Programmieren mit Statistischer Software **14. Block**
Grün **SS 2017**

14 Projektarbeit

Im Folgenden wollen wir Teile von Schepaschenko et al. (2017) mithilfe von R reproduzieren und an den zu dieser Publikation zur Verfügung gestellten Daten weitere Analysen durchführen. Der Artikel kann von der folgenden Webpage heruntergeladen werden:

`https://www.nature.com/articles/sdata201770`.

Die Daten sind verfügbar unter:

`https://doi.pangaea.de/10.1594/PANGAEA.871465`
`https://doi.pangaea.de/10.1594/PANGAEA.871491`

[redacted]

2. Versuchen Sie Tabelle 2 zu reproduzieren.
   - Schreiben Sie eine Funktion, die einen Vektor als Input nimmt und einen benannten Vektor der Länge 5 retouniert, die die Anzahl der nicht-`NA` Werte, der unterschiedlichen Werte für Buchstaben-Vektoren und Median, Minimum und Maximum für numerische Vektoren enthält und ansonsten `NA`. Überprüfen Sie innerhalb der Funktion, dass das Argument richtig spezifiziert wurde bzw. überlegen Sie, wie Vektoren konvertiert werden könnten, sodass sinnvolle Resultate zurückgegeben werden.
   - Wenden Sie die Funktion auf die Spalten des Datensatz mit den Bäumen an.
   - Vergleichen Sie die Ergebnisse und erläutern Sie, welche Unterschiede Sie feststellen.
   - Die Variable `ID_Plot` erlaubt die beiden Datensätze zu verknüpfen. Gibt es Beobachtungen bei den Bäumen, die nicht mit den Plots verknüpft werden können?
3. Versuchen Sie Tabelle 1 zu reproduzieren. Vergleichen Sie die Ergebnisse und erläutern Sie, welche Unterschiede Sie feststellen.
4. Bestimmen Sie jeweils die Länderverteilung für die Bäume und Grundstücke ("plots") und reproduzieren Sie damit Tabelle 3. Visualisieren Sie die Verteilungen auch geeignet.

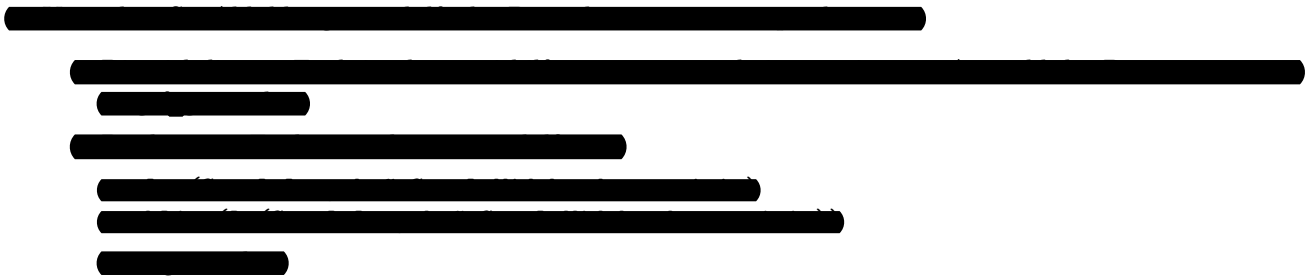
[redacted]

6. Bestimmen Sie jeweils die Baumgattungsverteilung für die Bäume und Grundstücke ("plots") und versuchen Sie damit Tabelle 5 zu reproduzieren. Verwenden Sie dazu jeweils den Beginn des Eintrags bei Tree species bei den Plots und Species bei den Bäumen. Visualisieren Sie die Verteilungen auch geeignet.
7. Erstellen Sie einen parallelen Boxplot der Biomasse über dem Boden gegeben Biom und vergleichen Sie ihn mit Abbildung 3.
8. Wir wollen die Variable Anzahl an Bäumen charakterisieren, die für einen Plot zur Verfügung stehen. Dabei ist das Problem, dass diese Variable sowohl genaue numerische Einträge enthält, als auch Bereiche, z.B. 2–3.

- Die Variable soll über Mittelwert, Median, Standardabweichung und Interquartilsabstand zusammengefasst werden. Schreiben Sie eine Funktion, die eine numerische Variable als Input hat und einen Vektor mit diesen Werten retouniert. Die Funktion soll ein `...` Argument haben, womit dann `na.rm` geeignet gesetzt werden kann.
- Schreiben Sie eine Funktion, die einen Buchstabenvektor als Input nimmt und diesen in einen numerischen umwandelt indem:
  - die in numerische Werte umwandelbaren Einträge umgewandelt werden;
  - Bereichswerte durch einen Repräsentanten ersetzt werden.

  Die Funktion soll ein Argument method haben, wie der Repräsentant bestimmt wird:
  - `"mean"`: Mittelwert des Bereichs
  - `"min"`: kleinster Wert des Bereichs
  - `"max"`: größter Wert des Bereichs
  - `"uniform"`: Gleichverteilt gezogen aus dem Bereich

  Die Funktion soll die Argumente geeignet überprüfen.
- Wenden Sie die Funktion auf die Variable Anzahl an Bäumen an und vergleichen Sie die Ergebnisse mit denen, wo nur die konkreten Werte verwendet wurden. Interpretieren Sie das Ergebnis.



Bitte geben Sie Folgendes ab:

- `PDF`-Datei, die den Bericht mit den Ergebnissen enthält,
- Das `R`-Skript sowie etwaige weitere Dateien, die zum Reproduzieren der `PDF`-Datei notwendig sind. Achten Sie dabei darauf, dass der Code gut formatiert, ausreichend kommentiert und gut lesbar ist.

Sie können das Projekt in Gruppen von 1–3 Personen machen. Bitte führen Sie Namen und Matrikelnummer am Deckblatt des Berichts an. Bei E-Mailabgabe ist es vorteilhaft, alle Projektteilnehmer anzuführen, damit diese auch etwaige Antworten erhalten.

Literatur

Schepaschenko, D., A. Shvidenko, V. Usoltsev, P. Lakyda, Y. Luo, R. Vasylyshyn, I. Lakyda, Y. Myklush, L. See, I. McCallum, S. Fritz, F. Kraxner, und M. Obersteiner (2017). A Dataset of Forest Biomass Structure for Eurasia. *Scientific Data 4*(170070), 1–11.