

erstellt: 6. Dezember 2018

Hausaufgabe (WaST3), ZHAW-SoE, HS 18

Wir arbeiten mit aktuellen Luftqualitätsmessungen vom Umwelt- und Gesundheitsschutz der Stadt Zürich (UGZ). Diese werden in der Datenbank OSTLUFT gespeichert und sind als *open data* erhältlich.

Sie können die Daten von hier beziehen:

<https://data.stadt-zuerich.ch/dataset/luftqualitaet-historisierte-messungen> (seit 2012)

Wir wollen uns die Feinstaubmessungen, NO₂, NO, Ozonmengen als auch metrologische Variablen ansehen.

Vor allem die Feinstaubmessungen sind von Relevanz. Als PM10 werden Partikel bezeichnet, deren Durchmesser weniger als 10 Tausendstel-Millimeter beträgt. Deshalb können diese Partikel tief in die feinsten Verästelungen der Lunge eindringen und von dort zum Teil in die Lymph- und Blutbahnen. Ihre zerklüftete Struktur ermöglicht eine Anlagerung von weiteren giftigen Substanzen. Dieses Schadstoffgemisch besteht aus einer Vielzahl von chemischen Verbindungen mit teils Krebs erzeugender Wirkung z.B. Russ. Es verursacht in den Atemwegen lokale Entzündungen und kann so zu schwerwiegenden Auswirkungen auf die Gesundheit führen.

Sie können in Gruppen bis zu drei Personen zusammenarbeiten. Bitte schreiben Sie von keiner anderen Gruppe/Klasse ab um Punkteabzug zu vermeiden.

Aufgabe 1: Datenimport und Datenmanagement

Vorweg: dies ist die zeitintensivste und schwierigste aller Aufgaben.

Laden Sie sich die Daten herunter und importieren Sie die Daten in R. Sie werden einige Probleme beim Import und Datenaufbereitung zu lösen haben. Achten Sie darauf, dass die Daten in R richtig dargestellt werden, die entsprechenden Variablennamen aufweisen, die Messwerte vom Typ **numeric** sind und kategoriale Variablen den Typ **factor** aufweisen. Statistiker wollen Reproduzierbarkeit, die Originaldaten sollen nicht angegriffen werden, nur (R) Code sollte die Daten nach Import verändern. Verwenden Sie deshalb ausschliesslich R Code und lassen Sie das csv unverändert. Achten Sie auf long-Format bzgl. den Standorten (Stichwort: *tidy data*). Falls Sie die Datenaufbereitung in R nicht schaffen, nehmen Sie einen Punkteabzug in Kauf und bereiten Sie die Daten *von Hand* auf, damit Sie fortfahren können.

In anderen Worten: lesen Sie die Datenbeschreibung und nehmen Sie sich Zeit die Daten schön für Ihre Analysen aufzubereiten.

Bemerkung: Nachdem das Wetter sich nicht gross zwischen den Stationen unterscheiden wird und es keine näher gelegenen metrologischen Stationen geben wird, nehmen wir die Wetterdaten der Station Stampfenbachstrasse als Wetterinformation gültig für alle Feinstaubmessstationen.

Tipps: möglicherweise ist die Zeichenkodierung dieses Files nicht konform mit Ihrer Einstellung. Achten Sie auf Funktionsparameter *encoding* und probieren Sie zum Beispiel `latin1`. Falls Sie nicht verstehen, was ich damit meine, googlen Sie nach *Encoding R*.

Der Datensatz könnte zum Beispiel so aussehen (erste drei Zeilen):

```
##      datum O3_max_h1 ozon_warn  NO2   NO  PM10      station    T
## 1 2012-01-01    43.46      nein 26.54 7.83 18.04 Stampfenbachstrasse 8.12
## 2 2012-01-01     NaN    <NA>   NaN  NaN   NaN  Rosengartenstrasse 8.12
## 3 2012-01-01    43.75      nein 12.70 1.62   NA      Heuberg 8.12
##      Hr      p  WVs StrGlo RainDur
## 1 86.9 968.65 2.01  39.61    95.6
## 2 86.9 968.65 2.01  39.61    95.6
## 3 86.9 968.65 2.01  39.61    95.6
```

Aufgabe 2: Visualisieren und deskriptive Beschreibung der Daten

Visualisieren Sie für alle Messstationen in einer Grafik den zeitlichen Verlauf (2012 bis heute) der Konzentrationen von SO₂, CO, NO, NO₂, PM10 sowie der Konzentration des höchsten Stundenmittels von O₃. Was fällt ihnen auf (max. 1-2 Sätze)? [sätze noch](#)

Aufgabe 3: Fehlende Werte im Datensatz [R Paket: VIM, mice, nwns](#)

Gibt es fehlende Werte im Datensatz? Wieviele? Visualisieren Sie den Anteil der fehlenden Werte pro Variable sowie den Anteil der fehlenden Werte für Kombinationen von Variablen.

Kombinationen von Variablen? → zb immer x1 und x3
haben NA wieviel kommt diese Kombi vor

Wird mehrmals?!

Aufgabe 4: Grenzwertüberschreitungen

Der Jahresmittel-Grenzwert für PM10 liegt bei einer Konzentration von $20\mu\text{g}/\text{m}^3$. Der Tagesmittel-Grenzwert von PM10 liegt bei $50\mu\text{g}/\text{m}^3$ und darf höchstens einmal pro Jahr überschritten werden.

Wie oft wird der Tagesmittel-Grenzwert an welcher Station überschritten? Stellen Sie die Anzahl der überschrittenen Tagesmittel-Grenzwerte für alle drei Stationen über die Zeit in einer Grafik dar. Wie interpretieren Sie die Grafik (1-2 Sätze)?

In welchen Jahren und an welchen Stationen ist der Anteil der Tage mit Grenzwertüberschreitungen signifikant grösser als zufällig? Sie dürfen für die Beantwortung dieser Frage annehmen, dass die Messwerte iid sind, obwohl dies natürlich bei Zeitreihen im Allgemeinen nicht der Fall ist.

t verteilt

?

h0: keine Grenzwertüberschreitungen = 0
h1: Grenzwertüberschreitung > 0
Poisson.test auf jede einzelne Station, Jahr

Aufgabe 5: Einflussgrößen

Visualisieren Sie in **einer** Grafik den Zusammenhang zwischen der Regendauer und der Feinstaubkonzentration pro Station. Wie interpretieren Sie die Grafik?

Ist die Feinstaubkonzentration gemittelt über alle Messstationen statistisch signifikant tiefer an Tagen mit Regen als an Tagen ohne Regen ($\alpha = 0.01$)?

Aufgabe 6: Bevölkerungsbefragung

Auf https://data.stadt-zuerich.ch/dataset/prd_stez_bevoelkerungsbefragung finden Sie eine Bevölkerungsbefragung aus Zürich (2015). Finden Sie relevante Variablen welche auf die Feinstaubbelastung Einfluss haben könnte? Ein indirekter Proxy könnte zum Beispiel die Zufriedenheit mit dem öffentlichen Grünraum im Quartier sein. Testen Sie, ob es Unterschiede in der Zufriedenheit gibt? (bzgl. der 4 Quartiere wo sich die Messstationen befinden)

Unter

https://data.stadt-zuerich.ch/dataset/prd_stez_bevoelkerungsbefragung

finden Sie die Daten zu einer Bevölkerungsbefragung aus Zürich (2015). Verwenden Sie die Daten aus dem File *bevoelkerungsbefragung_2015-stadtentwicklung-zuerich.csv*. Wählen Sie daraus eine Variable ('fxxxxxSort'), welche auf die Feinstaubbelastung Einfluss haben könnte. Die Zufriedenheit mit dem öffentlichen Grünraum im Quartier könnte zum Beispiel ein indirekter Proxy sein.

Unterscheiden sich die Antworten der befragten Personen in den 3 unterschiedlichen Stadtkreisen wo sich die Messstationen für Feinstaubmessungen befinden signifikant ($\alpha = 0.01$)? Fassen Sie zur Beantwortung dieser Frage die Antworten aus der Bevölkerungsbefragung mit den Noten 1 - 3 zusammen, also alle Antworten mit den Noten 1-3 werden zu Antworten mit Note 3.

Um die Daten der Feinstaubimmissionen mit den Antworten aus der Bevölkerungsbefragung zu verknüpfen, müssen sie den Datensatz der Luftschadstoffe zuerst mit dem folgenden Datensatz verknüpfen:

<https://data.stadt-zuerich.ch/dataset/adressen/resource/2b60acd2-aa4e-4a09-b9ce-2b838download/adressen.csv>.

Durch diese Verknüpfung wissen Sie in welchen Stadtkreisen die Messstationen liegen.

Wie deuten Sie das Resultat (1-2 Sätze)?