Einführung in Umweltanalysen mit R

Unterlagen für Lehrpersonen

K. Rosin

Institution:	ETH Zürich, Institut für Verhaltenswissenschaften
Unterrichtsform:	Leitprogramm
Thema:	Einführung in Umweltanalysen mit R
Fachgebiet:	Umweltwissenschaften
Unterrichtsstufe:	Hochschule, Fachhochschule, Gymnasium
Vorkenntnisse:	Grundlagen in Statistik und Umweltwissenschaften
Bearbeitungsdauer:	Sechs Lektionen (je etwa drei Stunden)
Autor:	K. Rosin, D-BAUG, rosin@ifu.baug.ethz.ch
Betreuung:	Dr. G. Cerletti
Kolektorat:	B. Bründler, D. Rosin
Fassung:	Oktober 2008

Inhaltsverzeichnis

1. Verwendung des Leitprogramms 4
1.1. Einführung 4
1.2. Unterlagen und Inhalt 5
1.3. Aufgaben der Lehrperson 7
1.4. Abschlussarbeit 8

2. Ziele des Unterrichts 10
2.1. Lernziele des gesamten Leitprogramms 10
2.2. Lernziele zu Kapitel 2 10
2.3. Lernziele zu Kapitel 3 11
2.4. Lernziele zu Kapitel 4 11
2.5. Lernziele zu Kapitel 5 12
2.6. Lernziele zu Kapitel 6 12
2.7. Lernziele zu Kapitel 7 13

3. Kommentare zu den Zusatzaufgaben 14
3.1. Zusatzaufgaben zu Kapitel 2 14
3.2. Zusatzaufgaben zu Kapitel 3 15
3.3. Zusatzaufgaben zu Kapitel 4 15
3.4. Zusatzaufgaben zu Kapitel 5 16
3.5. Zusatzaufgaben zu Kapitel 6 19
3.6. Zusatzaufgaben zu Kapitel 7 22

4. Quellen 30
4.1. Quellen für die Studierenden 30
4.2. Zusätzliche Publikationen 30
4.3. Daten für die Kapiteltests 32

5. Literaturverzeichnis 34

1. Verwendung des Leitprogramms

1.1. Einführung

Adressaten

Das vorliegende Leitprogramm richtet sich an Studierende der Fachrichtungen Umweltlehre, Hydrologie, Geographie, Forstwirtschaft und Biologie an Hochschulen oder Fachhochschulen. Die Adressaten sind Studierende im ersten oder zweiten Semester, welche später in ihrem Studium Umweltdaten analysieren werden.

Voraussetzungen

In diesem Leitprogramm wird ein minimales Wissen bezüglich Umweltwissenschaften und Statistik vorausgesetzt. Das bedeutet, dass den Studierenden Umweltsysteme wie beispielsweise der Wasserkreislauf bekannt sind. Bezüglich Statistik wird vorausgesetzt, dass die Studierenden über das Wissen einer einfach gehaltenen Statistikveranstaltung verfügen. Deshalb werden Begriffe wie Mittelwert oder Standardabweichung in diesem Leitprogramm nicht erklärt. Hingegen sind Regressionsanalysen den meisten Studierenden in unteren Semestern nicht in detaillierter Weise bekannt. Darum werden diese in diesem Leitprogramm kurz erläutert.

Ablauf des Unterrichts

Die Studierenden bearbeiten selbstständig die sechs Kapitel dieses Leitprogramms (Kapitel 1 besteht nur aus einer kurzen Einführung). Nach jedem Kapitel legen sie einen Kapiteltest ab. Pro Kapitel sind drei Stunden Bearbeitungszeit vorgesehen, inklusive 20 Minuten für den Kapiteltest. Es wird erwartet, dass die Studierenden für die Kapitel 2 (inklusive Kapitel 1), 3 und 4 weniger Zeit benötigen als für die folgenden Kapitel.

Infrastruktur

In dem vorliegenden Leitprogramm erarbeiten sich die Studierenden R-Kenntnisse anhand von praktischen Beispielen. Deshalb ist es notwendig, dass die Studierenden über einen eigenen Computer-Arbeitsplatz mit Internetzugang verfügen. Das vorliegende Leitprogramm wurde für Windows-Betriebssysteme erstellt. Deshalb wird empfohlen, dass die Studierenden das Leitprogramm nicht auf Computern mit anderen Betriebssystemen bearbeiten. Nach jedem Kapitel legen die Studierenden einen Kapiteltest ab. Die dazu verwendeten Computer sollten etwas abseits stehen, über Internetzugang verfügen, und mit einem Drucker verbunden sein. Bei einer Gruppengrösse von 30 Studierenden sollten für die Kapiteltests mindestens vier Computer zur Verfügung gestellt werden.

Es wird empfohlen, dass auf den Computern (auch auf denjenigen, die für den Kapiteltest verwendet werden) die folgende Ordnerstruktur angelegt wird:

- Ablage der Daten: *C:/R/Leitprogramm/Daten.*
- Ablage der Grafiken: *C:/R/Leitprogramm/Grafiken.*

1.2. Unterlagen und Inhalt

Allgemeine Unterlagen

- *Leitprogramm.pdf*: Diese Dokumentation wird von den Studierenden verwendet, und enthält zu jedem Kapitel Lernziele, einführende Beispiele, Übungsbeispiele, Merkpunkte, Zusatzaufgaben und eine Lernkontrolle. In dieser Datei sind auch Lösungen zu den Beispielen und der Lernkontrolle vorhanden.
- *Merkblatt.pdf*: Das Merkblatt enthält wichtige R-Merkpunkte, welche nach den Kapiteln dieses Leitprogramms geordnet sind. Die Studierenden dürfen dieses Merkblatt bei den Kapiteltests verwenden. Es enthält einen Bereich, in dem die Studierenden eigene Notizen anbringen können.

Unterlagen Studierende

Der Ordner *Unterlagen Studierende* enthält folgende Unterlagen:

- Ordner *1_Daten_Leitprogramm*: Daten, die zum Bearbeiten des Leitprogramms benützt werden.
- Ordner *2_Code_Leitprogramm*: R-Codes, die zum Bearbeiten des Leitprogramms benützt werden.

Unterlagen Lehrperson

Der Ordner *Unterlagen Lehrperson* enthält folgende Unterlagen:

- *Unterlagen_Lehrpersonen.pdf:* Dieses Dokument beinhaltet vor allem die Lernziele des Leitprogramms, Kommentare zu den Zusatzaufgaben sowie Quellenangaben.
- Ordner *1_Code_Lernkontrollen*: R-Codes zu den Lernkontrollen des Leitprogramms.
- Ordner *2_Kapiteltests*: Pro Kapitel sind vier Kapiteltests von gleichem Schwierigkeitsgrad erstellt worden. Der Ordner enthält zudem Lösungen zu den Kapiteltests (inkl. Bewertungsvorschlag).
- Ordner *3_Daten_Kapiteltests*: Daten, die für den Kapiteltest verwendet werden.
- Ordner *4_Code_Kapiteltests*: R-Codes zu den Kapiteltests.

Inhaltliche Schwerpunkte

Die folgenden Themen werden in Bezug auf Umweltanalysen als zentral eingestuft. Sie werden deshalb im vorliegenden Leitprogramm besonders stark gewichtet:

- Datenimport/-export (Kapitel 3)
- Datensätze indizieren (Kapitel 4)
- Fehlende Werte (Kapitel 4)
- Funktionen in R (Kapitel 5)
- Zeit- und Datumsformate (Kapitel 5)
- Vektorwertig programmieren (Kapitel 5)
- Deskriptive statistische Kenngrössen (Kapitel 6)
- Grafiken in R (Kapitel 6)
- Häufigkeitsverteilungen (Kapitel 7)
- Lineare Regression (Kapitel 7)

Basierend auf diesen Schwerpunkten beinhalten die Kapitel 5 bis 7 mehr Zusatzaufgaben. Die schnellen Studierenden sollen ihre Zeit bei den wichtigen Themen einsetzen können! Als Fortsetzung des R-Unterrichts nach diesem Leitprogramm werden unter anderem die folgenden Inhalte empfohlen: Statistische Tests, Varianzanalyse, weiterführende Regressionsanalysen, Geostatistik, GIS und Zeitreihenanalysen.

Unterrichtsmethoden

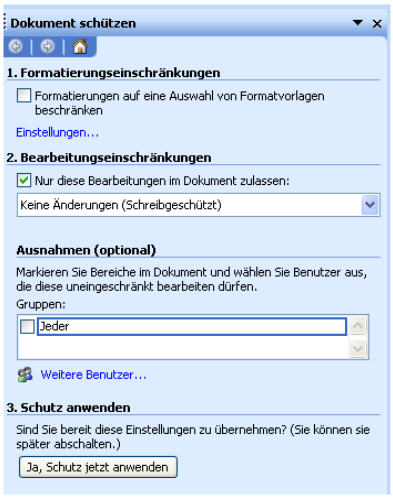
Die Studierenden erarbeiten sich in selbstständiger Arbeit Kenntnisse in R. Dabei lesen sie einführende Beispiele, holen sich Informationen aus dem Internet und lösen mit R Übungsbeispiele. Auf Gruppenarbeiten wurde wegen den wahrscheinlich unterschiedlichen Bearbeitungsgeschwindigkeiten der Studierenden verzichtet. Zudem wird nur in den Zusatzaufgaben auf Bücher verwiesen, weil nicht erwartet werden kann, dass von diesen Büchern ein „Klassensatz“ zur Verfügung gestellt wird.

1.3. Aufgaben der Lehrperson

Vorbereitung

Es wird empfohlen, dass die Lehrperson vor dem Unterricht unter anderem Folgendes vorbereitet:

- Auf den Computern sollten die neusten Versionen von R und TINN-R installiert sein. Falls aktuellere Versionen als R 2.9.2 verwenden werden, sollte überprüft werden, ob sich dadurch Änderungen im Leitprogramm ergeben. Kritische Punkte sind die Zeitzone im *POSIXct*-Format, oder die Argumente der *bmp*-Funktion.
- Die Lehrperson sollte die Excel-Einstellungen für Trennungs- und Dezimalzeichen kennen. Diese Einstellungen oder die verwendeten Daten müssen allenfalls angepasst werden.
- In dem vorliegenden Leitprogramm werden die Englischen Tastaturbezeichnungen *shift*, *ctrl* und *enter* verwendet. Diese müssen allenfalls im Leitprogramm durch suchen/ersetzen geändert werden.
- Im Leitprogramm sollte der nächstgelegene *CRAN-Mirror* verwendet werden.
- Die Lehrperson sollte auf den Computern eine sinnvolle Ordnerstruktur anlegen (siehe Kapitel 1.1). Es wird empfohlen, für die Kapiteltest- die gleiche Struktur zu verwenden.
- Die Lehrperson druckt die Leitprogramme und Merkblätter aus. Digitale Versionen von Leitprogramm und Merkblatt sollen ebenfalls auf den Computern zugänglich sein (auf den Kapiteltest-Computern nur das Merkblatt). Begründung: Die Studierenden sollen Befehle des Leitprogramms kopieren können.
- Die Kapiteltests werden aktiviert, indem die Dokumente mit einem Passwort geschützt werden (*Extras/Dokument schützen*). Die untenstehende Grafik zeigt die zu wählenden Einstellungen. Hinweis: Der Entwurfsmodus muss dabei ausgeschaltet sein.



Während dem Unterricht

Die Lehrperson verfolgt den Unterricht im Hintergrund, und steht den Studierenden im Sinn eines Tutors für allfällige Fragen und für den Kapiteltest zur Verfügung. Zudem wird empfohlen, dass die Lehrperson die zeitliche „Marschtabelle“ der Studierenden verfolgt, und in Notfällen sinnvolle Massnahmen ergreift. Während den Kapiteltests ist Internetzugang notwendig, beispielsweise um Pakete zu installieren oder die Online-Hilfe zu benützen. Die Lehrperson sollte aber darauf achten, dass die Studierenden während dem Kapiteltest keine Kommunikationsprogramme benützen.

Nachbereitung

Die Studierenden sollen ein Feedback zum Leitprogramm abgeben. Damit können das Zeitmanagement optimiert, und allfällige Unklarheiten beseitigt werden.

1.4. Abschlussarbeit

Es wird vorgeschlagen, dass die Studierenden am Ende des Leitprogramms nicht eine alles umfassende Prüfung absolvieren, sondern eine kleine Abschlussarbeit schreiben. Damit die Lehrperson frei über eine allfällige Abschlussarbeit entscheiden kann, wird diese im Leitprogramm nicht erwähnt.

Grundidee

Im vorliegenden Leitprogramm werden die Aktivitäten der Studierenden vorgegeben, damit sie sich grundlegende R-Kenntnisse aneignen. Es ist aber ebenfalls wichtig, dass Studierende die Inhalte verschiedener Kapitel dieses Leitprogramms in kombinierter Weise anwenden. Das kann in einer kleinen Abschlussarbeit mit relativ offener Zielstellung besser umgesetzt werden als mit einer grossen Prüfung.

Datenquellen

Den Studierenden wird entweder ein bestimmter Datensatz zur Verfügung gestellt, oder sie suchen selbst geeignete Daten. Beispielsweise können die folgenden Datenquellen verwendet werden:

- Institute for Atmospheric and Climate Science (IAC, 2008).
- Internationale Kommission zum Schutz des Rheins (IKSR, 2008).
- Bundesamt für Umwelt (BAFU, 2008).
- Global Runoff Data Centre (GRDC, 2008).
- United States Geological Survey (USGS, 2008).
- Environment Canada (2006).

Aufgabenstellung

Die Aufgabenstellung kann von der Lehrperson festgelegt werden. Es wird aber empfohlen, inhaltlich offene Fragestellungen zu wählen. So sollen die Studierenden eine Hypothese wählen (beispielsweise: Abflusskoeffizienten sind abhängig von der Regenintensität), und diese anhand der Daten testen und diskutieren.

Form

Die Länge der kleinen Abschlussarbeit beträgt maximal fünf Seiten. Sie soll wie eine wissenschaftliche Arbeit aufgebaut sein (Einführung, Zielsetzung, Methoden, Resultate, Diskussion, Literaturverzeichnis). Zudem soll sie mindestens eine Grafik einer Zeitreihe sowie eine Lineare Regressionsanalyse enthalten.

2. Ziele des Unterrichts

In den untenstehenden Tabellen werden die Lernziele des gesamten Leitprogramms sowie der einzelnen Kapitel (ohne Kapitel 1, Arbeitsanleitung) dargelegt. Die operationalisierten Lernziele werden nur für die entsprechenden Kapitel, nicht aber für das gesamte Leitprogramm diskutiert.

2.1. Lernziele des gesamten Leitprogramms

Leitidee

Analysen von Umweltdaten weisen oft ähnliche Strukturen auf. Beispielsweise werden Daten beschrieben, oder Regressionsmodelle an die Daten angepasst. Spezifisch auf Umweltdaten abgestimmte Informatikwerkzeuge erleichtern die Analysen beträchtlich, und sind relativ einfach zu erlernen. Bei Studierenden des ersten Semesters bestehen bezüglich Computerkenntnissen beträchtliche Unterschiede. Die Möglichkeiten der Lehrpersonen individuelle Defizite einzelner Studierender zu beseitigen sind gering. Das vorliegende Leitprogramm soll den Studierenden ein Grundwissen des Programms R vermitteln, das sich bestens eignet, um umweltrelevante Probleme zu analysieren.

Dispositionsziele

Die Studierenden sind in der Lage, Umweltdaten mit dem Programm R zu ordnen, aufzubereiten, zu analysieren sowie graphisch darzustellen.

2.2. Lernziele zu Kapitel 2

Leitidee

Die Studierenden machen sich mit den beiden Programmen R und TINN-R bekannt. Zudem ist es wichtig, dass sie von Beginn an wissen, dass Informationen und Hilfestellungen nicht bloss von diesem Leitprogramm oder der Lehrperson zu erhalten sind.

Dispositionsziele

Die Lernenden sollen Grundfunktionen von R und TINN-R bedienen können, und wissen, wo sie bei Problemen Hilfe erhalten.

Operationalisierte Lernziele

Die Studierenden

- können R und TINN-R korrekt aufstarten und beenden (K2).
- beherrschen das Laden und Installieren von Zusatzpaketen (K3).
- können drei Möglichkeiten aufzählen, wo Sie Informationen finden, um sich bei Problemen selbst zu helfen (K1).

2.3. Lernziele zu Kapitel 3

Leitidee

In R werden Daten in verschiedenen Formaten und Strukturen verwendet. Für die Lernenden ist es wichtig, die Möglichkeiten und Grenzen der einzelnen Datentypen und -strukturen zu kennen, damit sie später bei eigenen Analysen sinnvolle Datenformate verwenden.

Dispositionsziele

Die Studierenden sollen abschätzen können, welche Datentypen und -strukturen für welche Analysen geeignet sind. Sie sollen Ihre Daten in das Format der entsprechenden Objekte bringen können.

Operationalisierte Lernziele

Die Studierenden

- können anhand von Beispielen drei grundlegende Datentypen sowie vier Datenstrukturen erklären (K2).
- können in eigenen Worten beschreiben, was passiert, wenn arithmetische Operatoren auf Matrizen angewendet werden (K4).
- sind in der Lage, Daten in R zu importieren und exportieren (K3).

2.4. Lernziele zu Kapitel 4

Leitidee

Beim Messen umweltrelevanter Prozesse fallen oft grosse Datenmengen an. Diese können mit R (im Gegensatz beispielsweise zu Excel) effizient bearbeitet werden. Deshalb sind systematische Datenanalysen ein zentraler Aspekt dieses Leitprogramms.

Dispositionsziele

Die Studierenden können nach dieser Lektion selbstständig zielgerichtete Datenanalysen durchführen.

Operationalisierte Lernziele

Die Studierenden

- können in eigenen Worten das „Indizieren von Datensätzen“ erklären (K2).
- sind in der Lage, einen beliebigen Datensatz nach bestimmten Argumenten effizient zu analysieren (K3).
- können in Datensätzen fehlende Werte ersetzen oder eliminieren (K3).

2.5. Lernziele zu Kapitel 5

Leitidee

Umweltdaten sollen oft nicht nur analysiert werden, sondern dienen als Grundlage für Berechnungen. R stellt dazu viele vordefinierte Funktionen zur Verfügung. Die Studierenden sollen die Wichtigsten davon kennen. Zudem sollen sie sich grundlegende Programmierkenntnisse aneignen, um effiziente Analysen durchführen zu können. Dabei nehmen Zeit- und Datumsformate eine Sonderstellung ein: Die Studierenden sollen lernen, dass für die in Umweltanalysen sehr wichtigen Zeitformate spezifische Funktionen verwendet werden.

Dispositionsziele

Die Studierenden sollen nach diesem Kapitel über Grundkenntnisse bezüglich Funktionen und Programmieren in R verfügen. Das soll ihnen unter anderem ermöglichen, Daten in Zeit- und Datumsformat zu analysieren.

Operationalisierte Lernziele

Die Studierenden

- sind in der Lage, vordefinierte Funktionen verwenden, und selbst einfache Funktionen zu schreiben (K2).
- können für zwei Zeitformate in eigenen Worten erklären, wie Berechnungen durchgeführt sowie Daten importiert und exportiert werden (K2).
- können beurteilen, wann in einem R-Code Schleifen verwendet werden, und wann vektorwertig programmiert wird (K6).

2.6. Lernziele zu Kapitel 6

Leitidee

Wenn natürliche Prozesse gemessen werden, können schnell grosse Datenmengen anfallen. Es ist schwierig, daraus Erkenntnisse zu gewinnen. Erst das Beschreiben charakterischer Merkmale der Daten führt allenfalls zu einem besseren Verständnis der Umweltsysteme. Deshalb sollen die Studierenden lernen, wie Umweltdaten sinnvoll beschrieben werden können. Dabei kommt Grafiken eine zentrale Bedeutung zu, was auch in diesem Kapitel des Leitprogramms zum Ausdruck kommen soll.

Dispositionsziele

Die Studierenden sollen in der Lage sein, Daten verschiedenster Art mit geeigneten statistischen Funktionen zu beschreiben, und qualitativ hochwertige Grafiken zu erzeugen.

Operationalisierte Lernziele

Die Studierenden können

- in eigenen Worten formulieren, wie Daten mit Funktionen und Grafiken bezüglich Lage und Streuung beschrieben werden (K2).
- den Aufbau von Grafiken variieren, indem sie Mehrfachdarstellungen erzeugen oder die Grafikränder verändern (K3).
- mit Funktionen und Grafiken zwei Datenreihen miteinander vergleichen. (K3).

2.7. Lernziele zu Kapitel 7

Leitidee

Umweltanalysen werden nicht nur durchgeführt, um gemessene Daten zu beschreiben, sondern auch um Modelle zu erstellen. Modelle haben den Vorteil, dass sie auch für Prognosen verwendet werden können. Zudem helfen Modelle, Zusammenhänge zu verstehen, und komplexe Systeme auf wesentliche Beziehungen zu reduzieren. Deshalb sollen die Studierenden in diesem Kapitel zwei in Umweltanalysen oft verwendete Modelle (Häufigkeitsverteilungen und Regressionsmodelle) kennen lernen.

Dispositionsziele

Die Studierenden können für ihre Daten Häufigkeitsverteilungen und Regressionsmodelle erstellen, und anschliessend diese Modelle testen. Sie sollen zudem die Aussagekraft von Modellen abschätzen können.

Operationalisierte Lernziele

Die Studierenden können

- für eine diskrete Verteilung eine praktische Anwendung nennen, und beschreiben, wie diese in R berechnet wird (K2).
- in R stetige Verteilungen an Daten anpassen, und die Verteilungen testen (K3).
- einfache oder multiple lineare Regressionsmodelle erstellen (K3).
- die Aussagekraft von Regressionsanalysen beurteilen (K6).

3. Kommentare zu den Zusatzaufgaben

Dieses Kapitel beinhaltet Lösungen, Erklärungen und Bemerkungen zu den Zusatzaufgaben. Diese sollen der Lehrperson helfen, allfällige Fragen der Studierenden zu beantworten.

3.1. Zusatzaufgaben zu Kapitel 2

Zusatzaufgabe 1

Die folgenden Kurzbefehle können für die Studierenden nützlich sein. Dabei bedeuten die +-Zeichen, dass die Tasten gleichzeitig gedrückt werden.

Kurzbefehl	*Erklärung*
alt + r, r, enter	R starten
alt + r, s, s, s	Auswahl an R senden (s = selection)
alt + r, s, a	Alles an R senden (a = all)
alt + c	Auskommentieren
alt + n	Auskommentieren rückgängig (erstes # Zeichen entfernt)
shift+ctrl+c	In den Spaltenmodus wechseln (c = column)
shift+ctrl+n	In den normalen (nicht Spalten-)Modus wechseln (n = normal)
alt + a, b, i	Texteinzug (i = indent)
alt + e + u	Rückgängig machen (u = undo)
ctrl + f	Suchen (f = find)
ctrl +r	Ersetzen (r = replace)
F3	Weiter suchen (nach ctrl + f)
alt + s, enter, u	Text zu Grossbuchstaben umwandeln (u = upper case)
alt + s, enter, l	Text zu Kleinbuchstaben umwandeln (l = lower case)

Die folgenden bekannten Kurzbefehle können auch in TINN-R verwendet werden:

Kurzbefehl	*Erklärung*
ctrl + x	Ausschneiden
ctrl + c	Kopieren
ctrl + v	Einfügen

Zusatzaufgabe 2

Für die Studierenden könnten die folgenden Dokumentationen hilfreich sein:

- Kuhnert und Venables (2001): *An Introduction to R: Software for Statistical Modelling & Computing*. Gute Beschreibungen, hohe Qualität, anschauliche Beispiele, deckt inhaltlich eine grosse Bandbreite ab.
- Verzani (2005): *Simple R. Using R for Introductory Statistics.* Einfache Beschreibungen, etwas unordentlich dargestellt, viele R-Codes zum Erstellen von Grafiken.
- Robinson (2008). *IcebreakR.* Gute Einführungen zu hierarchischen Modellen sowie linearer und nicht linearer Regression.

Zusatzaufgabe 3

Mit repos kann der Mirror festgelegt werden: `install.packages("ads", repos = "http://stat.ethz.ch/CRAN/")`

3.2. Zusatzaufgaben zu Kapitel 3

Zusatzaufgabe 1

Die folgenden Argumente können hilfreich sein, um einen Datensatz auf spezifische Art und Weise einzulesen:

Argument	*Erklärungen/Beispiele*
header	*header = TRUE*: Die erste Zeile des Datensatzes wird als Überschrift verwendet. *header = FALSE*: Alle Zeilen werden als Daten eingelesen.
nrows	*nrows = 20* bedeutet, dass nur die ersten 20 Zeilen eingelesen werden.
skip	*skip = 10* bewirkt, dass die ersten 10 Zeilen nicht eingelesen werden.
na.strings	*na.strings = ("-9999")* bewirkt, dass alle *-9999* als *NA* in R eingelesen wird. Dieses Argument ist hilfreich, wenn in den Daten fehlende Werte nicht mit *NA* bezeichnet sind.
blank.lines.skip	Mit *blank.lines.skip = TRUE* werden leere Zeilen nicht eingelesen.

Zusatzaufgabe 2

Es kann vorkommen, dass die Spalten eines Datensatzes nicht durch ein Zeichen (wie beispielsweise ein Komma) getrennt sind. Falls pro Spalte jede Zeile die gleiche Anzahl Zeichen aufweist, kann dieser Datensatz mit der *read.fwf*-Funktion eingelesen werden. *fwf* steht für *fixed width formatted* (Deutsch: Format mit fester Breite).

Zusatzaufgabe 3

Mit der *append*-Funktion können Elemente an einer bestimmten Stelle eingefügt werden. Beispiel: *append(3:7, rep(NA, 4, after = 2)*.

Zusatzaufgabe 4

Diese Tabelle enthält Matrix-spezifische Funktionen (Kreuzprodukt, Invertierung, Diagonalelemente, Eigenvektoren, Eigenwerte, usw.).

Zusatzaufgabe 5

In diesem Kapitel wird unter anderem aufgezeigt, wie Daten mit der *scan*-Funktion eingelesen werden können (Seite 102). Diese Funktion ist sehr flexibel, aber auch schwieriger anzuwenden als *read.table()*.

3.3. Zusatzaufgaben zu Kapitel 4

Zusatzaufgabe 1

Mit der *subset*-Funktion werden in diesem Beispiel Teile des Dataframes ausgewählt. Dabei bewirkt der *%in%* Operator, dass nur Schwermetalle und Ionen zur Auswahl stehen. Mit dem *select*-Argument werden die zu verwendenden Spalten ausgewählt.

Zusatzaufgabe 2

Die *split*-Funktion bewirkt eine Aufspaltung der Daten. Als Resultat wird eine Liste erstellt. Beispiel: *split (Mess, Mess$Bezeichnung2)*.

Zusatzaufgabe 3

Der Ausschluss geschieht entweder über einen negativen Index (mit der *which*-Funktion oder direkt), oder über das Ausrufezeichen-Symbol einer logischen Abfrage.

Zusatzaufgabe 4

Mit *attach()* werden die Daten in den Suchpfad eingehängt. Beispielsweise können Spaltennamen der Dateframes danach als Objekte abgefragt werden. *detach()* hängt die Daten aus dem Suchpfad aus. Vorsicht, die Daten sind nicht komplett entfernt.

Zusatzaufgabe 5

Die *all*-Funktion gibt nur dann *TRUE* aus, wenn alle Elemente des Objekts die Bedingung erfüllen, die *any*-Funktion wenn mindestens ein Element die Bedingung erfüllt.

Zusatzaufgabe 6

Mit der *xor*-Funktion können „entweder-oder" Abfragen durchgeführt werden. Es wird *TRUE* ausgegeben, wenn entweder die eine oder die andere Bedingung erfüllt ist. Keine oder mehrere erfüllte Bedingungen ergibt die Ausgabe *FALSE*.

3.4. Zusatzaufgaben zu Kapitel 5

Zusatzaufgabe 1

Grundsätzlich funktioniert diese Funktion gleich wie die Funktion *round1* in Kapitel 5.2: Die Funktion *sign()* gibt das Vorzeichen eines Elements aus. *x+ sign(x)*0.5* bedeutet, dass von negativen Elementen 0.5 subtrahiert wird; Null bleibt unverändert, und zu positiven Elementen wird 0.5 addiert. Danach rundet die *trunc*-Funktion auf ganze Zahlen.

Der einzige Unterschied zwischen *round1* und *round2* ist der Faktor *10^deci*, dessen Wirkungsweise im folgenden Beispiel veranschaulicht wird: Gegeben sind x = 3.141592 und *deci* = 3. Der Faktor *10^deci* = 1000 bewirkt, dass nicht 3.141592 sondern die Zahl 3141.592 gerundet wird. Nach dem Runden wird die Zahl 3141 wieder durch 1000 geteilt, was 3.141 ergibt.

Zusatzaufgabe 2

```
round3 <- function(x) { 0.05 * trunc( x / 0.05 + sign(x)*0.5 )  }
round4 <- function(x) { 0.05 * round( x / 0.05 )  }
```

Hinweis: Bei sehr kleinen Zahlen ist bei der Anwendung der *trunc*-Funktion Vorsicht geboten, weil dann kleine Differenzen zwischen dem numerischen und theoretischen Resultat ein falsches Runden bewirken.

Zusatzaufgabe 3

Ligges (2007) enthält auf Seite 54 eine anschauliche Zusammenfassung von Zeichenkettenfunktionen. Die *sub*- und *gsub*-Funktionen sind zudem in Crawley (2008) auf Seite 82 erklärt. Hier ein paar einfache Beispiele zu diesen Funktionen:

```
cat("bonjour")                  #Ausgabe (Konsole oder Dateien)

letters                         #Buchstaben von a bis z
grep("[x-z]", letters)          #Index von x, y und z

Text <- "abcd abcd abcd"
gsub("b","B",Text)              #alle b's ersetzen
sub("b","B",Text)               #nur das erste b's ersetzen
```

Zusatzaufgabe 4

Diese Publikation stellt die Zeit- und Datumsklassen *Date*, *chron* und *POSIXct* auf anschauliche Weise vor. Am Schluss der Publikation (Seite 32) ist eine hilfreiche Zusammenfassung zu finden. In diesem Leitprogramm wurde die Klasse *chron* nicht verwendet, vor allem weil die Grafikmöglichkeiten mit *chron* beschränkt sind. Zudem können mit den Klassen *Date* und *POSIXct* sowohl einfache, wie auch komplexe Zeit- und Datumsformate abgedeckt werden.

Zusatzaufgabe 5

while-Schleifen werden ausgeführt, solange ein Kriterium erfüllt ist. Dieser Schleifentyp eignet sich beispielsweise für Iterationen.

Zusatzaufgabe 6

Es kann beispielsweise die folgende Funktion geschrieben werden:

```
Umkehren <- function(x){
Name1 <- strsplit(x, " ")                       #Namen aufteilen (Liste!)
Name2 <- strsplit(Name1[[1]], NULL)             #Buchstaben aufteilen
Name3 <- lapply(Name2, rev)                     #Reihenfolge tauschen
Name4 <- sapply(Name3, paste, collapse = "")    #Vereinfachen
Name5 <- tolower(Name4)                         #Kleinschreibung
paste(toupper(substring(Name5, 1, 1)),          #erster Buchstabe gross
substring(Name5, 2), sep = "", collapse = " ")}

Umkehren("Markus Weiler")                       #Anwendung
```

Der „Trick“ dieser Funktion besteht darin, dass mit *strsplit* eine Liste mit den Komponenten *Vorname* und *Nachname* erzeugt wird. Die nachfolgenden Funktionen werden (mit *lapply*) auf die beiden Komponenten separat angewendet. Mit *sapply* wird die Liste anschliessend wieder vereinfacht, indem sie auf eine Dimension reduziert wird.

Zusatzaufgabe 7

Mit der *diff*-Funktion werden Differenzen zwischen den benachbarten Elementen eines Vektors berechnet. Die Länge des Resultats ist deshalb um ein Element kürzer als der ursprüngliche Vektor. Mit der *cumsum*-Funktion werden kumulative Summen der Elemente des Vektors berechnet.

```
a <- c(1, 3, 6, 5)
diff(a)
cumsum(a)
```

Zusatzaufgabe 8

Das Resultat der *pmin*- und *pmax* Funktionen ist ein Vektor gleicher Länge wie der ursprüngliche Vektor, welcher die Extremwerte der Abfrage enthält.

```
pmin(5:1, 4)
pmax(5:1, 4)
```

3.5. Zusatzaufgaben zu Kapitel 6

Zusatzaufgabe 1

- http://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html
  Die Webseite von Paul Murrell enthält eine systematische Anordnung von Grafiken, die sich an Murrell (2006) orientiert. Man muss sich durch die verschiedenen Kapitel durchklicken, um eine bestimmte Grafik zu finden.
- http://addictedtor.free.fr/graphiques/thumbs.php
  Mit Hilfe der Miniaturansichten kann eine gesuchte Grafik schnell gefunden werden. Hingegen sind nur eher selten verwendete Grafiktypen erhältlich.
- http://bm2.genes.nig.ac.jp/RGM2/index.php?pageID=1
  Die Grafik-Auswahl dieser Webseite ist gross. Es ist aber schwierig, eine bestimmte Grafik zu finden.

Zusatzaufgabe 2

Ein quadratischer Plot-Bereich kann mit pty = "s" erzeugt werden.

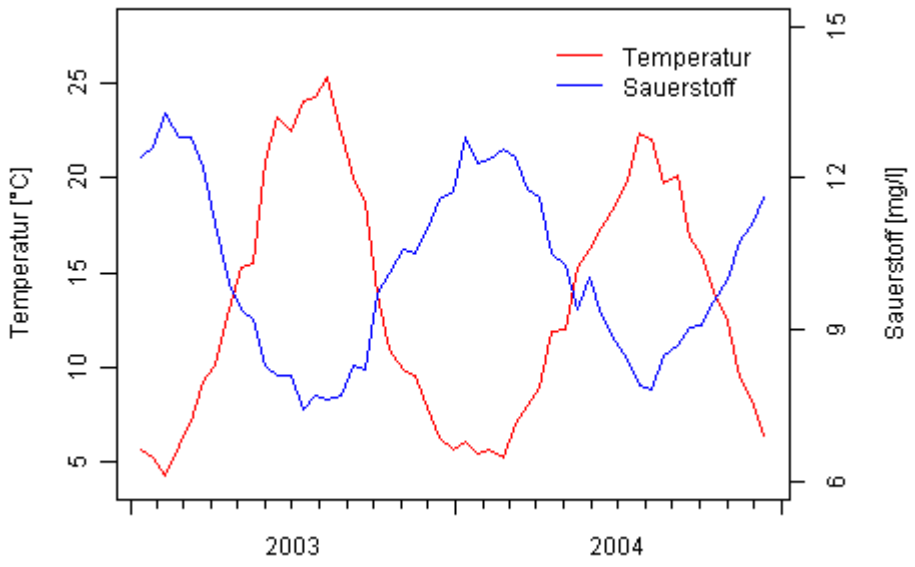
Zusatzaufgabe 3

Es kann beispielsweise der folgende Code verwendet werden:

```
#Daten importieren
    O2_PFAD <- "E:/R/3_Daten/Weil_O2.csv"
    O2 <- read.table(O2_PFAD, sep = ";", dec = ",", header = T)
    names(O2); dim(O2)
    head(O2)
#Datum vorbereiten
    W_Datum1 <- paste(O2$Datum, O2$Jahr, sep = "")
    W_Datum2 <- as.Date(W_Datum1, format="%d.%m.%Y")
    Beginn <- as.Date("1.1.2003", format="%d.%m.%Y")
    Ende <- as.Date("1.1.2005", format="%d.%m.%Y")
#Minimum und Maximum der Primär- und Sekundärachse definieren
    Max1 <- 28
    Min1 <- 4
    Max2 <- 15
    Min2 <- 6
#Skalierfunktion (entspricht einer "Skalier-Geraden" durch zwei Punkte)
    Skalieren <- function(x, Ma1, Mi1, Ma2, Mi2)
    {x * (Ma1-Mi1)/(Ma2-Mi2) + Ma1 - Ma2*(Ma1-Mi1)/(Ma2-Mi2) }
#Punkte und Achsenbeschriftung skalieren
    W_O2_S <- Skalieren (W_O2, Max1, Min1, Max2, Min2)
    W_O2_Achse <- Skalieren ( seq(6, 15, 3), Max1, Min1, Max2, Min2)
    W_O2_Label <- seq(6, 15, 3)
```

```
#Grafik
    bmp("Sekundaerachse.bmp", width=460, height=320)
        par( mar = c( 3.1, 4.1, 2.1, 4.5))
        plot(W_Datum2, O2$Temp, xaxt="n",
            ylim = c(Min1, Man1), ylab = "Temperatur [°C]",
            type = "l", col = "red")
        points(W_Datum2, W_O2_S, type = "l", col = "blue", pch = 16)
        axis(4, at=W_O2_Achse, labels = W_O2_Label)
        mtext("Sauerstoff [mg/l]", side=4, line=3)
        axis.Date(1, at=seq(Beginn, Ende, "months"), labels = F, tcl = -0.3)
        axis.Date(1, at=seq(Beginn, Ende, "years"), labels = F, tcl = -0.5)
        axis.Date(1, at=seq(Beginn+365/2, Ende-365/2, "years"),
            format = "%Y", labels = T, tcl = 0)
        legend(x = as.Date("01.04.2004", "%d.%m.%Y"), y = 28, bty = "n",
            c("Temperatur","Sauerstoff"), col = c("red","blue"),lty = 1)
    dev.off()
```

Die erzeugte Grafik sieht wie folgt aus:

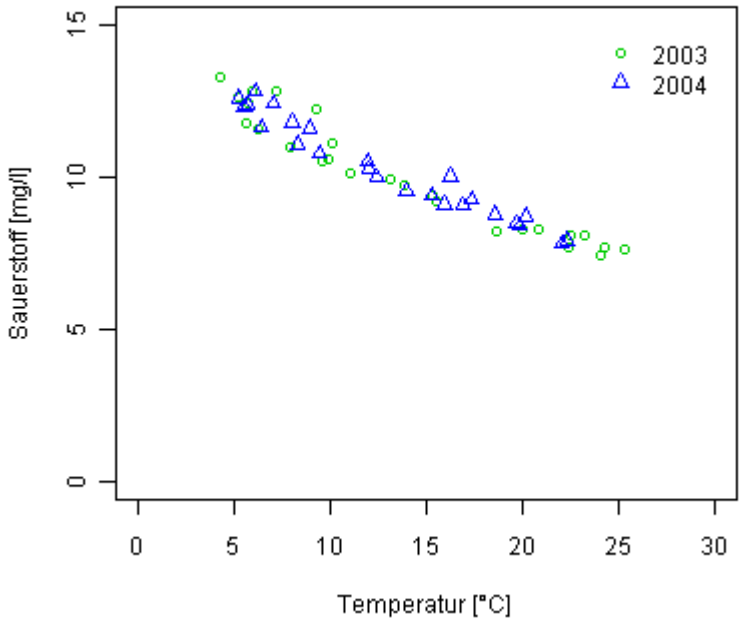


Zusatzaufgabe 4

Mit der *locator*-Funktion können Koordinaten auf der Grafikfläche ermittelt werden. Diese können beispielsweise verwendet werden, um die Position einer Legende festzulegen. Mit *locator(2)* werden die Koordinaten von zwei Punkten ermittelt.

Zusatzaufgabe 5

Die Parameter *pch* und *col* werden anhand des Jahres der Komponente *W_Jahr* festgelegt. Die Grafik sieht wie folgt aus:



Zusatzaufgabe 6

Die Balken-Darstellungen der barplot-Funktion sind den Histogrammen ähnlich. Im Gegensatz zu diesen muss die Klasseneinteilung ausserhalb der Funktion durchgeführt werden. Dafür gibt es mehr Darstellungsoptionen (Beispiele: horizontale Balken, Balken nebeneinander).

Zusatzaufgabe 7

Schiefe und Kurtosis von *W_Temp* können wie folgt berechnet werden:

```
install.packages("fBasics")
require(fBasics)
skewness(W_Temp)
kurtosis(W_Temp)
```

Zusatzaufgabe 8

Damit pro Fluss ein Boxplot erstellt wird, muss jedem Abflusswert (*Alle_Abfluss*) ein Faktor mit dem Namen des jeweiligen Flusses (*Alle_Namen*) zugeordnet werden.

Zusatzaufgabe 9

Mit der Funktion *image.plot* im Paket *fields* können auf einfache Weise räumlich-verteilte Daten (Grid-Raster) dargestellt werden.

3.6. Zusatzaufgaben zu Kapitel 7

Zusatzaufgabe 1

Die folgende Tabelle erläutert Funktionsweisen oder Anwendungsbereiche von relativ häufig verwendeten, in diesem Leitprogramm aber nicht diskutierten Verteilungen.

Funktion	*Beschreibung*	*Erläuterungen*
_geom()	Geometrische Verteilung	Spezialfall der Binomialverteilung; Beispiel: Wie viele Jahre vergehen, bis ein bestimmtes Ereignis eintritt?
_hyper()	Hypergeometrische Verteilung	Ähnlich wie die Binomialverteilung, aber die Versuche sind nicht unabhängig („ohne Zurücklegen“).
_multinom()	Multinomiale Verteilung	Ähnlich wie die Binomialverteilung, aber es sind mehr als zwei Zustände möglich.
_nbinom()	Negative Binomialverteilung	Ähnlich wie die Binomialverteilung. Hier wird aber die Wahrscheinlichkeit des k-ten Erfolgs im x-ten Versuch berechnet.
_pois()	Poissonverteilung	Ähnlich wie die Binomialverteilung, aber die Anzahl Versuche ist hoch, die Eintretenswahrscheinlichkeit gering.
_beta()	Betaverteilung	Wird in Bayesscher Statistik verwendet; ist die konjugierte a-priori Verteilung der Binomialverteilung.
_chisq()	Chiquadratverteilung	Sie Summe von quadrierten standard-normalverteilten Zufallsvariablen ist Chi-quadrat-verteilt. Wird beispielsweise bei Anpassungstests verwendet.
_exp()	Exponentialverteilung	Die Dauer von zufälligen Zeitintervallen ist oft exponentialverteilt.
_f()	F-Verteilung	Der Quotient zweier chi-quadrat-verteilten Zufallsvariabeln ist F-verteilt. Anwendung: Varianzanalyse (ANOVA), Vergleich von zwei Standardabweichungen.
_frechet()	Frechet-Verteilung	Extremwertverteilung.
_gev()	GEV-Verteilung	Extremwertverteilung.
_t()	t-Verteilung	Bei unbekannter Standardabweichung, ist der standardisierte Mittelwert normalverteilter Daten t-verteilt. Anwendung: Vergleich von zwei Mittelwerten.

Zusatzaufgabe 2

Mit der *choose*-Funktion wird der Binomialkoeffizient berechnet:

$$choose(n,k) = \frac{n!}{(n-k)!\cdot k!}$$

In der Grafik auf Seite 123 des Leitprogramms entspricht der Binomialkoeffizient der Anzahl der grauen Pfade.

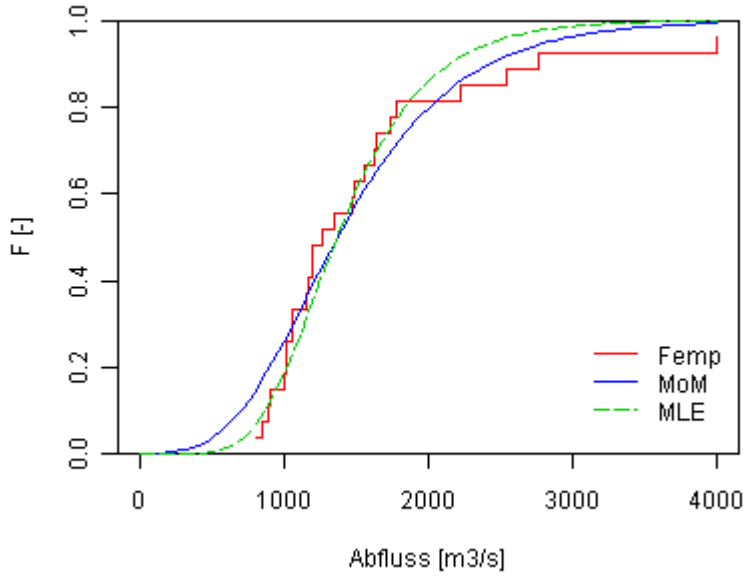
Zusatzaufgabe 3

Die *MLE*-Parameter der Gumbelverteilung können mit dem folgenden Code berechnet werden:

```
require(stats4)                 #wegen der mle-Funktion
require(evd)                    #wegender gumbel-Funktion

MW_Pi <- mean(Pi$Abfluss)
ST_Pi <- sd(Pi$Abfluss)

a_Pi1 <- MW_Pi - 0.577216/pi*sqrt(6)*ST_Pi
b_Pi1 <- sqrt(6)/pi*ST_Pi

MLE_GUMB1 <- function (a=a_Pi1, b=b_Pi1)
      {-sum(log(dgumbel(Pi$Abfluss, a, b)))}
Resultat <- mle(MLE_GUMB1)

a_Pi2 <- as.numeric(coef(Resultat)[1])
b_Pi2 <- as.numeric(coef(Resultat)[2])

qgumbel(1-1/100,a_Pi1,b_Pi1)
qgumbel(1-1/100,a_Pi2,b_Pi2)
```

Die MoM-Parameter (*a1* = 1168.02, *b1* = 559.6144) unterscheiden sich deutlich von den MLE-Parametern (*a2* = 1216.1577, *b2* = 414.2496), die Abschätzung des hundertjährlichen Hochwassers ebenfalls (MoM: 3742 m^3/s, MLE: 3122 m^3/s).

Dieser Vergleich zeigt unter anderem die beschränkte Aussagekraft solcher Abschätzungen auf. Die Unsicherheiten sind besonders bei den (für Hochwasserabschätzungen wichtigen) hohen Quantilen beträchtlich. Auf der nächsten Seite befindet sich eine Grafik mit den beiden Verteilungsfunktionen (MoM/MLE) sowie der empirischen Verteilung.



Diese Grafik ist mit folgendem Code erstellt worden:

```
bmp("Pine_Gumbel1.bmp", width = 400, height = 320)
        par(mar = c(5.1, 4.1, 2.1, 2.1))
                plot(sort(Pi$Abfluss),Femp,
                        type = "s", col = "red",
                        xlim = c(0,4000), xlab = "Abfluss [m3/s]", ylab = "F [-]")
                curve(pgumbel(x,a_Pi1,b_Pi1),
                        from=0, to=4000, col = "blue", add = T)
                curve(pgumbel(x,a_Pi2,b_Pi2), lty = 5,
                        from=0, to=4000, col = "green3", add = T)
                legend("bottomright", bty = "n", lty = c(1, 1, 5) ,
                        col = c("red", "blue", "green3"),
                        c("Femp","MoM","MLE"))
dev.off()
```

Zusatzaufgabe 4

Ricci (2005) stellt in übersichtlicher Weise vor, wie Verteilungsfunktionen geschätzt werden (von Modellwahl, über Parameterschätzung, zu Tests). Zu den Stärken diese Dokumentation gehören die anschauliche Einführung der gebräuchlichsten Modelle sowie die Auswahl an Testverfahren, um auf Normalverteilung zu prüfen.

Zusatzaufgabe 5

Mit diesem Paket können der Anderson-Darling-, Cramer-von Mises-, Kolmogorow-Smirnow-, Chi-Quadrat- und Shapiro-Francia-Test durchgeführt werden, um auf Normalverteilung zu prüfen.

Zusatzaufgabe 6

Die t.test-Funktion kann wie folgt angewendet werden, um auf Unterschiede zwischen zwei Mittelwerten zu testen:

```
x <- 1:10
y <- 1:11
z <- 15:20
t.test(x,y)
t.test(x,z)
```

Beispiele für Test-Funktionen: chisq.test und wilcox.test:

```
x <- c(89,37,30,28,2)
p <- c(40,20,20,15,5)
chisq.test(x, p = p, rescale.p = TRUE)

x <- c(1.83,  0.50,  1.62,  2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29
wilcox.test(x, y, paired = TRUE, alternative = "greater")
```

Zusatzaufgabe 7

Die *matplot*-Funktion kann wie folgt verwendet werden:

```
W_Temp2 <- sqrt(W_Temp)
W_Modell_2 <- lm (W_O2 ~ W_Temp2)
SEQ <- seq(2,5,0.1)
CONF_W <- predict.lm(W_Modell_2, data.frame(W_Temp2= SEQ),
            interval="confidence", level= 0.95)
matplot(SEQ, CONF_W, type = "l", lty = c(1,2,2), col = "black")
```

Zusatzaufgabe 8

Operatoren werden in einem Modell grundsätzlich verwendet, um Variablen zu dem Modell hinzuzufügen, sie aus dem Modell herauszunehmen, oder um Wechselwirkungen zwischen Variabeln einzufügen. Innerhalb *I()*, werden Operatoren tatsächlich als arithmetische Operatoren verwendet. Beispielsweise wird das in x_2 quadratische Modell $y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i}{}^2 + ... + \varepsilon_i$ wie folgt geschrieben:

```
y ~ x1 + I(x2 ^2)
```

Zusatzaufgabe 9

A) Verstehen der geometrischen Verteilung:

Mit der geometrischen Verteilung wird die Wahrscheinlichkeit berechnet, dass ein Ereignis (beispielsweise ein Hochwasserereignis) mit einer Auftretens-wahrscheinlichkeit von *p* nach *n*-Jahren auftritt:

$$P(n) = p \cdot (1-p)^{n-1},\ n = 1, 2, 3, \ldots$$

Das heisst, dass während *n-1* Jahren kein Ereignis auftritt. Das Ereignis kommt erst im *n*-ten Jahr vor. Die geometrische Verteilung kann von der Binomialverteilung abgeleitet werden:

$$P(n,k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

Dabei gilt $k = 1$, weil das Ereignis nur einmal auftritt. Zudem wird der Binomialkoeffizient $\binom{n}{k}$ weggelassen, weil nur ein Fall von Interesse ist, nämlich derjenige, bei dem das Ereignis im *n*-ten Jahr auftritt, nachdem es in den *n-1* vorhergehenden Jahren nicht vorgekommen ist. Der Erwartungswert der geometrischen Verteilung ist definiert als

$$E[N] = \frac{1}{p}$$

Das heisst beispielsweise, dass ein Ereignis mit einer jährlichen Auftretenswahrscheinlichkeit von 0.05 im Mittel alle 20 Jahre auftritt. Obwohl dies vernünftig klingt, ist es auf den ersten Blick nicht einfach nachzuvollziehbar. Deshalb wird die Erwartungswertberechnung in R veranschaulicht. Hinweis: In R ist die geometrische Verteilung wie folgt definiert:

$$P(x) = p \cdot (1-p)^{x},\ x = 0, 1, 2, \ldots$$

Es wird folgendes in R eingegeben:

```
n <- 1:1000
x <- n-1
p <- 1/20

Pr <- dgeom(x,p)
E <- n * Pr
A <- cbind(n, Pr, E)
A [1:10,]
sum(E)
```

Die Ausgabe sieht wie folgt aus:

```
> A [1:10,]
       n          Pr         E
 [1,]  1 0.05000000 0.0500000
 [2,]  2 0.04750000 0.0950000
 [3,]  3 0.04512500 0.1353750
 [4,]  4 0.04286875 0.1714750
 [5,]  5 0.04072531 0.2036266
 [6,]  6 0.03868905 0.2321343
 [7,]  7 0.03675459 0.2572822
 [8,]  8 0.03491686 0.2793349
 [9,]  9 0.03317102 0.2985392
[10,] 10 0.03151247 0.3151247
> sum(E)
[1] 20
```

Dabei enthält die Spalte Pr die Wahrscheinlichkeit, dass das Ereignis erst im *n*-ten Jahr auftritt. Die Jahre *n*, in denen das Ereignis zum ersten Mal auftritt, werden mit der Wahrscheinlichkeit, dass das Ereignis im *n*-ten Jahr auftritt, gewichtet. Das ist in der Spalte *E* dargestellt. Die Summe von E über alle Jahre von 1 bis 1000 (theoretisch von 1 bis unendlich) ergibt den Wert 20! In einem nächsten Schritt soll die Erwartungswertdefinition bewiesen werden.

B) Beweis des Erwartungswerts der geometrischen Verteilung

Der Erwartungswert wird wie folgt berechnet:

$$(1)\quad E[N] = \sum_{n=1}^{n_\infty} n \cdot p \cdot (1-p)^{n-1}$$

Dabei bedeutet das Symbol $n_\infty = n_{unendlich}$, dass über alle Jahre aufsummiert wird. Weil p nicht von n abhängt, kann p vor das Summenzeichen gezogen werden. Zudem wird die Auftretenswahrscheinlichkeit p teilweise mit der Nichtauftretens-Wahrscheinlichkeit $q = 1 - p$ ersetzt.

$$E[N] = p \cdot \sum_{n=1}^{n_\infty} n \cdot (1-p)^{n-1}$$

$$(2)\quad E[N] = p \cdot \underbrace{\sum_{n=1}^{n_\infty} n \cdot (q)^{n-1}}_{(3)}$$

Gleichung (3) kann wie folgt ausgeschrieben werden:

$$\sum_{n=1}^{n_\infty} n \cdot (q)^{n-1} = 1 \cdot q^0 + 2 \cdot q^1 + 3 \cdot q^2 + 4 \cdot q^3 + \ldots$$

$$(4)\quad \sum_{n=1}^{n_\infty} n \cdot (q)^{n-1} = 1 + 2 \cdot q + 3 \cdot q^2 + 4 \cdot q^3 + \ldots$$

An dieser Stelle wird ein Exkurs gemacht.

Gegeben ist die folgende Summe S_n, welche anschliessend mit x multipliziert wird.

$$S_n = 1 + x + x^2 + x^3 + x^4 + \ldots$$

$$x \cdot S_n = x + x^2 + x^3 + x^4 + x^5 \ldots$$

Falls $0 < x < 1$, gilt folgendes, wenn die Summe xS_n von S_n subtrahiert wird:

$$\begin{array}{rl} & S_n = 1 + x + x^2 + x^3 + x^4 + \ldots \\ - & x \cdot S_n = x + x^2 + x^3 + x^4 + x^5 \ldots \\ \hline & S_n - x \cdot S_n = 1 \end{array}$$

Daraus folgt:

$$S_n \cdot (1 - x) = 1$$

$$S_n = \frac{1}{1 - x}$$

Fazit:

$$1 + x + x^2 + x^3 + x^4 + \ldots = \frac{1}{1 - x}$$

$$1 + x + x^2 + x^3 + x^4 + \ldots = (1 - x)^{-1}$$

Nun wird die Gleichung auf beiden Seiten nach x abgeleitet (der letzte Faktor auf der rechten Seite der Gleichung ist die innere Ableitung von *1-x*):

$$0 + 1 \cdot x^0 + 2 \cdot x^1 + 3 \cdot x^2 + 4 \cdot x^3 + \ldots = (-1) \cdot (1 - x)^{-2} (-1)$$

$$1 + 2 \cdot x + 3 \cdot x^2 + 4 \cdot x^3 + \ldots = (1 - x)^{-2}$$

$$(5) \quad 1 + 2 \cdot x + 3 \cdot x^2 + 4 \cdot x^3 + \ldots = \frac{1}{(1 - x)^2}$$

Aus Gleichung (4) folgt mit Gleichung (5):

$$\sum_{n=1}^{n_\infty} n \cdot (q)^{n-1} = \frac{1}{(1 - q)^2}$$

Damit folgt aus Gleichung (2):

$$E[N] = p \cdot \frac{1}{(1-q)^2}$$

$$E[N] = p \cdot \frac{1}{p^2}$$

$$E[N] = \frac{1}{p}$$

4. Quellen

4.1. Quellen für die Studierenden

Die wichtigste externe Informationsquelle für die Studierenden ist die R-Webseite (www.r-project.org). Dort sind vor allem die Seiten zu „Frequently Asked Questions“ sowie „Documentations and Manuals“ von Interesse. Zudem wird empfohlen, den Studierenden die Bücher Ligges (2007) und Crawley (2007) zur Verfügung zu stellen.

4.2. Zusätzliche Publikationen

Die untenstehende Tabelle listet die Publikationen auf, die zum Erstellen dieses Leitprogramms verwendet wurden. Dabei sind diejenigen Veröffentlichungen gekennzeichnet, die der Lehrperson besonders zu empfehlen sind, um mit R zu arbeiten.

Publikation	*Für R besonders empfehlenswert bezüglich…*						
	Ein-führung	*Daten-analysen*	*Grafiken*	*Program-mieren*	*Geo-statistik*	*Zeitreihen analysen*	*Bayessche Statistik*
Albert (2009)							X
Behr (2005)							
Bivand et al. (2008)					X		
Braun und Murdoch (2007)							
Bürgisser (2003)							
Chambers (2008)		X		X			
Cowpertwait und Metcalfe (2009)						X	
Crawley (2005)							
Crawley (2007)	X						
Cryer und Chan (2008)						X	
Dalgaard (2002)	X						
Diggle und Ribeiro (2007)							
Dolić (2004)							
Everitt und Hothorn (2006)							
Faes (2007)							
Faraway (2004)							
Faraway (2005)	X						
Frey und Frey-Eiling (2003)							
Frey et al. (2008)							
Grothendieck und Petzoldt (2004)	X						
Kirchgraber et al. (2003)							

Publikation	*Für R besonders empfehlenswert bezüglich…*						
	Ein-führung	*Daten-analysen*	*Grafiken*	*Program-mieren*	*Geo-statistik*	*Zeitreihen analysen*	*Bayessche Statistik*
Kuhnert und Venables (2005)	X						
Ligges (2007)	X						
Maindonald und Braun (2003)	(X)						
Marin (2007)							
Murrell (2006)	X						
Pfaff (2008)						(X)	
Ricci (2005)	X						
Robert und Casella (2010)				X			
Robinson (2008)							
Schlittgen (2001)							
Shumway und Stoffer (2006)						X	
Sachs und Hedderich (2006)							
Sarkar (2008)			(X)				
Spector (2008)		X					
Verzani (2005)							
Wickham (2009)			X				
Zidek und Le (2006)							

4.3. Daten für die Kapiteltests

Für die Kapiteltests wurden die folgenden, öffentlichen Datenquellen verwendet:

Dateiname	*Beschreibung*	*Quelle*
Abfluss_2001.xls	Abfluss des Rheins bei Weil am Rhein in m^3/s im Jahr 2001 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Abfluss_2002.xls	Abfluss des Rheins bei Weil am Rhein in m^3/s im Jahr 2002 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Abfluss_2003.xls	Abfluss des Rheins bei Weil am Rhein in m^3/s im Jahr 2003 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Abfluss_2004.xls	Abfluss des Rheins bei Weil am Rhein in m^3/s im Jahr 2004 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Baden.csv	Tagesmittelwerte der Wassertemperatur der Limmat in Baden vom 1.6. bis 31.8. 2007.	Bundesamt für Umwelt (BAFU, 2008)
Chlorophyll.csv	Mittlere und maximale Konzentration von Chlorophyll_a in µg/l im Rhein im Jahr 2000; neun verschiedene Standorte.	Internationale Kommission zum Schutz des Rheins (IKSR, 2000)
Duck_Creek.csv	Maximale instantane Jahresabflüsse des Duck Creek (Stationsnummer: 08NH016); nicht reguliertes Gewässer; 34 Werte von 1947 bis 2006.	Environment Canada (2006)
HydroU.csv	11 hydrologische Untersuchungsgebiete der Schweiz. Kenngrössen: GEW (Gewässername), HOEHE (mittlere Einzugsgebietshöhe), WALD (Waldanteil in %), GLET (Gletscheranteil in %), SEE (Seeanteil in %), DURCH (mittlere Durchlässigkeit in cm/s), GEWD (Gewässernetzdichte in km/km^2), BEVD (Bevölkerungsdichte in Personen/km^2), P06 (Gebietsniederschlag 2006 in mm), PMAI06 (Gebietsniederschlag Mai 06 in mm), Q06 (Gebietsabfluss 2006 in mm), QMAI06 (Gebietsabfluss im Mai 06 in mm). TMAI (mittlere Wassertemperatur im Mai 2006 in °C), TAUG (mittlere Wassertemperatur im August 2006 in °C).	Bundesamt für Umwelt (BAFU, 2006) Bundesamt für Umwelt (BAFU, 2008)
Naehrstoffe.csv	Mittlere Konzentration von Phosphat und Nitrat in mg/l im Rhein im Jahr 2000; sieben verschiedene Standorte.	Internationale Kommission zum Schutz des Rheins (IKSR, 2000)

Dateiname	*Beschreibung*	*Quelle*
Schwebstoffe_2001.csv	Schwebstoffe des Rheins bei Weil am Rhein in mg/l Jahr 2001 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Schwebstoffe_2001.csv	Schwebstoffe des Rheins bei Weil am Rhein in mg/l Jahr 2001 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Schwebstoffe_2002.csv	Schwebstoffe des Rheins bei Weil am Rhein in mg/l Jahr 2001 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Schwebstoffe_2003.csv	Schwebstoffe des Rheins bei Weil am Rhein in mg/l Jahr 2001 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Schwebstoffe_2004.csv	Schwebstoffe des Rheins bei Weil am Rhein in mg/l Jahr 2001 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Stickstoff.csv	Stickstoffkonzentrationen (in mg-N/l) des Rheins bei Weil am Rhein von 1995 bis 2004 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)
Surprise_Creek.csv	Maximale instantane Jahresabflüsse des Surprise Creek (Stationsnummer: 08DA005); nicht reguliertes Gewässer; 34 Werte von 1967 bis 2006.	Environment Canada (2006)
Weil_Metallionen.csv	Calcium-, Kalium-, Magnesium-, und Natrium-Konzentration (in mg/l) des Rheins bei Weil am Rhein in den Jahren 2002 und 2003 (14-tägige Werte).	Internationale Kommission zum Schutz des Rheins (IKSR, 2008)

5. Literaturverzeichnis

1 Albert J. 2009. Bayesian computation with R. Springer, Berlin. ISBN: 978-0-387-92297-3

2 Behr A. 2005. Einführung in die Statistik mit R. WiSo Kurzlehrbücher. Vahlen, München.

3 Bivand, R.S., Pebesma, E. J., Gómez-Rubio V. 2008. Applied Spatial Data Analysis with R. Use R. Springer, Berlin ISBN: 978-0-387-78170-9.

4 Bundesamt für Umwelt (BAFU) 2006. Hydrologisches Jahrbuch der Schweiz. http://www.bafu.admin.ch/php/modules/shop/files/pdf/phpKNnhzt.pdf. 1. August 2008.

5 Bundesamt für Umwelt (BAFU) 2008. Hydrologische Daten von 361 Stationen. http://www.hydrodaten.admin.ch/d/index.htm?lang=de. 1. August 2008.

6 Braun J. und Murdoch D. J. 2007. A first course in statistical programming with R. Cambridge University Press, Cambridge.

7 Bürgisser D. 2003. Grundlagen der organischen Chemie. Ein Leitprogramm für den Grundlagen-Chemieunterricht der Sekundarstufe II. Neue Kantonsschule Aarau.

8 Chambers J. M. 2008. Software for Data Analysis: Programming with R (Statistics and Computing). Springer, Berlin.

9 Cowpertwait, P. S. P. und Metcalfe, A. V. 2009. Introductory Time Series with R. Use R. Springer, Berlin. ISBN: 978-0-387-88697-8.

10 Crawley M. J. 2005. Statistics an introduction using R. Wiley & Sons, New York.

11 Crawley M. J. 2007. The R book. Wiley & Sons, New York.

12 Cryer D. J. und Chan K.-S. 2008. Time series analysis with applications in R. Springer, Berlin. ISBN: 978-0-387-75958-6.

13 Dalgaard, P. 2002. Introductory statistics with R. Springer, Berlin.

14 Diggle P. J. und Ribeiro P. J. 2007. Model-based geostatistics. Springer, New York.

15 Dolić D. 2004. Statistik mit R. Einführung für Wirtschafts- und Sozialwissenschaftler. Oldenbourg Verlag, München, Wien.

16 Environment Canada 2006. Water Survey of Canada. http://www.wsc.ec.gc.ca/hydat/H2O/index_e.cfm?cname=main_e.cfm http://scitech.pyr.ec.gc.ca/climhydro/mainContent/main_e.asp?province=bc, 1. August 2008.

17 Everitt B. S. und Hothorn T. 2006. A handbook of statistical analyses using R. Chapman & Hall/CRC, Boca Raton.

18 Faes G. 2007. Einführung in R. Ein Kochbuch zur statistischen Datenanalyse mit R. BoD, Norderstedt.

19 Faraway J. J. 2004. Linear Models with R. Chapman & Hall/CRC, Boca Raton.

20 Faraway J. J. 2005. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman & Hall/CRC, Boca Raton.

21 Frey K. und Frey-Eiling A. 2003. Allgemeine Didaktik. Vorlesungsunterlagen. Institut für Verhaltenswissenschaften, ETH Zürich, Zürich.

22 Frey K., Frey-Eiling A., Mandrin, P., Preckel D. 2008. Manual zur Entwicklung von Leitprogrammen. Institut für Verhaltenswissenschaften, ETH Zürich, Zürich.

23 Global Runoff Data Centre (GRDC) 2008. http://grdc.bafg.de/servlet/is/Entry.987.Display/. 1. August 2008.

24 Grothendieck und Petzoldt 2004. R Help Desk, Date and Time Classes in R. Volume 4/1, June 2004, Seite 29. http://cran.r-project.org/doc/Rnews/

25 Institute for Atmospheric and Climate Science (IAC), ETH Zürich. 2008. http://www.iac.ethz.ch/research/rietholzbach/datasets. 1. August 2008.

26 Internationale Kommission zum Schutz des Rheins (IKSR) 2000. Plankton im Rhein 2000. Bericht Nr. 129d. Internationale Kommission zum Schutz des Rheins. Koblenz. http://www.iksr.org/index.php?id=30. 1. August 2008.

27 Internationale Kommission zum Schutz des Rheins (IKSR) 2008. Gewässergütedaten. http://www.iksr.org/index.php?id=71, 1. August 2008.

28 Kirchgraber U., Bettinaglio M., Stoffer D. und Weber C. 2003. Lineare Gleichungssysteme. Ein Leitprogramm in Mathematik. ETH-Leitprogramme. ETH Zürich, Zürich.

29 Kuhnert P. und Venables B. 2005. An Introduction to R: Software for Statistical Modelling & Computing. CSIRO Mathematical and Information Sciences Cleveland, Australia.

30 Ligges U. 2007. Programmieren mit R. Springer-Verlag, Heidelberg.

31 Maindonald J. und Braun J. 2003. Data Analysis and Graphics Using R. Cambridge University Press, Cambridge.

32 Marin J.-M. 2007. Bayesian core: a practical approach to computational Bayesian statistics. Springer, New York.

33 Murrell P. 2006. R Graphics. Chapman & Hall/CRC, Boca Raton.

34 Pfaff, B. 2008. Analysis of Integrated and Cointegrated Time Series with R. Use R. Springer, Berlin. ISBN: 978-0-387-75966-1.

35 Ricci V. 2005. Fitting Distributions with R. http://www.r-project.org/

36 Robert, P. und Casella G. 2010. Introducing Monte Carlo Methods with R. Use R. Springer, Berlin. ISBN: 978-1-4419-1575-7.

37 Robinson, R. 2008. icebreakeR. Department of Mathematics and Statistics. University of Melbourne, Melbourne.

38 Sachs L. und Hedderich J. 2006. Angewandte Statistik: Methodensammlung mit R. Springer, Berlin.

39 Sarkar, D. 2008. Multivariate Data Visualization with R. Use R. Springer, Berlin. ISBN: 978-0-387-75968-5.

40 Schlittgen R. 2001. Angewandte Zeitreihenanalyse. Oldenbourg, München.

41 Shumway R. H. und Stoffer D. S. 2006. Time Series Analysis and Its Applications With R Examples. Springer, New York. ISBN 978-0-387-29317-2.

42 Spector P. 2008. Data Manipulation with R. Springer, New York.

43 United States Geological Survey (USGS) 2008. http://water.usgs.gov/data/ 1. August 2008.

44 Verzani, J. 2005. Simple R. Using R for Introductory Statistics. Chapman & Hall/CRC, Boca Raton.

45 Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Use R. Springer. ISBN: 978-0-387-98140-6.

46 Zidek J. V. und Le N. D. 2006. Statistical Analysis of Environmental Space-Time Processes. Springer, New York.