Räumliche Statistik

Hauptklausur am 16.02.2016

Name: ____________________

Matrikelnummer: ____________________

Studiengang (Hauptfach)**:** ____________________

Angestrebter Abschluss: ☐ Bachelor ☐ Master ☐ Andere: ____________

Prüfungsordnung: ☐ 2007 ☐ 2010 ☐ Andere: ____________

Ich bestätige, dass ich die Hinweise auf der Rückseite zur Kenntnis genommen und die Klausurangabe auf Vollständigkeit überprüft habe. Ich bin mit der Veröffentlichung meines Klausurergebnisses in Moodle in der Form < Matrikelnummer >< Note > einverstanden. (Falls nicht, den letzten Satz bitte streichen.)

Unterschrift: ____________________

Aufgabe	1	2	3	4	5	Gesamt	**Note:**
Punkte	9	7	16	11	17	60	
Davon erreicht							

Hinweise zur Klausur:

1. Überprüfen Sie zunächst, ob Ihre Klausurangabe vollständig ist. Die Klausurangabe sollte aus 9 Seiten (ohne Deckblatt) mit 5 Aufgaben bestehen.

2. Die Bearbeitungszeit beträgt 90 Minuten. In den ersten 30 Minuten und in den letzten 15 Minuten ist keine vorzeitige Abgabe vorgesehen.

3. Es können maximal 60 Punkte erreicht werden. Ein nachvollziehbarer Lösungsweg ist Voraussetzung zum Erlangen der vollen Punktzahl.

4. Es sind folgende Hilfsmittel zugelassen:
   - Ein DIN-A4-Blatt mit handschriftlichen Notizen auf Vorder- und Rückseite
   - Nichtprogrammierbarer Taschenrechner
   - Bei Bedarf ein Wörterbuch

5. Verwenden Sie für Ihre Lösungen ausschließlich die Ihnen zur Verfügung gestellten karierten Kanzleibögen und kennzeichnen Sie jeden Kanzleibogen (einmal) mit Ihrem Namen und Ihrer Matrikelnummer.

6. Es darf nicht mit Bleistift geschrieben werden.

7. Legen Sie einen Lichtbildausweis und einen aktuell gültigen Studienausweis bereit.

8. Bei Unterschleif erfolgt eine Meldung an das Prüfungsamt. Sie sind verpflichtet, durch Ihr Verhalten jegliche Missverständnisse diesbezüglich auszuschließen.

9. Verlassen Sie den Prüfungsraum erst, nachdem Sie der Aufsicht die Klausur persönlich übergeben haben. Für den Eingang der kompletten Klausur bei der Aufsicht sind Sie selbst verantwortlich.

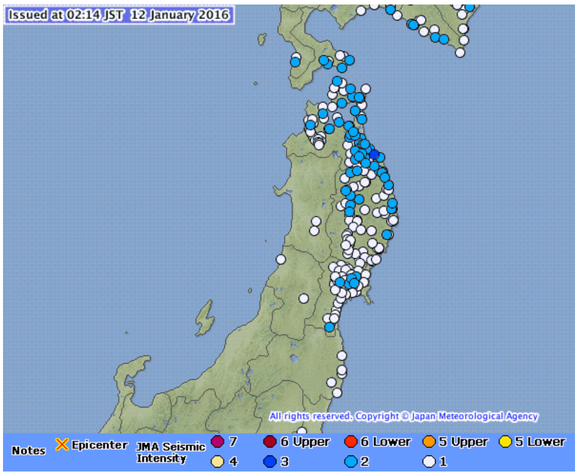
10. Bitte verlassen Sie nach der Klausur den Hörsaaltrakt zügig und leise, damit Sie die Teilnehmer anderer Klausuren nicht stören.

Viel Erfolg!

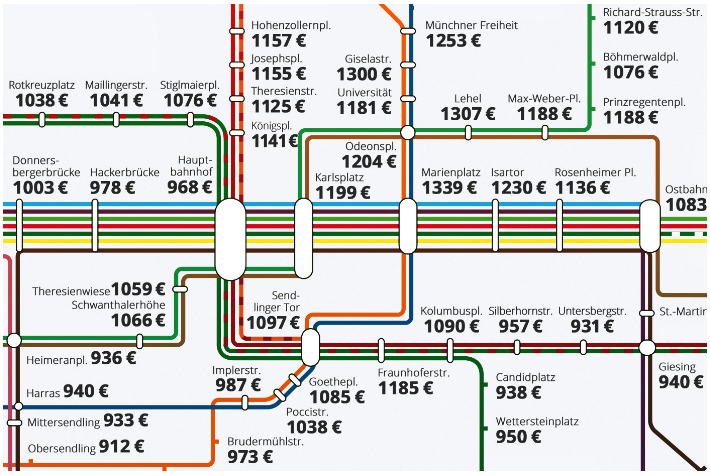
Aufgabe 1 (9 Punkte)

Entscheiden Sie für jeden der folgenden Datensätze, um welche Art von räumlichen Daten es sich jeweils handelt und begründen Sie Ihre Entscheidung.

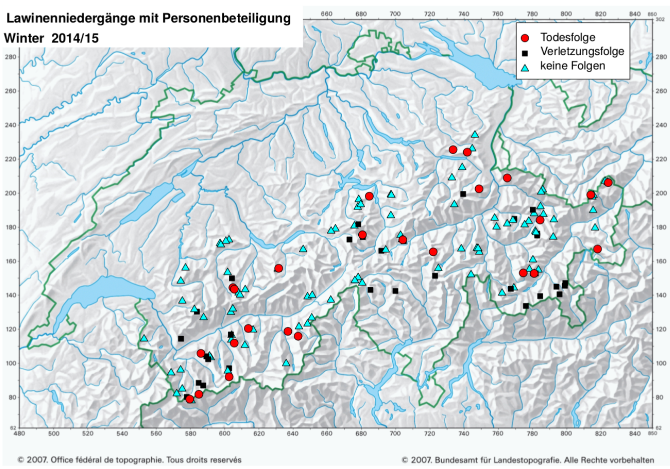
1. Stärke eines Erdbebens am 12. Januar 2016 in Japan gemäß der *JMA Seismic Intensity* Erdbebenskala an ausgewählten Messstationen:



2. Durchschnittliche Kaltmiete für eine 30 Jahre alte 2-Zimmer Wohnung mit $70m^2$ in München in Abhängigkeit von der nächstgelegenen U-/S-Bahn Station (Ausschnitt) . Nehmen Sie hier an, dass sich jede Wohnung eindeutig einer U-/S-Bahn Station zuordnen lässt:



3. Lawinenniedergänge in der Schweiz im Winter 2014/15 mit Personenbeteiligung, aufgeschlüsselt nach Art der Folge:



Quellen: Japan Meteorological Agency, Immobilienscout24, Institut für Schnee- und Lawinenforschung SLF

Lösung:

1. Erdbeben: Festgelegte Koordinaten (Messstationen) mit Daten (Stärke des Bebens am 12. Januar 2016, kategoriale Information) an jeder Lokation → Geostatistische Daten.

2. Mietpreise: Aggregierte Daten (Kaltmiete) über die U-/S-Bahnstationen in München. Stationen, die von derselben Linie bedient werden und direkt aufeinander folgen, sind benachbart und bilden insgesamt ein irreguläres Gitter → Irreguläre Gitterdaten.

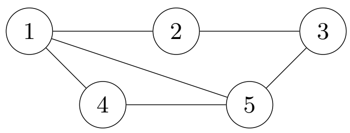
3. Lawinen: Zufällige, nicht festgelegte Koordinaten der Lawinenabgänge mit jeweils zusätzlicher Information (Todesfolge / Verletzungsfolge / keine Folge) an den Lokationen → Markierter Punktprozess.

Aufgabe 2 (7 Punkte)
Bezeichne $X = (X_1, \ldots, X_5)$ ein intrinsisches Gauß-Markov Zufallsfeld mit Mittelwert 0 und Präzisionsmatrix

$$\Sigma^{-1} = \frac{1}{\tau^2} \begin{pmatrix} 3 & -1 & 0 & -1 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & 0 & -1 \\ -1 & 0 & 0 & 2 & -1 \\ -1 & 0 & -1 & -1 & 3 \end{pmatrix}.$$

(a) Geben Sie den zu X gehörenden Nachbarschaftsgraphen an. [3]

Lösung:



(b) Prüfen Sie, ob jeweils (bedingte) Unabhängigkeit vorliegt. Begründen Sie Ihre Antwort. [4]

1. $X_3 \perp X_4 | X_{\{1,5\}}$
2. $X_1 \perp X_3 | X_2$

Lösung:

1. Bedingt unabhängig, da $C = \{1, 5\}$ die Mengen $A = \{3\}, B = \{4\}$ separiert, d.h. globale Markov-Eigenschaft ist erfüllt.
2. Nicht bedingt unabhängig, da $A = \{1\}$ und $B = \{3\}$ über $5 \notin C = \{2\}$ verbunden sind, d.h. globale Markov-Eigenschaft ist nicht erfüllt.

Aufgabe 3 (16 Punkte)

Gegeben sei ein stationärer Gauß-Prozess W in $\mathbb{R}^d$, $d \in \mathbb{N}$ mit Mittelwert $\mu = 0$, Varianz σ^2 und Korrelationsfunktion ρ. Für beliebige Lokationen $s_1, \ldots, s_n \in \mathbb{R}^d$ sei $w = (W(s_1), W(s_2), \ldots, W(s_n))$.

(a) Geben Sie die Kovarianzmatrix von w an. [2]

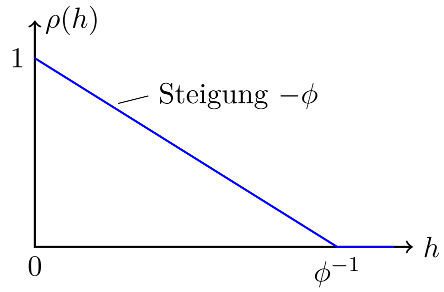
Lösung:
$\Sigma = \mathrm{Cov}(w) = \sigma^2 R$ mit $r_{ij} = \rho(\|s_i - s_j\|_2),\ i, j = 1, \ldots, n.$

(b) Für die Korrelationsfunktion ρ soll folgendes parametrische Modell angenommen werden: [5]

$$\rho(h) = \begin{cases} 1 - \phi h & |h| \leq \phi^{-1} \\ 0 & |h| > \phi^{-1} \end{cases} \qquad \phi > 0.$$

Welchen Einfluss haben große/kleine Werte von ϕ auf die räumliche Struktur von W?
Hinweis: Für die Beantwortung dieser Frage kann eine Skizze hilfreich sein.

Lösung:
Skizze:



- ϕ groß, z.B. $\phi = 10$: ρ fällt schnell ab, d.h. Korrelation zwischen $W(s)$ und $W(t)$ ist schon für kleine Abstände $\|s - t\|_2$ eher niedrig (für $\|s - t\|_2 > \phi^{-1} = 0.1$ sind $W(s), W(t)$ sogar unabhängig), d.h. W hat eine kleinräumige, 'rauhe' Struktur
- ϕ klein, z.B. $\phi = 0.1$: ρ fällt nur langsam ab, d.h. die Korrelation zwischen $W(s)$ und $W(t)$ ist auch für größere Abstände $\|s - t\|_2$ noch relativ hoch, d.h. W hat eine großflächige, 'glatte' Struktur.

Betrachten Sie nun das gewöhnliche Kriging-Modell

$$Y(s) = \mu + W(s) + \varepsilon(s).$$

Dabei bezeichnet $\mu \in \mathbb{R}$ den (konstanten) Mittelwert, W ist der oben genannte Gauß-Prozess und $\varepsilon(s)$ sei ein von W unabhängiger Prozess mit $\varepsilon(s) \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$.

(c) Zeigen Sie, dass für das (Semi-)Variogramm von Y gilt [6]

$$\gamma(h) = \sigma^2(1 - \rho(h)) + \tau^2, \quad h \geq 0.$$

Lösung:

Seien $s, t \in S$ so gewählt, dass $\|s-t\|_2 = h \geq 0$.

$$\begin{aligned}
\gamma(h) &= \frac{1}{2} \operatorname{Var}\left(Y(s) - Y(s+t)\right) \\
&= \frac{1}{2} \operatorname{Var}\left(W(s) + \varepsilon(s) - W(s+t) - \varepsilon(s+t)\right) \\
&\overset{\text{unabh.}}{=} \frac{1}{2} \operatorname{Var}\left(W(s) - W(s+t)\right) + \frac{1}{2} \operatorname{Var}\left(\varepsilon(s) - \varepsilon(s+t)\right) \\
&\overset{\varepsilon \text{ iid}}{=} \frac{1}{2} \left[\operatorname{Var}\left(W(s)\right) + \operatorname{Var}\left(W(s+t)\right) - 2 \operatorname{Cov}\left(W(s), W(s+t)\right)\right] \\
&\quad + \frac{1}{2} \operatorname{Var}\left(\varepsilon(s)\right) + \frac{1}{2} \operatorname{Var}\left(\varepsilon(s+t)\right) \\
&\overset{\text{Vor.}}{=} \frac{1}{2} \left[\sigma^2 + \sigma^2 - 2\sigma^2 \rho(h)\right] + \frac{1}{2}\tau^2 + \frac{1}{2}\tau^2 \\
&= \sigma^2 - \sigma^2 \rho(h) + \tau^2 \\
&= \sigma^2 (1 - \rho(h)) + \tau^2.
\end{aligned}$$

(d) Bestimmen Sie den *Sill* von Y. [3]

Lösung:
$\lim_{h \to \infty} \gamma(h) = \lim_{h \to \infty} \sigma^2(1 - \rho(h)) + \tau^2 = \sigma^2 + \tau^2.$

Aufgabe 4 (11 Punkte)
Zur Untersuchung der Malaria-Krankheit wurde in Afrika eine groß angelegte Studie mit $N = 2.500$ Teilnehmern durchgeführt. Für jeden Teilnehmer $i = 1, \dots, N$ liegen folgende Informationen vor:

Variable	Beschreibung
b_i	Mit Malaria assoziierter Blutwert (annähernd normalverteilt)
x_i	Kovariablen mit linearem Effekt (Alter, Geschlecht, Anzahl evtl. vorausgegangener Malaria-Infektionen)
r_i	Region, in der der Teilnehmer wohnt (codiert von $1, \dots, 10$)

(a) Geben Sie ein geeignetes statistisches Modell für den Blutwert b_i als Zielgröße an und erklären Sie stichwortartig alle in der Modellgleichung enthaltenen Terme. Gehen Sie dabei besonders auf die Modellierung der räumlichen Information ein! [5]

Lösung:

$$b_i = \mu + x_i^\top \beta + \gamma(r_i) + \varepsilon_i, \quad i = 1, \dots, N$$

- b_i: Zielgröße (Blutwert, annähernd normalverteilt)
- μ: Intercept
- $x_i^\top \beta$: Lineare Effekte β der Kovariablen x_i
- $\gamma(r_i)$: Regionsspezifischer räumlicher Effekt. Modellierung von $\gamma = (\gamma(1), \dots, \gamma(10))$ über Gauß-Markov-Zufallsfeld
- ε_i Fehlerterm

Nach einigen Jahren wird eine Nachfolgerstudie durchgeführt. Dank verbesserter technischer Ausstattung können nun statt der Region r_i die Koordinaten s_i des Wohnorts des Patienten bis auf wenige Meter genau erfasst werden. Alle weiteren Variablen (b_i, x_i) werden wie in der Vorgängerstudie erhoben.

(b) Passen Sie ihr Modell aus (a) an die neue Datensituation an. Erklären Sie stichwortartig alle neu hinzugekommenen Modellterme und geben Sie eine geeignete Form der Modellierung für s_i an. [3]

Lösung:

$$b_i = \mu + x_i^\top \beta + f(s_i) + \varepsilon_i, \quad i = 1, \dots, N$$

Neuer Modellterm: $f(s_i)$, glatter räumlicher Effekt. Modellierung z.B. über

- (stationäres) Gaußfeld $f \to$ Kriging
- Basisfunktionenansatz $f(s) = \sum_{k=1}^{K} \alpha_k b_k(s)$ mit gegebenen Basisfunktionen b_k.

(c) Wie ändert sich das Modell, wenn statt b_i die binäre Zielgröße [3]

$$m_i = \begin{cases} 1 & \text{Patient } i \text{ ist an Malaria erkrankt,} \\ 0 & \text{sonst} \end{cases}$$

betrachtet werden soll? Erklären Sie auch hier alle neu hinzugekommenen Modellkomponenten (Stichworte genügen).

Lösung:

$$\mu_i = \mathbb{E}\left(m_i | x_i, s_i\right) = h(\eta_i)$$
$$\eta_i = \mu + x_i^\top \beta + f(s_i)$$

Neue Modellkomponenten:

- h: Linkfunktion, z.B. Logit-Link
- m_i: Bedingter Erwartungswert von m_i *Formel oben genügt auch...*
- η_i: Linearer Prädiktor

Aufgabe 5 (17 Punkte)

Sei $\{N(B),\ B \in \mathcal{B}\}$ ein homogener Poisson-Prozess in $\mathbb{R}^2$ mit Intensität $\lambda > 0$. Seien weiterhin $B_1, B_2 \in \mathcal{B}$ gegeben mit $B_1 \cap B_2 = C \neq \varnothing$.

(a) Begründen Sie, warum $N(B_i) = N(B_i \setminus C) + N(C),\ i = 1, 2.$ [2]

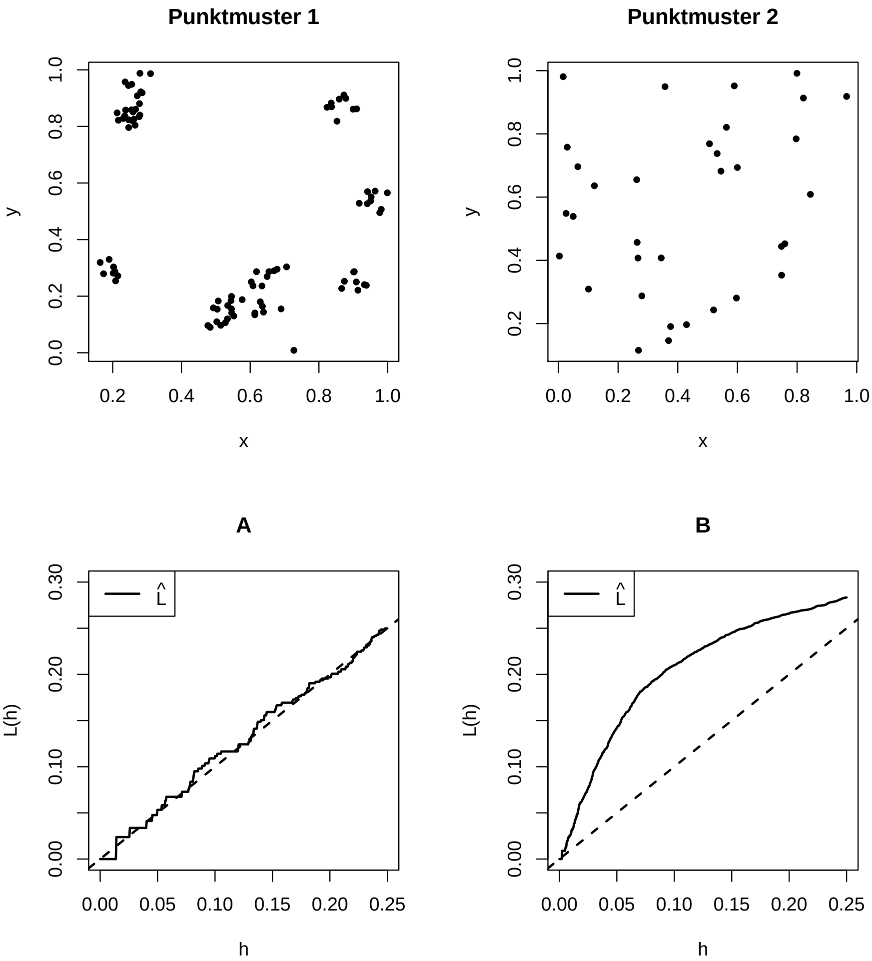
> **Lösung:**
> Verschiedene Möglichkeiten, z.B.
> - N ist (zufälliges) Zählmaß, damit gelten insbesondere alle Maßeigenschaften
> - $p(s)$ ausdünnen mit $\mathbb{1}_C$ bzw. $\mathbb{1}_{B_i \setminus C}$
> - $N(B)$ beschreibt die Anzahl der Ereignisse von N in B. Da $B_i = (B_i \setminus C)\dot{\cup}C$, setzt sich die Anzahl der Ereignisse in B_i additiv zusammen aus der Anzahl der Punkte in $B_i \setminus C$ und C.

(b) Bestimmen Sie $\text{Cov}(N(B_1), N(B_2))$. [6]

> **Lösung:**
>
> $$\begin{aligned}\text{Cov}(N(B_1), N(B_2)) &= \text{Cov}\left(N(B_1 \setminus C) + N(C), N(B_2 \setminus C) + N(C)\right) \\ &= \text{Cov}\,(\underbrace{N(B_1 \setminus C), N(B_2 \setminus C)}_{B_1\setminus C, B_2 \setminus C \text{ disjunkt}}) + \text{Cov}\,(\underbrace{N(B_1 \setminus C), N(C)}_{B_1 \setminus C, C \text{ disjunkt}}) + \\ &\quad + \text{Cov}\,(\underbrace{N(C), N(B_2 \setminus C)}_{C, B_2 \setminus C \text{ disjunkt}}) + \text{Cov}\left(N(C), N(C)\right) \\ &\overset{*}{=} \text{Var}(N(C)) \overset{**}{=} (\lambda\nu(C))^2 = \lambda^2\nu(C)^2.\end{aligned}$$
>
> *: Für A_1, A_2 disjunkt sind $N(A_1), N(A_2)$ unabhängig, d.h. $\text{Cov}(N(A_1), N(A_2)) = 0$.
> **: $N(C) \sim \text{Poi}(\lambda\nu(C))$ mit $\nu(\cdot)$ Lebesguemaß auf $\mathbb{R}^2$.

(c) Die folgende Abbildung zeigt je eine Realisierung von N und einem anderen Prozess M auf dem Einheitsquadrat. Zusätzlich wurde für jedes Punktmuster die geschätzte L-Funktion berechnet (vgl. Abb. A und B; zur besseren Orientierung ist die Winkelhalbierende als gestrichelte Linie mit angegeben). [7]

1. Ordnen Sie die L-Funktionen den Punktmustern zu und begründen Sie Ihre Wahl.
2. Bei welchem Punktmuster handelt es sich wohl um die Realisierung von N?



Lösung:

Zuordnung:

- Punktmuster 1, L-Funktion B:
  Das Punktmuster weist deutliche Clusterbildung auf. Die L-Funktion ist für kleine Werte von h deutlich größer als die Winkelhalbierende, die der theoretischen L-Funktion des homogenen Poisson-Prozesses entspricht. Dies spricht für Clusterbildung.
- Punktmuster 2, L-Funktion A:
  Das Punktmuster lässt weder Clusterbildung noch Regularität erkennen. Dies deutet auf einen homogenen Poisson-Prozess hin. Die geschätzte L-Funktion stimmt nahezu mit der Winkelhalbierenden überein, was ebenfalls weder Clusterbildung noch Regularität indiziert.

Die Realisierung von N entspricht also Punktmuster 2.

(d) Kann man in der vorherigen Aufgabe allein aufgrund der geschätzten L-Funktion auf das Vorliegen eines homogenen Poisson-Prozesses schließen? (Kurze Begründung genügt.) [2]

Lösung:

Nein, Gegenbeispiel z.B. durch Zellenprozess von Baddeley und Silverman (vgl. Übung).