Praktikumsbericht Data Science 1

Robin Mayer, Nils Nover, Mariella Zunker

26. Juni 2020

Inhaltsverzeichnis

1 Fragestellung und Datensätze **2**

2 Datenaufbereitung **2**

3 Datenauswertung **3**
3.1 Betrachtung der beiden Datensätze 3
3.2 Betrachtung der Werte und lineare Regression 4
3.3 Nicht-lineare Regression 5

4 Diskussion **6**

5 Quellen **6**

Code und Daten für dieses Projekt sind einsehbar über das GitHub Repository:
https://github.com/RaumschiffGS/Data-Science-Project

1 Fragestellung und Datensätze

Die Frage der Schädlichkeit von Stickoxiden und deren Zusammenhang mit Dieselmotoren hat in der Vergangenheit die Debatte um die Verkehrswende dominiert. Um einen Überblick darüber zu bekommen, sollten in diesem Projekt Antriebsdaten mit NO2-Werten verglichen werden.
Zur Analyse wurden jahrweise .xlx-Datensätze mit NO2-Werten des Umweltbundesamtes (UBA, Abb. 1, (1)) sowie .pdf-Dateien des Kraftfahrtbundesamtes (KBA, Abb. 2. (2)) genutzt.

bundesland	jahr	station	name	umgebungstyp	emissionstyp	jahresmittel	maxstundenwert
Brandenburg	2002	'DEBB001'	Burg (Spreewald)	vorstädtisches Gebiet	Hintergrund	10	69
Brandenburg	2002	'DEBB006'	Cottbus-Süd	städtisches Gebiet	Hintergrund	19	107
Brandenburg	2002	'DEBB009'	Forst	vorstädtisches Gebiet	Hintergrund	16	83
Brandenburg	2002	'DEBB021'	Potsdam-Zentrum	städtisches Gebiet	Hintergrund	21	111

Abbildung 1: Auszug aus dem NO2-Datensatz.

land	rb	Stadt	Insgesamt	Benzin	Diesel	Gas (einschl. bivalent)	Hybrid	Elektro
BADEN-WUERTTEMBERG	STUTTGART	08111 STUTTGART,STADT	298.172	182.451	111.148	1.893	1.788	814
BADEN-WUERTTEMBERG	STUTTGART	08115 BOEBLINGEN	244.396	155.578	85.499	1.452	1.336	474
BADEN-WUERTTEMBERG	STUTTGART	08116 ESSLINGEN	319.920	208.669	107.640	1.907	1.301	325

Abbildung 2: Auszug aus dem KFZ-Datensatz.

2 Datenaufbereitung

Zur sinnvollen Verarbeitung wurden die .pdf-Dateien ins .xslx-Format konvertiert. Aufgrund der großen Unterschiede der Formate musste manuell viel nachgebessert

werden. Die .xlsx-Dateien wurden im Anschluss in ein .csv-Format konvertiert und so eingelesen.
Aufgrund der Verfügbarkeit bestehender Tools wurde für die Auswertung Python mit den Analysetools Pandas sowie Numpy und Matplotlib gewählt und das Jupyter Notebook als interaktive Programmierumgebung genutzt.
Nach dem Einlesen wurden die Datensätze auf Vollständigkeit und Fehler durchsucht. Dabei wurde auf nicht erfasste Datenpunkte sowie offensichtliche Abweichungen wie negative Zahlen geachtet. Zudem wurden alle als String codierten Variablen überprüft und vereinheitlicht. So wurden alle Namen in Großbuchstaben und ohne Umlaute dargestellt, sowie Rechtschreibfehler und Unterschiede in der Darstellung von Doppelnamen korrigiert. Trennzeichen wurden vereinheitlicht und Zahlenwerte als Float gecastet.
Messwerte, für die kein Ort angegeben war, wurden aus dem Datensatz gelöscht, da eine Zuordnung in der verfügbaren Zeit nicht möglich war. Schließlich wurde aus den einzelnen Dateien für jedes Jahr eine Gesamtdatei für alle Jahre erstellt.
Die Aufbereitung der Daten nahm viel Zeit in Anspruch. Mit mehr Kapazitäten hätten Fehlerbehebung und Vervollständigung der Daten noch stärker verfolgt werden können, indem z.B. alle Daten zugeordnet werden oder die Verteilung der Werte genauer betrachtet wird. In Abbildung 3 sind beispielhaft einige Daten als Heatmaps dargestellt.

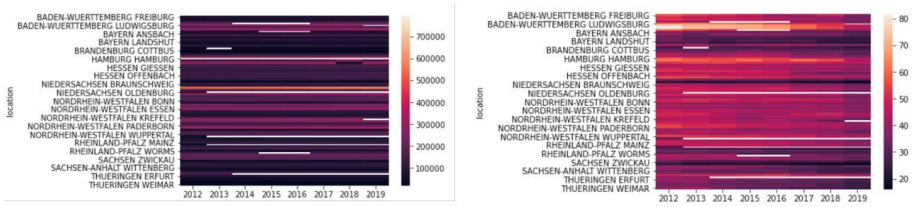


Abbildung 3: Ausschnitt aus den Daten als Heatmaps dargestellt. Links die NO2-Werteentwicklung, rechts die Entwicklung der Autoanzahlen.

3 Datenauswertung

3.1 Betrachtung der beiden Datensätze

Die Datensätze wurden zuerst isoliert betrachtet. Zur Messung der NO2-Werte existierten Daten aus den Jahren 2002-2019, KfZ-Daten konnten von 2012-2019 genutzt werden. Die Datensätze enthielten auch Zuordnungen der einzelnen Orte zu verschiedenen Abstufungen der Urbanität wie "vorstädtisches Gebiet", die zur Vereinfachung zu den drei Kategorien "städtisch", "vorstädtisch" sowie "ländlich" zusammengefasst wurden. Für einen ersten Überblick wurden die NO2-Werte über die Zeit geplottet (Abb. 4). In der Tendenz stimmt das Ergebnis mit der Auswertung des UBA (3) überein, die absoluten Werte weisen leichte Abweichungen auf. Dies ist wohl auf die veränderte Einteilung der Orte in die Kategorien zurückzuführen.

Zur weiteren Analyse wurden die Datensätze im Anschluss zusammengefügt. Dafür wurde pro Jahr und Bundesland überprüft, ob eine Stadt im NO2-Datensatz auch im KfZ-Datensatz existiert und bei Übereinstimmung wurden neue Spalten hinzugefügt. Dieses Merging stellte eine größere Schwierigkeit dar als vermutet, da

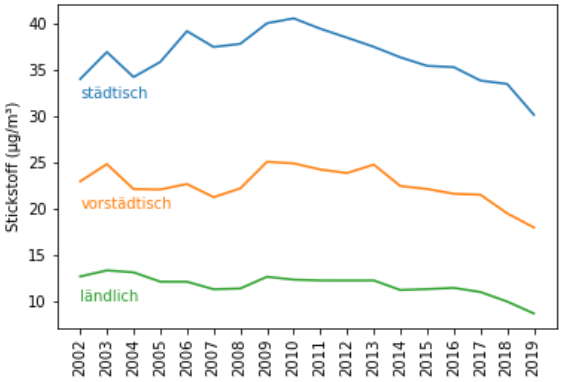


Abbildung 4: Entwicklung der NO2-Werte über die Zeit. Blau: Städtisch. Gelb: Vorstädtisch. Grün: Ländlich.

die Datensätze nicht die gleiche Zuordnung der Werte zu den Orten enthielten. So bezogen sich die Daten des KBA auf die regulären Landkreise, die Messungen des UBA waren jedoch lediglich dem nächsten Ort zugewiesen. Diese Zuordnung führte dazu, dass beim Merging ländliche Gebiete weniger gut zugeordnet werden konnten als größere Städte (Abb. 5).

Zudem ist zu beachten, dass einige Städte in den Datensätzen mehrfach vorkommen (z.B. Weimar Schwanseestr. und Weimar Steubenstr.). Hier wurde zur Vereinfachung der Mittelwert dieser Stationen verwendet. Das Merging beanspruchte einen Großteil der Zeit des gesamten Projektes und konnte dennoch nicht die angestrebte Vollständigkeit erreichen. Wäre mehr Zeit vorhanden, hätten noch mehr Versuche angestrengt werden können, alle Städte in den beiden Datensätzen zuzuordnen. Möglich wäre dabei, die Orte eindeutig mithilfe eines Karten-Tools zu ermitteln. Auch auf Ausreißer hätte mehr geachtet werden können. Zudem wäre es noch interessant gewesen, die zeitliche Entwicklung in den Daten zu betrachten und diese in gesellschaftliche Kontexte zu setzen, wie z.B. Gesetzesänderungen, Umweltzonen und lokale Unterschiede.

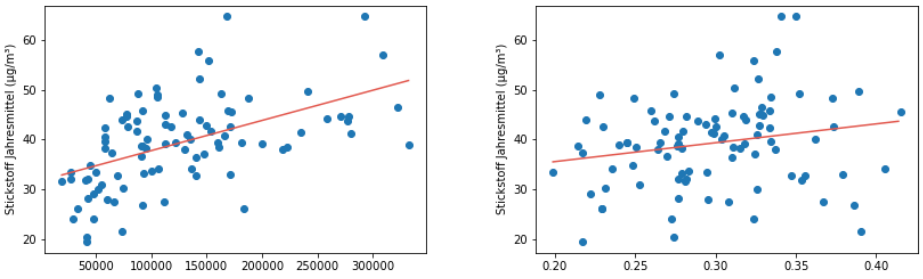
umgebungstyp	
ländlich Gebiet	13
ländlich regional	48
ländlich stadtnah	39
städtisches Gebiet	2131
vorstädtisches Gebiet	292

Abbildung 5: Anzahl an Orten in den Kategorien, die zugeordnet werden konnten.

3.2 Betrachtung der Werte und lineare Regression

Im Anschluss an die Zuordnung wurden die NO2-Werte gegen die Fahrzeugzahlen aufgetragen. Dabei konnte bereits ein grober Trend erkannt werden. Mithilfe linearer Regression wurde versucht, den Zusammenhang genauer zu beschreiben (Abb. 6). Außerdem zeigten sich im Bereich der Fahrzeugzahlen einige extreme Ausreißer, welche bei der weiteren Analyse nicht berücksichtigt wurden.

Das Ergebnis wurde mithilfe von drei Error-Parametern evaluiert. Dabei zeigte sich ein mittlerer absoluter Fehler von 6,00, ein MSE von 56,78 sowie ein R2-Score von 0,26. Die gefundene Regressionsgerade beschreibt den gefundenen Zusammenhang somit weniger gut.
Dann wurde der Anteil an Dieselfahrzeugen betrachtet und gegen die NO2-Werte



(a) Fahrzeuge Insgesamt (b) Dieselanteil

Abbildung 6: NO2-Werte in Abhängigkeit der Fahrzeugzahlen (a) und des Dieselanteils (b).

aufgetragen. Eine grobe Betrachtung des Plots ließ auf eine eher zufällige Verteilung mit geringfügiger linearer Tendenz schließen. Auch durch lineare Regression konnte kein starker Zusammenhang festgestellt werden (Abb. 6).

Die Evaluation ergab einen absoluten durchschnittlichen Fehler von 6,61, einen MSE von 73,85 sowie einen R2-Score von 0,04, und lässt darauf schließen, dass kein sinnvoller Erkenntnisgewinn durch lineare Regression geschaffen wurde.

3.3 Nicht-lineare Regression

Die Größe des Stadtgebietes und physikalische Eigenschaften der Gase wie Diffusion könnten dafür sorgen, dass sich der Zusammenhang mit einer Wachstumskurve anstelle einer Gerade genauer beschreiben lässt. Die Auslastung der Straßen ist je nach Größe auf einen bestimmten Verkehrsdurchfluss begrenzt. Neue Fahrzeuge in der Stadt sorgen hier also nicht gleichzeitig für einen hören NO2 Wert. Daher ist es möglich, dass eine Sättigungskurve die NO2-Entwicklung besser beschreibt. Es wurde also noch eine nicht-lineare Regressionsanalyse durchgeführt (Abb. 7).

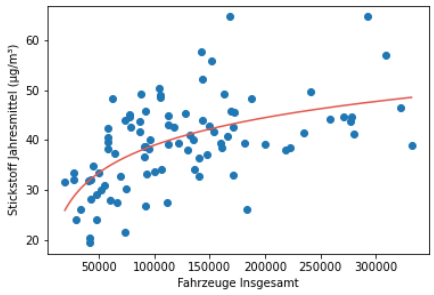


Abbildung 7: Nicht-lineare Regression des Zusammenhangs von Fahrzeugzahlen und NO2-Werten.

Bei dieser Analyse wurde nach Überprüfung verschiedener Funktionen schließlich eine log-Funktion als bester Fit gewählt. Für diese ergab sich ein durchschnittlicher Fehler von 5.73 und ein MSE von 52.11. Eine solche Kurve bestärkt die vorherige These einer möglichen Sättigung der NO2-Werte bei vielen Fahrzeugen.

4 Diskussion

Der Zusammenhang zwischen NO2-Werten und Fahrzeugzahlen wurde durch die Daten zwar gezeigt, doch die Regressionsanalyse zeigt, dass für eine genaue Beschreibung zusätzliche Daten benötigt werden. So ist zweifelhaft, ob Autos auch tatsächlich nur an ihrem Meldeort gefahren werden. Gewerbe, die alle Fahrzeuge in einer Stadt melden, können die Ergebnisse verzerren. Es sollten Fahrzeugdaten verwendet werden, die an einer Stelle nachgewiesen wurden. Um statistische Sicherheit zu erlangen, wäre zudem ein Hypothesentest notwendig.
Bei der Analyse der Dieselfahrzeuge wurde kein zuverlässiger Zusammenhang gefunden. Allerdings kann die Einteilung in Diesel/Nicht-Diesel hierbei als zu grob betrachtet werden.
Insgesamt war vor allem die Datenaufbereitung und das Merging der Datensätze mit Schwierigkeiten verbunden. Mit mehr Zeit wäre es zudem möglich gewesen, sich die zeitliche Entwicklung anzusehen. Die Ergebnisse zeigten zwar sinkende NO2-Werte über die Zeit, doch von genauerem Interesse wären auch lokale Unterschiede wie Umweltzonen oder Ballungsräume.
Auch der Vergleich von urbanen und ländlichen Gebieten konnte aufgrund der Herausforderungen bei der Fusion der Datensätze nicht gezogen werden.

5 Quellen

(1) https://www.umweltbundesamt.de/themen/luft/luftschadstoffe/stickstoffoxide
(2) https://www.kba.de/DE/Statistik/Fahrzeuge/Bestand/ZulassungsbezirkeGemeinden/zulassungsbezirke_node.html
(3) https://www.umweltbundesamt.de/themen/luft/daten-karten/entwicklung-der-luftqualitaet#entwicklung-der-luftqualitat-in-deutschland