Abgabe 2: Atmospheric Pollution - Feinstaubbelastung in Graz

Die Stadt Graz hat durch ihre geografische Lage schon lange Probleme mit übermässiger Feinstaubbelastung und schlechter Luftqualität (siehe, z.B., Hörmann et al. (2005), Stadlober et al. (2008) und Hörmann et al. (2021)). Um diesem Zustand entgegen zu wirken und mit den EU Standards konform zu werden, gibt es langestehende Projekte in diesem Zusammenhang, in denen Methoden der Datenanalyse dazu verwendet werden, policy Entscheidungen zu lenken. Hierzu hat die Stadt Graz sehr sorgfältig gepflegte Messungen, die Daten zur Luftverschmutzung sowie meterologische und saisonale Einflüsse genau berücksichtigen. Es ist ebenfalls bekannt, dass Luftschadstoffe mit meteorologischen Faktoren stark zusammenhängen und eine starke Saisonalität aufweisen. Die Daten, die Sie betrachten, sind Teil dieser Kampagne und sind öffentlich auf der Webseite https://www.umwelt.steiermark.at/cms/ziel/2060750/DE/ für zahlreiche Stationen in der Steiermark verfügbar.

Die bereitgestellten Daten wurden bereits sauber aufbereitet und sind nun zur Analyse bereit. Wir konzentrieren uns auf die Station **Graz Don Bosco**. Die gegebene Information umfasst *Tagesmittelwerte* (TMW), die zwischen 2015 — 2020 gesammelt wurden. Die folgenden Spalten sind inkludiert:

Spalte	**Erklärung**
Index: *DateTime*	Datum, an dem die Tagesmittelwerte ermittelt werden
day_type	eine kategorische Variable, die den Tag in Wochentag, Samstag oder Sonntag\ Feiertag teilt
humidity	TMW der Luftfeuchtigkeit in %
temp	TMW der Temperatur in Grad Celsius
no2	TMW von Stickstoffdioxid (in $\mu g/m^3$)
pm10	TMW der Feinstaubpartikel mit Durchmesser von ≤ 10 micrometer (in $\mu g/m^3$)
prec	Summe des gesamten Niederschlags in ℓ/m^2
wind_speed	TMW der Windgeschwindigkeit in m/s
peak_velocity	Maximum der Windgeschwindigkeit an diesem Tag in m/s

Das Ziel dieser Übung ist es, die Daten zu analysieren und ein Verständnis für die Art der Zusammenhänge zu entwickeln und insbesondere die Unterschiede im Verhalten für die responses *pm10* und *no2* zu betrachten. Bearbeiten Sie hierzu die folgenden Punkte und geben Sie Antworten und Interpretationen!

Vorgangsweise

Laden Sie die Daten *feinstaubdataexercise.pickle* in Ihre Umgebung.

```
import pickle as pkl

with open('insert_path_to_your_file/feinstaubdataexercise.pickle', 'rb') as file:
    dailymeansdata = pkl.load(file)

print(dailymeansdata)
dailymeansdata["Graz-DB"].head()
```

Dies ist ein Dictionary mit den keys *Graz-DB* und *Kalkleiten*. Unter dem key *Graz-DB* ist das Dataframe enthalten, welches Sie in dieser Übung betrachten sollen.

Betrachten Sie die Variablen *pm10* und *no2* als Responses und die anderen Variablen als Prädiktoren. (Wichtig: Wir wollen Schadstoffe unabhängig von einander betrachten, daher die Responses nicht als Prädiktoren im jeweils anderen Modell verwenden. Also *no2* ist kein Prädiktor in dem Modell für *pm10* und umgekehrt.) **Verwenden Sie zur Modellbildung *ausschließlich* die Daten aus 2015 – 2019**. Wir verwenden die Daten aus 2020 zu einem späteren Zeitpunkt.

1. Untersuchen Sie die Zusammenhänge der einzelnen Prädiktoren mit den beiden responses. In welchen Bereichen befinden sich die Daten? Wir sehen die Daten über die Zeit verteilt aus? Gibt es missing data? Erstellen Sie aussagekräftige Plots und machen Sie sich ein Bild über die Datenlage. [5]
2. Erstellen Sie ein erstes Modell für *pm10* und *no2* mit Ihren Daten. Wir sieht das Modell aus? Evaluieren Sie und überlegen Sie, ob Verbesserungen gut wären! Wie sieht es diagnostisch aus? Welche Prädiktoren sind sinnvoll? (ANOVA! R^2!) [3]
3. Finden Sie heraus, was **Inversion** in einem meteorologischen Zusammenhang bedeutet und wie es sich auf Luftverschmutzung auswirken kann. Sie haben Daten aus einer Wetterstation in Kalkleiten bekommen. Ermitteln Sie die Differenz der Temperaturen zwischen der Station in Graz - der Station in Kalkleiten und fügen Sie diese Variable zu Ihrem Modell hinzu. Hilft das? [2]
4. Erstellen Sie Variablen, die vielleicht hilfreich sein können, z.B. eine Variable die *frost* (temperatur < 0) anzeigt, eine Variable die *inversion* anzeigt (temperatur_differenz < 0), eine Variable die starken Wind anzeigt (z.B. *wind_speed* > 0.6), eine Variable für das Jahr, etc. Versuchen Sie, *lagged values* für meteorologische Einflüsse einzufügen, z.B. die Temperatur des Vortages etc. Helfen diese Variablen? Adaptieren Sie Ihr Modell passend! [5]
5. Versuchen Sie, Ihr Modell durch eine Transformation der Response zu verbessern. Hier kann eine Wurzeltransformation hilfreich sein, also man modelliert $\sqrt{Y}$, wobei hier Y die Response bezeichnet. Hilft das Ihrem Modell? Wie sieht es diagnostisch aus? [3]
6. Wenn Sie mit Ihrem Modell zufrieden sind, verwenden Sie dieses, um eine Vorhersage für das Jahr 2020 zu treffen. Verwenden Sie dazu die Prädiktor-Information aus dem Jahr 2020 und ermitteln Sie die fitted values. [2]
7. Vergleichen Sie Ihre Vorhersage für 2020 mit den wahren Werten aus 2020 in einem passenden Plot über die ersten 6 Monate. Was ist hier passiert? Haben Sie eine Interpretation für das Phänomen? (Tipp: Remember Covid? Suchen Sie sich die genauen Daten der Lockdowns und ersten Lockerungen heraus und zeichnen Sie diese in Ihrer Darstellung ein.) [5]

Methoden

Verwenden Sie eine Mischung aus Kennzahlen und grafischen Darstellung zur Bearbeitung der Probleme. Modellieren Sie die Daten mithilfe von `statsmodels`. Verwenden Sie zur Darstellung gerne libraries Ihrer Wahl und versuchen Sie verschiedene Methoden auszuprobieren, um die Daten von allen Seiten zu beleuchten.

Abgabeformat

Bearbeiten Sie diese Aufgabe in Teams von 2 (bzw. 3). Geben Sie pro Team nur eine Abgabe ab. Eine passende Abgabe besteht aus:

- Ein Jupyter-Notebook, das den gesamten Source Code enthält und die Resultate in Markdown-Blöcken beschreibt
- Oder ein (oder mehrere) .py files für den Source Code **und** einem pdf, das die Resultate beschreibt.

Alternativ kann das Juypternotebook auch nur den Source Code beinhalten und ein Report kann separat als pdf abgegeben werden.

Der Code soll **ausführbar sein**.

Geben Sie unbedingt ein **.zip** File ab!

Bibliography

Hörmann, S., Jammoul, F., Kuenzer, T., & Stadlober, E. (2021). Separating the impact of gradual lockdown measures on air pollutants from seasonal variability. *Atmospheric Pollution Research*, *12*(2), 84–92. https://doi.org/10.1016/j.apr.2020.10.011

Hörmann, S., Pfeiler, B., & Stadlober, E. (2005). Analysis and Prediction of Particulate Matter PM10 for the Winter Season in Graz. *Austrian Journal of Statistics*, *34*(4), 307. https://doi.org/10.17713/ajs.v34i4.420

Stadlober, E., Hörmann, S., & Pfeiler, B. (2008). Quality and performance of a PM10 daily forecasting model. *Atmospheric Environment*, *42*(6), 1098–1109. https://doi.org/https://doi.org/10.1016/j.atmosenv.2007.10.073