Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
IDP Institut für Datenanalyse
und Prozessdesign

2020-6

CAS Datenanalyse

Prüfung Modul B Teil 2

Name:	Punkte:	Note:

Zeit

150 Minunten

Bewertung

Geben sie bei quantitativen Fragen immer alle benutzten Formeln an, bevor sie die Zahlen einsetzen und Schlussresultate berechnen. Erfragte Begründungen müssen in ganzen Sätzen verständlich ausformuliert sein. Fehlt eine Begründung, bzw. ein klar ersichtlicher Lösungsweg, so wird auch bei korrekten Zahlenresultaten höchstens die Hälfte des Punktemaximums vergeben.

Erlaubte Hilfsmittel:

open book, d.h. beliebige schriftliche Hilfsmittel sind erlaubt. Taschenrechner sind ebenfalls zugelassen. Ein Laptop darf zur Bearbeitung der Aufgaben eingesetzt werden. Auf dem Computer darf R-Studio und ein PDF-Reader betrieben werden. Bereits bestehende R-Files dürfen eingesetzt werden.

Viel Erfolg

Busfahrgäste in den USA (24 Punkte)

Laden Sie die Daten der folgenden zwei Files `riders.Rdata` und `testData.Rdata` in den R-Workspace.

```
load(".../riders.Rdata")
load(".../testData.Rdata")
```

Im File `riders.Rdata` findet sich ein Data Frame mit dem Namen `riders`. Dieses enthält die Variable `avgnumber`. Es ist die mittlere Anzahl Busfahrgäste pro Monat in Iowa City, USA. Es liegen Daten von September 1971 bis Dezember 1982 vor.

Im File `testData.Rdata` ist ein `ts`-Objekt mit dem Namen `testData` enthalten. Diese Daten benötigen Sie nur, falls Sie in der Teilaufgabe e) nicht selber einen Test-Datensatz erzeugen können. `testData` enthält die logarithmierte Anzahl Busfahrgäste pro Monat in Iowa City für das Jahr 1983.

a) [1 Punkt] Definieren Sie die Daten in `riders` gemäss den obigen Angaben in R sinnvoll und korrekt als Zeitreihe. Schreiben Sie den verwendeten R-Befehl auf das Lösungsblatt.

b) [1.5 Punkte] Betrachten Sie die Zeitreihe (`plot(...)`). Formulieren Sie eine Aussage, beginnend mit "Die Reihe ist nicht stationär, weil..". Drei Eigenschaften sind erwähnenswert.

Arbeiten Sie ab jetzt mit der log-transformierten Zeitreihe (`log(...)`)!

c) [2 Punkte] Führen Sie eine STL-Zerlegung der log-transformierten Reihe durch. Für die Saisonkomponenten probieren Sie die Argumente `s.window=3`, `s.window=15` und `s.window="periodic"`. Welche dieser drei Einstellungen bevorzugen Sie und wehalb?

d) Nehmen Sie nun den stationären Restterm aus der STL-Zerlegung mit `s.window=3` (ohne Rückschluss für Teilaufgabe c)). Beantworten Sie mit ACF-und PACF-Plot:

   i) [1 Punkt] ist das Anpassen eines AR(p)-Modells angebracht?

   ii) [1.5 Punkte] schlagen Sie 3 Modellordnungen vor, die man ausprobieren sollte.

   iii) [1 Punkt] welche Modellordnung wird aufgrund des AIC-Kriterium gewählt?

   iv) [2 Punkte] Kann mit dem AR-Modell aus iii) die zeitliche Korrelation im Restterm vollständig beschrieben werden. Begründen Sie mit einem Satz ihre Antwort.

e) [1 Punkt] Teilen Sie die logarithmierte Zeitreihe in ein Training- (Daten bis Ende 1981) und ein Test-Datensatz (Daten ab 1982). Geben Sie den dazu verwendeten R-Code an.

f) [2 Punkte] Passen Sie mit der `ets(...)` Funktion einen exponentiellen Glätter an den in der Teilaufgabe e) generierten Training-Datensatz an. Geben Sie die geschätzten Glättungsparameter α, β und γ an. Worin unterscheidet sich das gefundenen Modell gegenüber dem klassischen Holt-Winters-Verfahren?
Falls Sie in e) keinen Training-Datensatz erzeugen konnten, verwenden Sie dazu die ganze logarithmierte Zeitreihe.

g) [1 Punkt] Berechnen Sie, mit dem in f) gefundenen Modell eine Prognose der logarithmierten, mittlere Anzahl Fahrgäste aller Monate im Jahr 1982. Geben Sie den verwendeten R-Code dazu an.
Falls Sie in e) keinen Trainigs-Datensatz erzeugen konnten, berechnen Sie mit dem in f) gefundenen Modell eine Prognose der logarithmierten, mittleren Anzahl Fahrgäste aller Monate im Jahr 1983.

h) [2 Punkte] Vergleichen Sie die Vorhersagewerte mit dem in Teilaufgabe e) generierten Test-Datensatz. Wie gross ist der out-of-sample Root Mean Square Error der Vorhersage? Geben Sie den verwendeten R-Code dazu an.
Falls Sie in e) keinen Test-Datensatz erzeugen konnten, verwenden Sie das ts-Objekt `testData` um die Vorhersagewerte zu vergleichen.

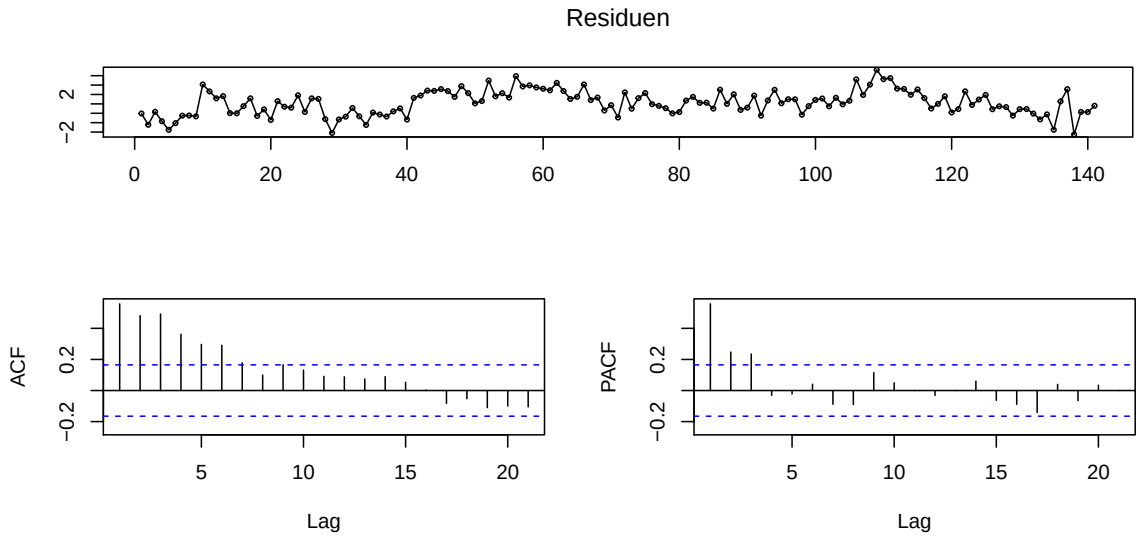
i) [1 Punkt] Was für ein Regressionsmodell würden Sie verwenden um die logarithmierte, mittlere Anzahl Fahrgäste pro Monat zu modellieren. Gehen Sie davon aus, dass Ihnen keine zusätzliche Information (Daten) zur Verfügung steht. Sie können das Modell in der R Formel-Schreibweise ($y \sim x$) angeben, mit eindeutiger Bezeichnung der Variablen.

j) [1 Punkt] Müssen Sie, wenn Sie nur an einer Prognose interessiert sind, überprüfen, ob die Residuen allenfalls noch zeitleiche Autokorrelation aufweisen?

k) [1 Punkt] Weisen die geschätzten Regressionskoeffizieten der Variablen `t` und `Monat` einen systematischen Fehler auf, falls die Residuen eine zeitliche Korrelation aufweisen?

l) [1 Punkt] Weisen die geschätzten Standardfehler der Variablen `t` und `Monat` einen systematischen Fehler auf, falls die Residuen eine zeitliche Korrelation aufweisen?

m) [3 Punkte] Wenn Sie die Korrelation der Fehler berücksichtigen möchten, müssen Sie die `gls`-Funktion in R verwenden. Angenommen die ACF und PACF der Residuen ihres Regressionsmodells in i) sehen gemäss der unteren Abbildung aus. Geben Sie den R-Code an, um ein `gls`-Modell zu schätzen. Sie müssen in dieser Aufgabe nichts berechnen. Es genügt wenn Sie den R-Code hinschreiben.



n) [1 Punkt] Angenommen die ACF und PACF der Residuen ihres Regressionsmodells sehen gemäss der oberen Abbildung aus. Werden die Standardfehler der Regressionskoeffizienten über- oder unterschätzt?