b) $w_{ij} = 0$? **[3 MARKS]**

iii. Describe two limitations of the *intrinsic CAR* model. **[2 MARKS]**

iv. Briefly describe a model that overcomes the limitations outlined in iii. above and write down its full conditional distribution $Z(\mathbf{s}_i)|\mathbf{Z}(-\mathbf{s}_i)$. **[3 MARKS]**

3. (a) Consider a spatial point process $Z = \{Z(A)|A \subset D\}$ defined on a spatial domain D.

i. Write down the modelling assumptions for an inhomogeneous Poisson process (IPP) with first order intensity function $\lambda(\mathbf{s})$. **[2 MARKS]**

ii. Write down the general form for a log-linear parametric model for $\lambda(\mathbf{s})$ ensuring you define all the quantities you specify. In practice, name one drawback of fitting such as model. **[3 MARKS]**

(b) Consider a spatial point process $Z = \{Z(A)|A \subset D\}$, where the domain D is a unit square whose four corners have coordinates $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. The first order intensity function for this process is given by $\lambda(\mathbf{s}) = s_1 + s_2 - s_1 s_2$, where the location $\mathbf{s} = (s_1, s_2)$. Thus, across the domain D both $s_1, s_2 \in [0,1]$.

i. Compute the first order intensity function at (a) $\mathbf{s} = (0, 0.5)$, (b) $\mathbf{s} = (0.5, 0)$, and (c) $\mathbf{s} = (0.5, 0.5)$. Hence or otherwise briefly describe the spatial pattern in the first order intensity function across the domain D. **[4 MARKS]**

ii. Consider a region $A \subset D$, write down the formula for the expected number of points that occurred in A, $\mathbb{E}[Z(A)]$. **[2 MARKS]**

iii. Let A denote the rectangle in the domain D defined by all points $\mathbf{s} = (s_1, s_2)$ such that $s_1 \in [0,1]$ and $s_2 \in [0.4, 0.5]$. That is, the four corners of A are $(0, 0.4)$, $(0, 0.5)$, $(1, 0.4)$, $(1, 0.5)$. Compute the expected number of points that occurred in A. **[3 MARKS]**

iv. Now suppose that the second order intensity function for this process is $\lambda_2(\mathbf{s}, \mathbf{t}) = 1$ for all $(\mathbf{s}, \mathbf{t}) \in D$. Compute the pair correlation function $\rho(\mathbf{s}, \mathbf{t})$ for points $(\mathbf{s}, \mathbf{t}) \in D$. **[3 MARKS]**

v. Are there any points $(\mathbf{s}, \mathbf{t}) \in D$ at which the pair correlation function computed in iv. is not a real number? If so what aspect of either the first or second order intensity functions is causing this problem. **[3 MARKS]**

CONTINUED OVERLEAF/

4. (a) A researcher is interested in estimating what factors impact the risk of respiratory disease in Scotland, and has collected a population-level summary data set at the areal unit level for the n intermediate zones that make up Scotland. Specifically, for each intermediate zone they collected the following data:

- The number of deaths due to respiratory disease from the population living in that areal unit.
- The expected number of deaths from respiratory disease based on the population size and its age-sex demographics.
- The average air pollution levels.
- Five different measures of poverty, including measures of average income, unemployment rate, average house price, average education level, and crime rates.

The researcher is unsure of how to undertake a spatial analysis of these data. Write down a step-by-step analysis plan of the steps the researcher should take when modelling these data. In doing so point out any pitfalls or problems that they might encounter and discuss how they could be solved. **[14 MARKS]**

(b) Consider an areal unit process $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$ with a corresponding binary $n \times n$ neighbourhood matrix $\mathbf{W}$, where $w_{ij} = 1$ if areas (i, j) are spatial neighbours (share a common border) and $w_{ij} = 0$ otherwise. Consider the conditional autoregressive (CAR) model $\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W})^{-1})$, where the precision matrix $\mathbf{Q}(\mathbf{W})$ is given by

$$\mathbf{Q}(\mathbf{W}) = [\mathrm{diag}(\mathbf{W1}) - \rho\mathbf{W}],$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones and ρ is a spatial dependence parameter with $\rho \in [0, 1)$. Derive the full conditional distribution $Z(\mathbf{s}_i)|\mathbf{Z}(-\mathbf{s}_i)$, where $\mathbf{Z}(-\mathbf{s}_i)$ denotes all the elements in $\mathbf{Z}$ except $Z(\mathbf{s}_i)$. **[6 MARKS]**

Hint
It may be helpful to remember that if $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ is partitioned into two subvectors whose joint distribution is given by

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \tau^2\mathbf{Q}^{-1} = \tau^2 \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}^{-1} \right).$$

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

$$\mathbf{Z}_1|\mathbf{Z}_2 \sim \mathrm{N}\left(\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\mathbf{Z}_2 - \boldsymbol{\mu}_2), \tau^2\mathbf{Q}_{11}^{-1}\right).$$

Total: 80 MARKS

END OF QUESTION PAPER.