

DSML Project Paper

Semesterarbeit DSML - SS 2022

Studenten	Iris Lüthi, Maja Velkova, Yannik Zimmermann
Dozenten	Prof. Dr. M. Krebs, Prof. Dr. P. Collova
Modul	DSML
Semester	6. Semester
Datum	10. Juni 2022

Inhaltsverzeichnis

1 Einleitung	3
2 Data Source, Datasets und Code	3
2.1 Dataset “Daily Temperature of Major Cities”	3
2.2 Dataset “Global Air Quality Index (AQI)”	4
2.3 Dataset “US Pollution”	5
2.4 Bedenken über die Datasets	5
2.5 Aufbau und Code der Arbeit	5
3 Analyse	6
3.1 Temperaturverlauf von 1995 - 2020	6
3.2 Luftqualität und Verteilung der Treibhausgase der USA	9
3.3 Vorhersagen zur Temperatur & Luftqualität	12
3.3.1 Vorhersage der Temperaturentwicklung	12
3.3.2 Vorhersage der Entwicklung des AQI's	12
3.4 Zusammenhänge zwischen Temperatur & Treibhausgasen	14
4 Resultate und Diskussion	16
5 Fazit	16
Literaturverzeichnis	17
Abbildungsverzeichnis	17
Selbstständigkeitserklärung	18

1 Einleitung

Im Rahmen des Moduls Data Science and Machine Learning (DSML) werden die Studierenden der Berner Fachhochschule dazu angehalten eine empirische Arbeit auf dem Gebiet der Datenanalyse oder des maschinellen Lernens unter Verwendung von R oder Python zu erarbeiten.

Das Autorenteam hat sich dazu entschieden, die Entwicklung der globalen Lufttemperaturen, des AQI (Air Quality Index) und der möglichen Einflüsse durch Treibhausgase zu untersuchen. Hierfür wurden drei Datasets ausgewählt, welche eine solche Analyse ermöglichen sollen. Zuerst wurden die weltweiten Veränderungen der Temperaturen analysiert. Darauf wurde eine fokussierte Analyse der Temperatur, Luftqualität und der anteiligen Zusammensetzung der Treibhausgase basierend auf amerikanischen Messungen durchgeführt. Auf die Analyse der Datasets folgt die Anwendung von ML-Algorithmen auf die Daten, um eine Vorhersage für die Luftqualität und die Temperaturentwicklung zu treffen. Ziel der Analyse und der Vorhersage ist es, anhand von Daten feststellen zu können, ob die Aussagen der Wissenschaft und den Medien in Bezug auf den Klimawandel durch Studenten nachgewiesen werden kann. Schlussendlich werden die Ergebnisse vom Autorenteam zusammengefasst und diskutiert, woraufhin ein Fazit folgt.

2 Data Source, Datasets und Code

Alle Datasets, die im Rahmen dieser Semesterarbeit genutzt werden, wurden auf der Plattform Kaggle veröffentlicht. Die Plattform bietet über 50'000 öffentliche Datasets. Diese können dabei direkt in Online-Notebooks analysiert werden. Kaggle bietet nebenbei einige Kurse zur Datenanalyse mithilfe verschiedenster Tools und Programmiersprachen. Im Anschluss an die Aufschlüsselung der verwendeten Datasets wird der Codeaufbau dieser Arbeit erläutert.

2.1 Dataset “Daily Temperature of Major Cities”

Das “Daily Temperature of Major Cities” Dataset besteht aus acht Spalten und 2,9 Millionen Zeilen. Es wurde zuletzt vor etwa zwei Jahren aktualisiert. In der originalen Datenstruktur sind folgende Spalten enthalten:

- Region
- Country
- State
- City
- Day
- Month
- Year
- AvgTemperature

Mithilfe des folgenden Befehls lässt sich herausfinden, wie viele Zeilen identisch sind. In diesem Dataset sind insgesamt 20'715 identisch und müssen daher vor der weiteren Verarbeitung gefiltert werden.

```
sum(duplicated(raw_city_temps))
```

Bei der Untersuchung des Datasets stellte sich heraus, dass sich einige der Temperaturwerte unterhalb von -90° Fahrenheit befinden. Aufgrund der unrealistischen Werte werden diese Zeilen für die weitere Analyse entfernt. Des Weiteren befinden sich mögliche Tippfehler in der Datenerhebung, da Temperaturwerte für die Jahre 200 und 201 erfasst wurden. Diese Werte werden ebenfalls gefiltert. Zuletzt wurden die Werte aus dem Jahr 2020 entfernt, da die Datenerfassung für dieses Jahr nicht abgeschlossen wurde. Die Bereinigung wurde mithilfe des folgenden Befehls vorgenommen:

```
# Data preprocessing for further analysis
city_temps = raw_city_temps %>%
  # remove duplicates
  distinct() %>%
  # remove temp values below -50f because they seem like default/null values
  filter(AvgTemperature > -50) %>%
  # remove values where year is below 1950 because those are probably typos (200, etc)
  filter(Year > 1950) %>%
  # remove year 2020 because of really low number of observations
  filter(Year != 2020)
```

Um die Temperaturwerte einfach interpretieren zu können, wurden diese in das Format Grad-Celsius transformiert. Für die Umwandlung wurde die Library «weathermetrics» verwendet. Diese bietet den Befehl «fahrenheit.to.celsius», welcher wie folgt angewendet wurde.

```
city_temps['AvgTemperatureInCelcius'] = fahrenheit.to.celsius(city_temps$AvgTemperature)
```

In der abschliessenden Struktur wurde somit lediglich die Spalte «AvgTemperatureInCelcius» ergänzt.

2.2 Dataset “Global Air Quality Index (AQI)”

Das “Global Air Quality Index (AQI)” Dataset besteht aus täglichen Messungen des AQI aus verschiedenen Städten der USA. Es besteht aus insgesamt acht Spalten und über zehn Millionen Zeilen und ist damit das grösste Dataset dieser Semesterarbeit. Das Dataset wurde zuletzt im Oktober 2021 aktualisiert. Folgende Spalten sind im Dataset enthalten:

- State Name
- Date
- AQI
- Category
- Defining Parameter
- Latitude
- Longitude
- Country Name

Auch dieses Dataset wurde auf sich wiederholende Werte untersucht. Es wurden jedoch nur einzigartige Messungen festgestellt. Mithilfe des «unique» Befehls wurden die möglichen Werte für die Kategorie und den definierenden Parameter festgestellt. Der AQI wird in die Kategorien «Good», «Moderate», «Unhealthy for Sensitive Groups», «Unhealthy», «Very Unhealthy» und Hazardous eingeteilt. Die definierenden Parameter bestehen aus «Ozone», «PM2.5», «PM10», «NO2», «CO» und «SO2».

```
unique(raw_air_quality_index$`Defining Parameter`)
unique(raw_air_quality_index$Category)
```

Aufgrund der guten Datenqualität wurden lediglich die Messwerte des Jahres 2021 entfernt, da diese unvollständig sind. Somit hat sich die Struktur des Datasets für die Analyse nicht verändert. Insgesamt wurden 40'283 von 10'158'517 Zeilen entfernt.

```
air_quality_index = raw_air_quality_index %>%
  filter(year(Date) != 2021)
```

2.3 Dataset “US Pollution”

Das “US Pollution” Dataset besteht aus Messungen von Treibhausgas-Anteilen in den USA. Es besteht aus 24 Spalten und über 600'000 Zeilen. Das Dataset wurde zuletzt vor 6 Jahren aktualisiert. Nachfolgend sind die relevanten Spalten aufgelistet. Einige der Spalten wurden aufgrund der Leserlichkeit nicht aufgelistet. Diese bestehen aus den Minimal- und Maximalwerten der Treibhausgase gemessen pro Stunde.

- Date
- Year
- Month
- Day
- State
- Country
- City
- O₃ Mean
- NO₂ Mean
- CO Mean
- SO₂ Mean

Im «US Pollution» Dataset wurden ebenfalls 1479 identische Messungen festgestellt. Zudem wurde die Datenerfassung für das Jahr 2021 nicht abgeschlossen. Das finale Dataset unterscheidet sich strukturell nicht vom originalen Dataset. Diese Werte wurden darauffolgend mit folgendem Befehl entfernt:

```
pollution_data = raw_pollution_data %>%
  # remove duplicates
  distinct() %>%
  # remove measurements from 2021
  filter(year(Date) != 2021)
```

2.4 Bedenken über die Datasets

Die Datasets wurden auf Kaggle nur sehr minimal beschrieben. Dementsprechend finden sich kaum Infos darüber, wie die Daten erhoben wurden. Dadurch fehlen dem Autorenteam Informationen über die Methodik der Datenerfassung. Da die Datenerfassung über einen längeren Zeitraum durchgeführt wurde, besteht auch die Möglichkeit, dass sich die Messmethode geändert hat. Durch die fehlenden Informationen über die Methodik können daher keine Aussage über mögliche Messfehler oder Messveränderungen getroffen werden.

Bei den Daten aus den «US Pollution» Dataset handelt es sich bereits um ein verarbeitetes Dataset, bei welchem die originalen Messwerte auf Tagesdurchschnitte umgerechnet wurden. Die nachfolgende Analyse nutzt teilweise Durchschnittswerte, wodurch die Daten weiter vereinfacht werden. Mögliche Trends könnten dadurch verzerrt wahrgenommen werden.

Laut der Beschreibung der Datasets wurden diese entweder durch die Universität von Dayton oder durch die EPA (Environmental Protection Agency) zur Verfügung gestellt. Welche Transformationen oder Veränderungen der originalen Daten vorgenommen wurde, ist nicht dokumentiert und öffnet die Tür für potenzielle Verfälschungen.

2.5 Aufbau und Code der Arbeit

Die Semesterarbeit wurde mithilfe eines RMarkdown-Files erstellt. Dies bietet den Vorteil, dass der Code und das resultierende Dokument in einem File erfasst werden können. Das Grundsetup für das RMarkdown-File wurde aufgrund der Lesbarkeit in ein eigenes Setup File verlagert. Für die initiale Analyse wurde ein eigenes

File namens «DSML_SS22_Luethi-Velkova-Zimmermann.R» erstellt, in welchem die Datasets durch einen iterativen Prozess analysiert und transformiert wurden. So konnten Versuche und Codebausteine, welche nicht im finalen Dokument genutzt wurden, trotzdem erhalten bleiben. Das R-Projekt besteht somit aus nachfolgender Filestruktur:

- DSML_SS22_Luethi-Velkova-Zimmermann.Rmd
- DSML_SS22_Luethi-Velkova-Zimmermann.R
- DSML_SS22_Luethi-Velkova-Zimmermann_Setup.R
- DSML_SS22_Luethi-Velkova-Zimmermann.Rproj
- DSML_SS22_Luethi-Velkova-Zimmermann.pdf

3 Analyse

In diesem Kapitel werden die Untersuchungen und Resultate aus den Datensätzen dargestellt. Wie in der Einleitung erwähnt, wird zuerst die Temperaturrentwicklung, weiter die Luftqualität, allfällige Prognosen und zum Schluss einen möglichen Zusammenhang zwischen den beiden Aspekten untersucht.

3.1 Temperaturverlauf von 1995 - 2020

Alle folgenden Grafiken wurden mithilfe der Library “ggplot2” generiert. Um eine einfache Lesbarkeit zu gewährleisten wird der Code nur von der folgenden Grafik im Dokument angezeigt. Alle weiteren Grafiken wurden mit dem selben Prinzip erstellt.

```
city_temps %>%
  # Apply plot based filters
  group_by(Year) %>%
  # Summarize certain values
  summarize_at(vars(AvgTemperatureInCelsius), list(AvgTemp=mean)) %>%
  # Generate ggplot
  ggplot(aes(Year,AvgTemp)) +
  geom_smooth(method = 'loess', formula = 'y ~ x') +
  geom_line() +
  # Define custom descriptions for x, y and title
  labs(
    x="Jahr",
    y="Durchschnittstemperatur in Celsius",
    title = "Gemessene Durchschnittstemperatur in Grossstädten rund um die Erde"
  ) +
  theme(aspect.ratio = 0.7)
```

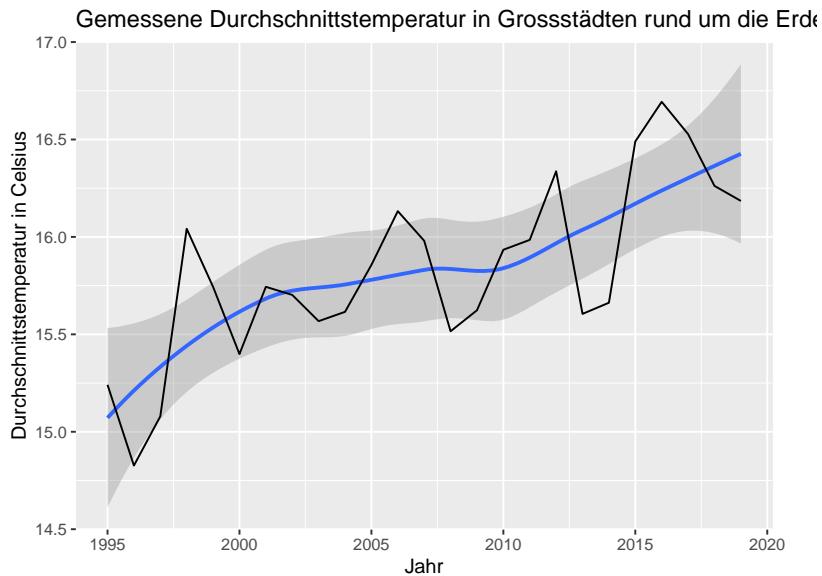


Abbildung 1: Temperatur Entwicklung von 1995 - 2020

Die erste Grafik wurde mit den veröffentlichten Temperaturwerten aus verschiedenen Grossstädten aus dem Dataset «Daily Temperature of Major Cities» generiert. Darauf ist die Entwicklung der weltweiten Durchschnittstemperaturen zwischen den Jahren 1995 bis 2020 ersichtlich. Es ist ein klar aufsteigender Trend zu erkennen. Jedoch ist zu beachten, dass die y – Achse eine Skala zwischen 14.5 und 17.0 Grad Celsius aufweist. Daher wird die Kurve drastischer angezeigt als bei einer grösseren Skala. Hätte die y-Achse beispielweise eine Skala zwischen 0 und 30 Grad, würde der Unterschied optisch geringer ausfallen. Die schwarze Linie zeigt die weltweiten Durchschnittstemperaturen der Jahre 1995-2020 an. Anhand dieser, ist zu erkennen, dass nicht jedes Jahr pauschal wärmer ist als das vorherige, sondern dass es Schwankungen zwischen den verschiedenen Jahren gibt. Die blaue Linie stellt den Trend der Temperaturänderungen dar. Obwohl also Schwankungen zwischen den Jahren zu sehen sind, zeigt der Trend eine klare Temperaturzunahme auf. Bei dieser Grafik ist wichtig mitzunehmen, dass die durchschnittliche weltweite Temperatur in 25 Jahren um ca. ein Grad Celsius erhöht hat. Da es sich bei der ersten Grafik, um den weltweiten Durchschnitt handelt, hat sich das Autorenteam die Frage gestellt, ob alle Kontinente den gleichen Trend aufweisen.

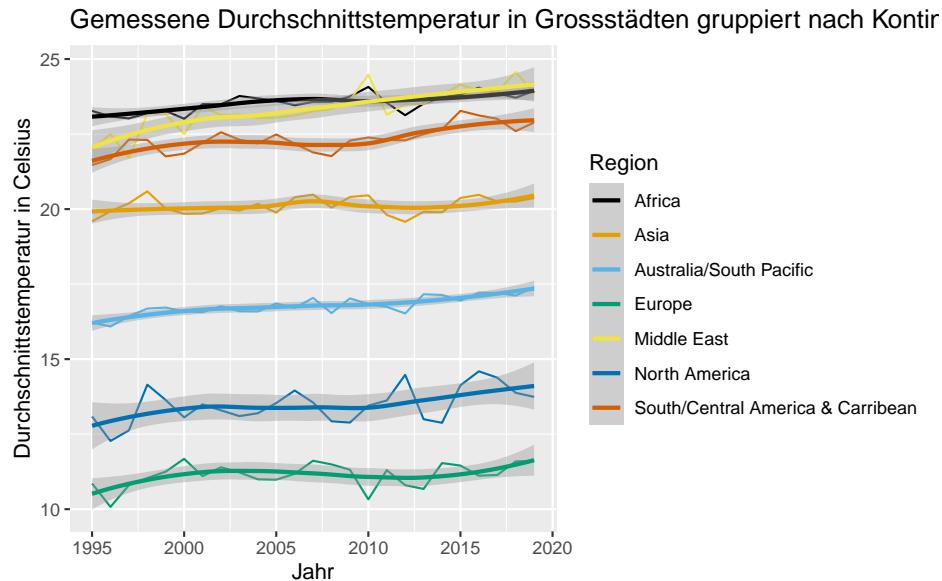


Abbildung 2: Durchschnittstemperaturen pro Kontinent bzw. Region

In der zweiten Grafik werden die durchschnittlichen Temperaturänderungen auf die Kontinente aufgeschlüsselt. Der allgemeine Trend ist, wie bereits in der ersten Grafik, auch hier zunehmend. Interessant zu sehen ist, dass nicht alle Kontinente beziehungsweise Regionen gleich stark von der Erwärmung betroffen sind. Die dunkelblaue Trendlinie von Nord Amerika ist stärker geneigt als die orangefarbene Linie der asiatischen Region. Auch wurde die Region Nord Amerika grösseren Temperaturschwankungen über die Jahre ausgestellt als die asiatischen Regionen. Der Nahe Osten ist deutlich am stärksten von den Temperaturerwärmungen betroffen. Um den Umfang der Analyse weiter einzuschränken, hat sich das Autorenteam im weiterführenden Teil dieser Arbeit auf den Amerikanischen beziehungsweise Nordamerikanischen Kontinent konzentriert.

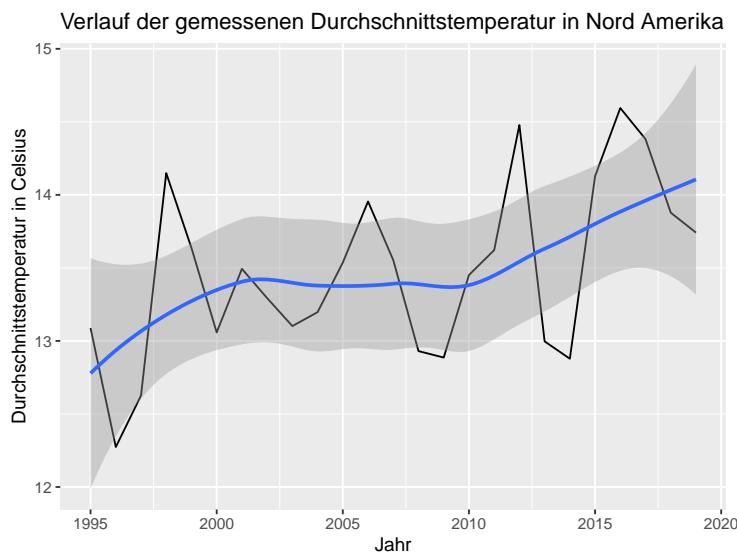


Abbildung 3: Durchschnittstemperaturen von Nordamerika

In der oben aufgeführten Grafik ist die Durchschnittstemperatur der Nordamerikanischen Staaten isoliert ersichtlich. Die Temperaturskala wurde verkleinert und die Zeitskala wurde analog den vorherigen Grafiken beibehalten. Hier ist nochmals ersichtlich, wie stark die Temperaturschwankungen in der oberen Hälfte des amerikanischen Kontinents ausfallen. Zudem ist auffallend, dass wie bei den weltweiten Temperaturen, die Durchschnittstemperaturen in Nordamerika um mindestens ein Grad Celsius gestiegen sind.

Im nächsten Kapitel wird der Umfang wiederum auf die USA reduziert. Für diese sind Datasets für die Luftqualität und Verteilung der Treibhausgase verfügbar. Das Ziel des Autorenteam ist es mit Hilfe dieser Datasets herauszufinden, ob in den Daten der Luftqualität und der Verteilung der Treibhausgase eine mögliche Ursache für die stetige Temperaturerwärmung zu finden ist.

3.2 Luftqualität und Verteilung der Treibhausgase der USA

In einem ersten Schritt wird in diesem Kapitel die allgemeine Luftqualität in ganz Amerika aufgezeigt. Weiter wird die durchschnittliche Luftqualität der USA indexiert dargestellt. Der Air Quality Index (AQI) dient der U.S. amerikanischen Regierung, um die gemessene Luftqualität zu kategorisieren. In der folgenden Abbildung wird die Indexierung detailliert erläutert (airnow.gov, k.A.).

AQI Basics for Ozone and Particle Pollution			
Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Abbildung 4: Air Quality Index der U.S. Environmental Protection Agency

Daraus ist zu entnehmen, dass höhere Werte des Indexes eine entsprechend schlechtere Luftqualität darstellen. Die gleichen Kategorien sind im untenstehenden Graph in Bezug auf die Beschriftung wieder zu erkennen.

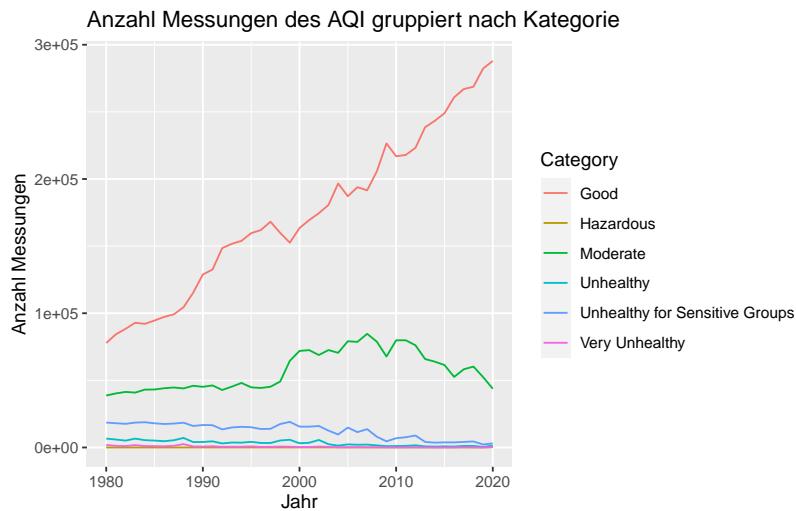


Abbildung 5: Luftqualitätsmessungen aufgeteilt in Kategorien

Zu den Stufen «Very Unhealthy», «Unhealthy», «Unhealthy for Sensitive Groups» und «Hazardous» gibt es über die aufgezeigten 40 Jahre wenig bis kaum Messungen. Diese sind zudem abnehmend. Die Stufe «Moderate» kommt nach einer kurzen Schwankung wieder auf ihren Ausgangswert zurück. Im Gegenteil steigt die Kategorie «Good» nahezu konstant an. Wichtig zu beachten ist es, dass die Anzahl der Messpunkte über die Jahre zugenommen haben. Trotz dieser Zunahme zeigt der Graph eine klare Verbesserung der Luftqualität. In der nächsten Grafik wird der durchschnittliche AQI visuell dargestellt.

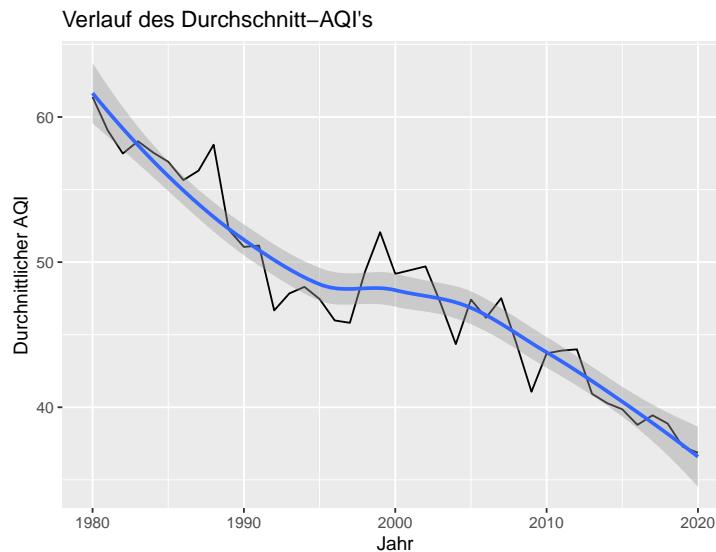


Abbildung 6: Durchschnitt des Luftqualität Index

Trotz starken Schwankungen in den Messungen des AQI, ist eine klare und konstante Abwärtsbewegung festzustellen. Der durchschnittliche AQI hat sich demnach zwischen 1980 und 2020 von ca. 60 auf ca. 40 abgesenkt. Aufgrund der Grösse der USA kann vermutet werden, dass die Durchschnittswerte des AQI nicht landesweit gleich sind. Aufgrund dessen werden die oben dargestellten Durchschnittswerte als nächstes auf die verschiedenen Staaten der USA verteilt dargestellt.

AQI Durchschnittswerte aufgeteilt in amerikanische Staaten

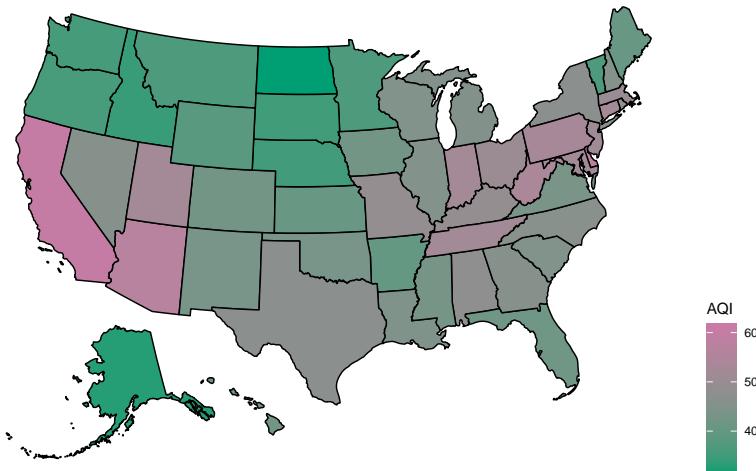


Abbildung 7: AQI Luftqualitätsskala der USA

Wie unter der Abbildung vier erwähnt, zeigen tiefe AQI Werte eine bessere Luftqualität auf als hohe Werte. In der aufgeführten Karte der USA ist sichtbar, dass Kalifornien und Arizona bei der Luftqualität am schlechtesten abschneiden. Tendenziell weisen die nordischen Staaten einen besseren AQI Wert auf als die südlichen. Dabei kann beachtet werden, dass Kalifornien, der am stärksten bevölkerte Staat der USA ist. Dies im Gegenteil zu Arizona, welcher weit weniger Menschen beherbergt. Möglichweise könnte hier eine Korrelation bestehen. Um die Erkenntnisse aus der Karte zu verdeutlichen, werden die AQI von zwei Staaten individuell angeschaut und verglichen.

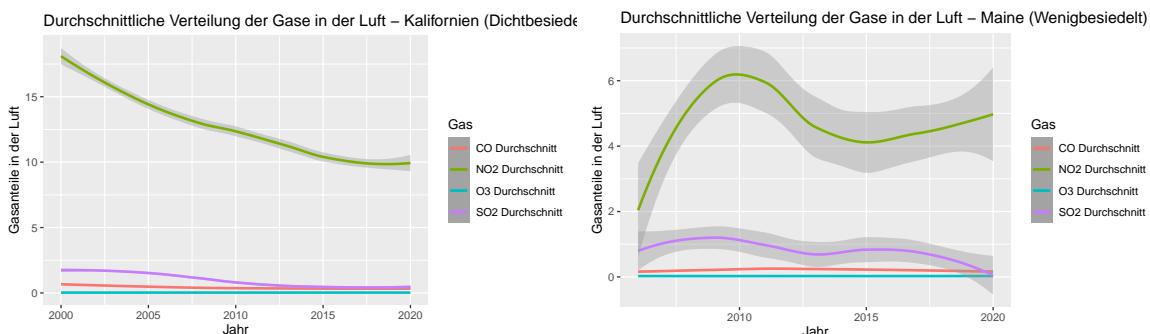


Abbildung 8: Gasanteile in der Luft: Kalifornien vs Maine

In der linken Grafik ist die durchschnittliche prozentuale Verteilung einer Auswahl an Treibhausgasen vom kalifornischen Staat ersichtlich. Auf den ersten Blick, kann festgestellt werden, dass die anteiligen Stickstoffdioxidwerte (NO₂) in den letzten zwanzig Jahren abgenommen haben. Der Anteil an Schwefeldioxid (SO₂) und Kohlenstoffmonoxid (CO) hat, zwar weniger deutlich, aber dennoch abgenommen. Die Ozonwerte (O₃) bleiben auf oder knapp über Null eine Konstante. Trotz dem allgemeinen Abwärtstrend werden durch die hohen Anteile an Treibhausgasen in der Luft, die Erkenntnisse der vorherigen Abbildung bestätigt. Kalifornien hat entsprechend sehr hohe Anteile an Treibhausgase in der Luft.

In der rechten Grafik sind die Werte eines weniger dicht besiedelten Staates zu sehen. In Maine wurde der Anteil der Gase von sieben Prozent im angezeigten Zeitraum nie überschritten. Da die Skala auf der Y-Achse

deutlich kleiner ist, sind auch stärkere Schwankungen beim Stickstoff- und dem Schwefeldioxid festzustellen. Im Allgemeinen konnte jedoch festgestellt werden, dass Maine einen tieferen Anteil an Treibhausgasen in der Luft aufweist.

3.3 Vorhersagen zur Temperatur & Luftqualität

Die analysierten Daten zeigen, dass der AQI abnimmt und die Temperatur zunimmt. Aufgrund dessen wird in folgendem Abschnitt mithilfe von Machine Learning Algorithmen und der Library «forecast» versucht eine Vorhersage der Temperatur und Luftqualität über die nächsten zehn Jahre zu treffen. Die Vorhersagen werden mit einem «Confident-Level» von 95 durchgeführt. Als Datengrundlage werden alle Daten aus dem «Daily Temperature of Major Cities» und dem «Global Air Quality Index (AQI)» Dataset genutzt. Die Library «forecast» wird aufgrund der einfachen Handhabung genutzt. Mithilfe einer sog. «Time-Series», einem speziellen Format, und unter Anwendung des «ARIMA (Auto-Regressive Integrated Moving Average)»-Modells werden die Daten auf den Trend, die saisonalen Veränderungen und die übrigen Messwerte aufgeteilt und analysiert.

3.3.1 Vorhersage der Temperaturentwicklung

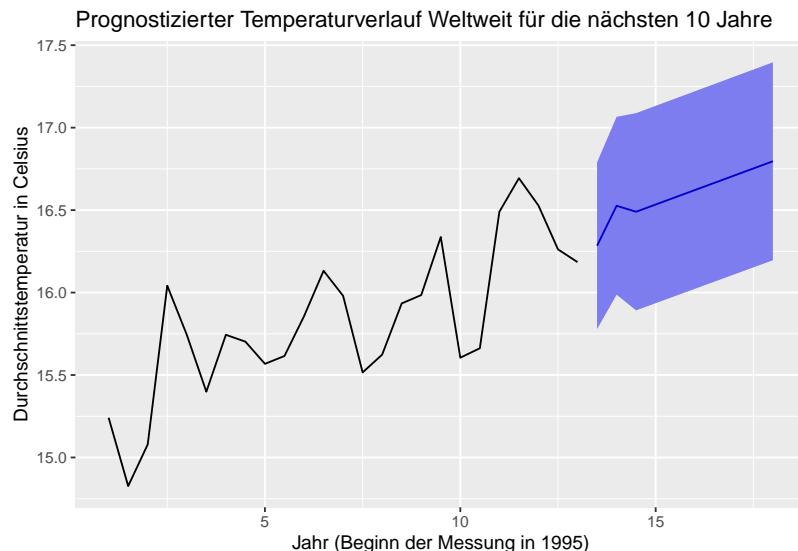


Abbildung 9: Prognostizierter Temperaturverlauf Weltweit für die nächsten 10 Jahren

Die obenstehende Grafik zeigt den prognostizierten Temperaturverlauf für die nächsten 10 Jahre. Dabei ist zu erkennen, dass die Temperatur um etwa ein weiteres Grad steigen könnte. Mit dem «predict» Befehl lassen sich unter anderem die Werte auslesen. Folgende Temperaturentwicklung wird somit für die nächsten 10 Jahre angegeben:

```
round(predict(city_temps_forecast_year)$mean[1:10], digits=1)  
## [1] 16.3 16.5 16.5 16.5 16.6 16.6 16.7 16.7 16.8 16.8
```

3.3.2 Vorhersage der Entwicklung des AQI's

Die nachfolgenden Grafiken zeigen die durch «forecast» erstellten Trends, saisonale Schwankungen und übrige Messabweichungen des «Global Air Quality Index (AQI)» Datasets. In der Trend Spalte der Grafik ist ein

klarer Abwärtstrend zu erkennen. Laut diesem Trend hat sich der durchschnittliche AQI Wert von 60 auf 40 minimiert. Auch in der saisonalen Darstellung ist zu erkennen, dass der AQI im Sommer allgemein höher ist als im Winter, dieser aber durchschnittlich mit jedem vergangenen Jahr abnimmt.

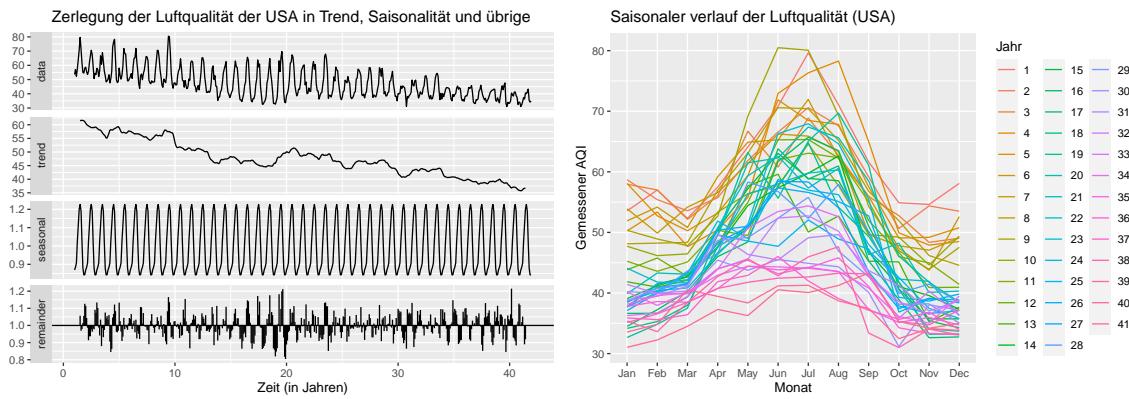


Abbildung 10: Zerlegung des Temperaturverlaufs der USA in Trend, Saisonalität und übrige Messabweichungen

Auf der Grafik 10 ist die Vorhersage des AQIs der nächsten zehn Jahre abgebildet. Folgende Befehle wurden genutzt, um das entsprechende Modell und die Grafik zu erstellen:

```
autoplot(aqi_forecast) +
  labs(
    x="Zeit (in Jahren)",
    y="Gemessener AQI",
    title = "Vorhersage der Luftqualität (USA) in den nächsten 10 Jahren"
  )
```

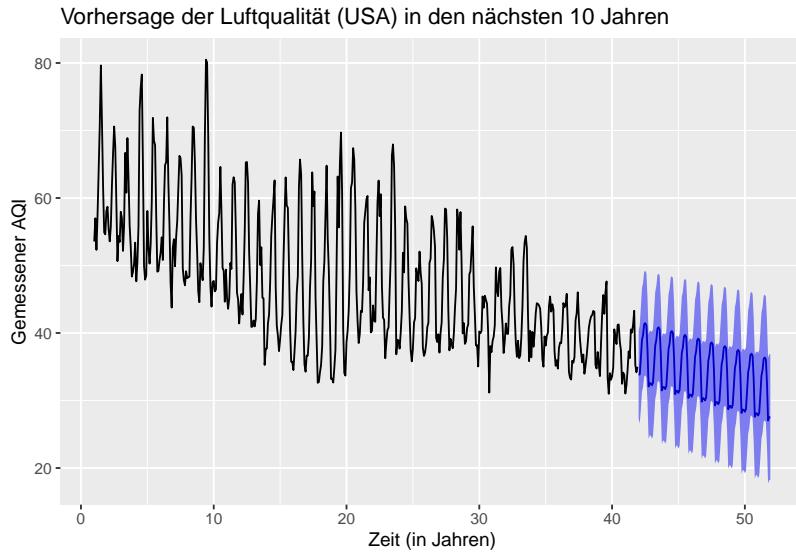


Abbildung 11: Vorhersage der Luftqualität (USA) in den nächsten 10 Jahren

Die Vorhersage beschreibt somit, dass sich der AQI Wert in den nächsten Jahren vermutlich um ca. 10 verkleinern wird. Durch den «predict» Befehl lassen sich auch für dieses Modell die durchschnittlich vorhergesagten

Werte des AQI's auslesen.

```
round(predict(aqi_forecast)$mean, digits = 1)
```

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 42 33.8 33.9 36.2 38.9 39.7 41.3 41.5 41.1 37.6 32.1 32.6 32.6
## 43 32.2 32.7 35.3 38.1 39.0 40.7 40.8 40.5 37.0 31.5 32.0 32.0
## 44 31.7 32.1 34.7 37.6 38.4 40.1 40.3 40.0 36.4 30.9 31.4 31.4
## 45 31.1 31.5 34.1 37.0 37.9 39.5 39.7 39.4 35.9 30.4 30.9 30.9
## 46 30.6 31.0 33.6 36.5 37.3 39.0 39.2 38.9 35.3 29.8 30.3 30.3
## 47 30.0 30.4 33.0 35.9 36.8 38.4 38.6 38.3 34.8 29.3 29.8 29.8
## 48 29.5 29.9 32.5 35.3 36.2 37.9 38.1 37.8 34.2 28.7 29.2 29.2
## 49 28.9 29.3 31.9 34.8 35.6 37.3 37.5 37.2 33.7 28.1 28.6 28.6
## 50 28.3 28.7 31.4 34.2 35.1 36.7 36.9 36.6 33.1 27.6 28.1 28.1
## 51 27.8 28.2 30.8 33.7 34.5 36.2 36.4 36.1 32.5 27.0 27.5 27.5
```

3.4 Zusammenhänge zwischen Temperatur & Treibhausgasen

In diesem Kapitel soll nun überprüft werden, ob eine Korrelation zwischen der Lufttemperatur und den Anteilen an Treibhausgasen erkennbar ist. Hierfür wurde ein neues Dataframe aus den Datasets «Daily Temperature of Major Cities» nach Kalifornien gefiltert und «Global Air Quality Index (AQI)» mittels inner Join auf die Datums-Variable erstellt. Die anteiligen Treibhausgas-Werte wurden hierfür auf Tagesbasis als Durchschnitt gerechnet. Durch den Befehl «pairs» konnte die nachfolgende Grafik erstellt werden.

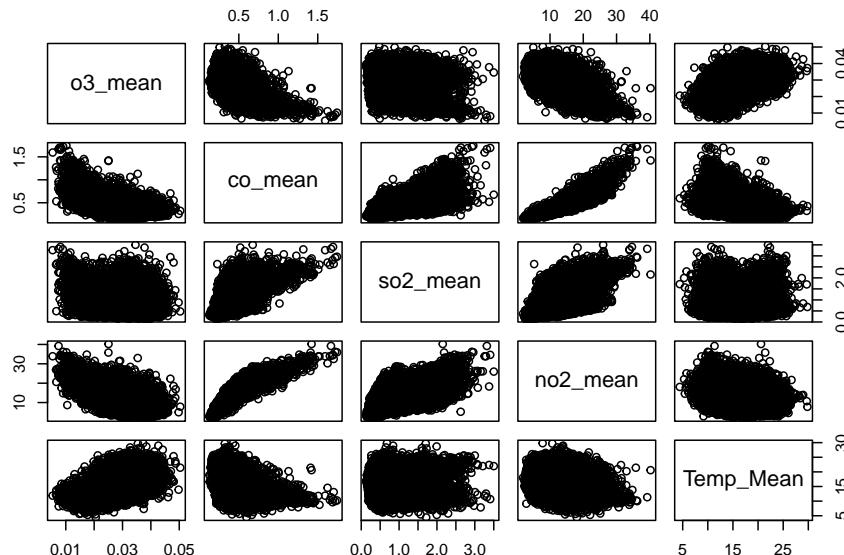


Abbildung 12: Zusammenhänge zwischen Temperatur und Treibhausgasen in Kalifornien

In den oberen 20 Scatterplots werden jeweils die Relationen zwischen den Durchschnittswerten der Treibhausgasen und der Temperatur in Kalifornien aufgezeigt. Zwischen dem Durchschnitt vom Kohlenstoffmonoxid (CO) und dem Stickstoffdioxid (NO₂) ist eine positive lineare Korrelation zu erkennen. Einen Zusammenhang zwischen den Gasen und der Temperatur hat das Autorenteam nicht erkennen können. Die nächste Grafik zeigt die Treibhausgas-Werte in Relation zu den Temperaturwerten und bestätigt diese Erkenntnis.

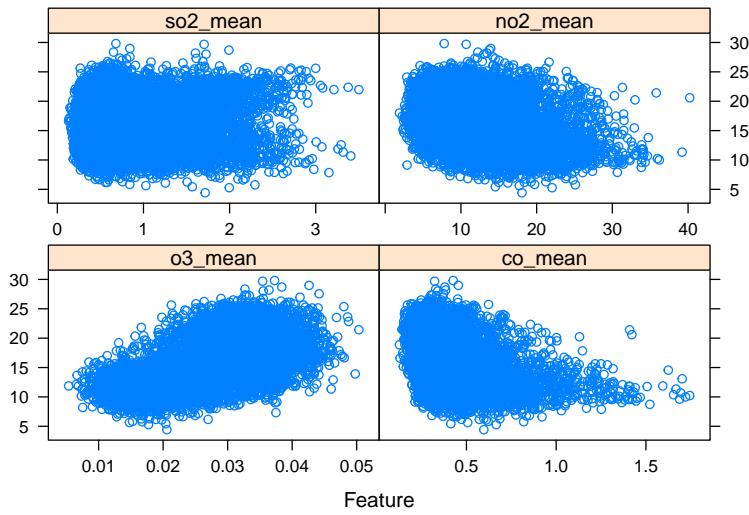


Abbildung 13: Zusammenhang zwischen den Durchschnittstemperaturen und den vier Gasen

Die obenstehende Grafik zeigt nun nur noch die Relation der vier Treibhausgase in Bezug zur kalifornischen Durchschnittstemperatur. Aus den aggregierten Daten wurde folgendes Model erstellt:

```
summary(temperature_and_pollution_model)

##
## Call:
## lm(formula = Temp_Mean ~ ., data = new_model_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.3084  -2.4882   0.1733   2.5448  12.9088 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 12.20012   0.28745  42.442 <2e-16 ***
## o3_mean     200.34541   7.01218 28.571 <2e-16 ***
## co_mean    -12.93372   0.46648 -27.726 <2e-16 ***
## so2_mean     2.27582   0.09349 24.343 <2e-16 ***
## no2_mean     0.15401   0.01636  9.412 <2e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.471 on 7292 degrees of freedom
## Multiple R-squared:  0.3953, Adjusted R-squared:  0.395 
## F-statistic: 1192 on 4 and 7292 DF,  p-value: < 2.2e-16
```

Obwohl das Model signifikant ist, kann in keinem der Scatterplots eine klare Korrelation erkannt werden. Das Model zeigt grundsätzlich einen guten P-Value(fit), da dieser unterhalb von 5% liegt, jedoch sagt der R-Squared Wert 0.39 aus, dass die Variablen eher nicht für eine befriedigende Vorhersage genutzt werden können.

4 Resultate und Diskussion

Im Kapitel 3.1 hat das Autorenteam festgestellt, dass trotz teils hohen Temperaturschwankungen zwischen den Kontinenten und den Jahren ein klar aufsteigender Trend zu erkennen ist. Es werden also weltweit zunehmend wärmere Temperaturen gemessen. Durch diese Analyse ist die Frage aufgekommen, ob alle Kontinente den gleichen Trend aufweisen. Das Autorenteam konnte feststellen, dass der Trend alle Regionen, wenn auch nicht überall gleich intensiv, betrifft. Der Nahe Osten zeigt dabei die stärkste Entwicklung der Temperaturerwärmungen auf. Um die Resultate der Temperaturerwärmung mit der Luftqualität in Zusammenhang zu setzen, hat sich das Autorenteam jedoch auf den amerikanischen Kontinent beziehungsweise auf die USA fokussiert.

Im Kapitel 3.2 hat das Autorenteam in erster Linie festgestellt, dass über die letzten Jahre im Allgemeinen immer mehr Messungen durchgeführt wurden. Die Vermutung besteht, dass diese Entwicklung auf die verfügbaren technischen Mittel zurückgeführt werden kann.

In einem weiteren Schritt wurde die Luftqualität nach dem AQI aufgeschlüsselt. Damit konnte das Autorenteam Vergleiche zwischen den Anteilen an verschiedenen Treibhausgasen in den Staaten der USA ziehen. Im Allgemeinen zeigen die Ergebnisse auf, dass sich die Luftqualität langsam verbessert. Der durchschnittliche AQI ist in stark besiedelten Staaten deutlich schlechter als in weniger besiedelten Staaten. Um dies zu bestätigen, wurden die Staaten Kalifornien und Maine direkt verglichen. Im Code wurden diverse andere Staaten verglichen, die immer ähnliche Ergebnisse aufgezeigt haben.

Schlussendlich hat die Autorengruppe versucht einige Vorhersagen zu erarbeiten. Zuerst sollte herausgefunden werden, wie sich die weltweiten Temperaturen entwickeln werden. Das Modell hat aufgezeigt, dass sich die Weltbevölkerung effektiv auf wärmere Zeiten gefasst machen kann. Weiter wollte das Autorenteam wissen, ob sich die Luftqualität weiterhin verbessern wird. Auch da hat das Modell bestätigt, dass der Trend weiterhin positiv sein wird. Im letzten Kapitel der Analyse haben die Verfasserinnen und der Verfasser dieser Arbeit versucht einen Zusammenhang zwischen der globalen Temperaturerhöhung und den Anteilen an Treibhausgasen in der Luft zu finden. Obwohl in der Wissenschaft belegt wird, dass die Temperaturerhöhung in der terrestrischen Atmosphäre durch die Zunahmen an schädlichen Treibhausgase gefördert wird, konnte bei den Analysen in dieser Arbeit kein Zusammenhang festgestellt werden. Das Modell war allgemein schwierig zu interpretieren. Das Autorenteam hatte angenommen, dass eine offensichtliche Korrelation ersichtlich wäre.

5 Fazit

In der vorliegenden Arbeit hat das Autorenteam anhand der Datenanalyse herausgefunden, dass die Temperatur in den nächsten Jahrzehnten stetig zunehmen wird. Zur gleichen Zeit wird, gemäß dem Dataset über den AQI, die Luftqualität in den ganzen USA besser. Es ist jedoch wissenschaftlich erwiesen, dass die schädlichen Treibhausgase diesem aufwärts Trend erheblich beisteuern. Das Autorenteam ist sich nicht sicher, warum diese Ergebnisse so ausfallen. Der Fehler könnte sowohl bei der Auswahl der Daten, im Erstellungsprozess der Grafen oder bei deren Interpretation liegen. Da nur sehr spärliche Informationen über die Datasets mitgeliefert wurden, könnte da schon die erste Fehlerquelle liegen. Mit der stetigen Zunahmen an Messungen über die letzten Jahren, wäre es sicherlich spannend diese Arbeit in 10 bis 20 Jahren noch einmal durchzuführen, um weitere Erkenntnisse zu erarbeiten.

Das Autorenteam hat festgestellt, dass es schwierig war, die Arbeit einzugrenzen und nicht jedes interessante Thema zu verfolgen. Es wäre sicherlich spannend eine tiefere Analyse bei der Luftzusammensetzung durchzuführen, um zu erfahren, welche weitere Erderwärmungsfaktoren vorhanden sind. Weiter könnte ein Dataset mit den Wetterdaten ausgewertet werden, um herauszufinden, ob ein Zusammenhang zwischen den Temperaturen und dem Wetter besteht. Die Gegenüberstellung der «Definierenden Parameter» aus dem AQI Dataset und den Werten des Pollution Datasets hätte bestimmt auch interessante Resultate geliefert.

Diese Semesterarbeit hat der Autorengruppe noch stärker verdeutlicht, dass sich alle mit dem Klimawandel beschäftigen sollten, da die globale Erderwärmung mit einfachen Datasets erwiesen werden kann. Dennoch

kennt sich die Autorengruppe in diesem Bereich zu wenig aus. Bestimmt wäre es wichtig, noch andere Daten in Betracht zu ziehen, um qualifizierte Aussagen zu treffen. Dies hätte den Umfang dieser Arbeit jedoch gesprengt.

Literaturverzeichnis

- Daily Temperature of Major Cities.* (k.A.). Abgerufen am 03. April 2022, von kaggle.com: <https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities>
- 1980-2021 Daily Air Quality Index from the EPA.* (k.A.). Abgerufen am 03. April 2022, von kaggle.com: <https://www.kaggle.com/datasets/threnjen/40-years-of-air-quality-index-from-the-epa-daily>
- U.S. Pollution Data.* (k.A.). Abgerufen am 03. April 2022, von kaggle.com: <https://www.kaggle.com/datasets/sogun3/uspollution>
- Air Quality Index (AQI) Basics.* (k.A.). Abgerufen am 09. Juni 2022, von airnow.gov: <https://www.airnow.gov/aqi/aqi-basics>

Abbildungsverzeichnis

Abbildung 1	Temperatur Entwicklung von 1995 - 2020	7
Abbildung 2	Durchschnittstemperaturen pro Kontinent bzw. Region	8
Abbildung 3	Durchschnittstemperaturen von Nordamerika	8
Abbildung 4	Air Quality Index der U.S. Environmental Protection Agency	9
Abbildung 5	Luftqualitätsmessungen aufgeteilt in Kategorien	10
Abbildung 6	Durchschnitt des Luftqualität Index	10
Abbildung 7	AQI Luftqualitätsskala der USA	11
Abbildung 8	Gasanteile in der Luft: Kalifornien vs Maine	11
Abbildung 9	Prognostizierter Temperaturverlauf Weltweit für die nächsten 10 Jahren	12
Abbildung 10	Zerlegung des Temperaturverlaufs der USA in Trend, Saisonalität und übrige Messabweichungen	13
Abbildung 11	Vorhersage der Luftqualität (USA) in den nächsten 10 Jahren	13
Abbildung 12	Zusammenhänge zwischen Temperatur und Treibhausgasen in Kalifornien	14
Abbildung 13	Zusammenhang zwischen den Durchschnittstemperaturen und den vier Gasen	15

Selbstständigkeitserklärung

Die Länge des vorliegenden Textes ab und inklusive Kapitelüberschrift 1 bis vor diesen Abschnitt beträgt 2846 Wörter.

Wir bestätigen, die vorliegende Arbeit selbstständig verfasst zu haben. Sämtliche Textstellen, die nicht von uns stammen, sind als Zitate gekennzeichnet und mit dem genauen Hinweis auf ihre Herkunft versehen. Die verwendeten Quellen (gilt auch für Abbildungen, Grafiken u.ä.) sind im Literatur- bzw. Quellenverzeichnis aufgeführt.

Bern, 10.06.2022



Iris Lüthi



Maja Velkova



Yannik Zimmermann