INWT Statistics

AP1 - Literaturrecherche

1. **Überblick über Ansätze in der Statistischen Modellierung**
    a. Klassische Statistik
    b. Maschinelles Lernen
    c. Künstliche Intelligenz

2. **Literatur zu Statistischen Modelle für die Luftschadstoffprognose**

3. **Fazit**

1. Überblick über Ansätze in der statistischen Modellierung

Es gibt viele verschiedene Ansätze für die statistische Modellierung von der Konzentration von Luftschadstoffen. Ganz allgemein gesehen sind die gemessenen Konzentrationen strikt positive, metrische Zielvariablen. D.h. sie können alle möglichen Zahlen > 0 annehmen und sind nicht entweder 0 oder 1, oder nur ganzzahlig wie z.B. die Anzahl Autos. Daher werden wir im folgenden statistische Modelle für metrische Variablen betrachten.

Im Folgenden wird eine Übersicht über die unterschiedlichen Modellierungsansätze aus der klassischen Statistik, des maschinellen Lernens, und der künstlichen Intelligenz gegeben.

a) Klassische Statistik

- Die Lineare Regression ist das Standardmodell der klassischen Statistik. Dabei wird die Zielvariable durch eine lineare Kombination von erklärenden Variablen modelliert. Die Parameter alpha (Achsenabschnitt) und beta (Effekt der Variable) werden dabei mit Hilfe von statistischen Methoden geschätzt. Die Funktion NO2 = alpha + beta * Verkehrsdichte könnte nach dem Schätzen beispielsweise NO2 = 22 + 3.5 * Verkehrsdichte heißen.

- Die Lineare Regression beachtet jedoch im Standardfall den zeitlichen Zusammenhang nicht. Die Tatsache, dass die aktuelle NO2 Konzentration ähnlich der NO2 Konzentration vor einer Stunde ist, wird nicht modelliert.
- Zeitreihen-Modelle wie ARIMA (Autoregressive Integrated Moving Average) Modelle berücksichtigen die zeitlichen Zusammenhänge. Im simpelsten Modell, dem AR(1) sieht die Modellgleichung wie folgt aus: NO2_heute = beta * NO2_gestern. Weitere Dynamiken und Variablen könnten ebenfalls in der Formel berücksichtigt werden. Komplexe Zeitreihen-Modelle haben in vielen Bereichen hervorragende Ergebnisse erzielt und werden z.B. in der Modellierung volkswirtschaftlicher Zusammenhänge verwendet. Auch für die Modellierung der Luftschadstoffprognose finden sie Verwendung, zeigen jedoch häufig eine schlechtere Performance, als Modelle die Maschinelles Lernen nutzen.

b) Maschinelles Lernen

- Modelle des Maschinellen Lernens (ML) sind in der Lage, große Mengen an Daten hervorragend zu nutzen. Ein besonders hervorstechender Modellierungsansatz beruht auf Entscheidungsbäumen. Entscheidungsbäume ermöglichen die Modellierung vieler Interaktionen und die Identifizierung komplexer Zusammenhänge, indem “Regeln” aus den Daten gelernt werden. z.B.: Wenn Anzahl der KFZ pro Stunde > 1000 und Zeit seit letztem Regen > 3 Tage dann ist NO2 erhöht.
- Die Lernfähigkeit der ML-Modelle bringt jedoch auch neue Herausforderungen mit sich. ML-Modelle sind anfällig für Overfitting. Beim Overfitting lernt das Modell alle Besonderheiten der vorliegenden Daten, und ist nicht mehr generalisierbar. Es lernt nur ganz genau, die vorliegenden Daten zu beschreiben, und die Eigenheiten der Stichprobe zu verstehen. Die generell geltenden Zusammenhänge werden nicht gelernt.
- Ensemble Methoden lindern das Problem des Overfitting und verbessern die Vorhersagefähigkeit, indem sehr viele Modelle geschätzt und kombiniert werden. Sogenannte Random Forests sind eine Kombination

vieler Entscheidungsbäume. Dadurch kann identifiziert werden, welche Regeln wirklich allgemein gelten und immer wieder gefunden werden, und welche nur Eigenheiten der Stichprobe waren.

- Neben Ensemble Methoden hat vor allem das "Gradient Boosting" zu einer weiteren, erheblichen Verbesserung der Vorhersagefähigkeit der existierenden Random Forest Modelle geführt. Anstatt nach einer Schätzung mit dem Ergebnis zufrieden zu sein, fügt das Gradient Boosting weitere Modelle hinzu, um den Vorhersagefehler weiter zu reduzieren.
- XGBoost (eXtreme Gradient Boosting) ist ein generelles, erprobtes Modell, welches mit Hilfe von Gradient Boosting viele kleine "schwache" Regressionsbäume schätzt und es erlaubt, Modelle mit sehr hoher Vorhersagefähigkeit zu schätzen.

c) Künstliche Intelligenz (KI)

- Künstliche Neuronale Netze und Artifizielle Intelligenz (AI) erfreuen sich aktuell großer Beliebtheit da sie hervorragende Ergebnisse in Forschungsbereichen erzielen, in denen es zuvor wenig Fortschritt gab. Die stark angestiegene Rechenleistung erlaubt es, Modelle zur Sprach-, Bild- und Textanalyse zu erstellen, die hervorragendes Verständnis für komplexe Zusammenhänge erlernen.
- Für die NO2 Prognose werden oft sogenannte RNN (Recurrent Neural Network), oder LSTM (Long-Short-Term-Memory) Modelle verwendet. Im Vergleich zu anderen KI Modellen muss bei der NO2 Prognose die Reihenfolge der Beobachtungen mit berücksichtigt werden, sodass hier der Fokus auf der Modellierung der Sequenz der beobachteten Daten liegt. Daher haben die Modelle einen ähnlichen Aufbau zu den klassischen ARIMA Modellen. Es wird versucht mit künstlichen Neuronalen Netzen die Zeitreihe zu modellieren. Der RNN/LSTM Ansatz zur NO2-Prognose konnte jedoch bisher nicht die hervorragenden Ergebnisse der anderen KI Methoden aus Sprach, Bild und Texterkennung reproduzieren. Oftmals sind klassische statistische Methoden ähnlich gut, und Machine Learning

Methoden wie XGBoost erzielen ähnlich gute oder sogar bessere Ergebnisse.

2. Literatur zu Statistischen Modelle für die Luftschadstoffe

Im Folgenden wird eine Auswahl wichtiger wissenschaftlicher Fachliteratur kurz vorgestellt, die für die Luftschadstoffprognose relevant ist. Die Artikel sind chronologisch absteigend geordnet.

- Gilik et al (2022)
  - https://doi.org/10.1007/s11356-021-16227-w
  - In dem Artikel werden verschieden KI basierte Methoden (LSTM, CNN) zur stündlichen Vorhersage von PM10, NO2, und andere Luftschadstoffe für Istanbul, Kocaeli (eine benachbarte Stadt) und Barcelona verglichen. Die Modelle, die Meteorologie berücksichtigen, liefern bessere Ergebniss, als solche ohne meteorologische Parameter. Der Transfer des Modells von Kocaeli nach Istanbul funktioniert besser als von Barcelona nach Istanbul, da in *benachbarten Städten ähnliche Bedingungen* herrschen.

- Park et al. (2021)
  - DOI: https://doi.org/10.3390/su132413782
  - Park et al. schätzen drei auf boosting basierende ML Modelle (Gradient Boosting Machines, Extreme Gradient Boosting (XGBoost), lightGradient Boosting Machines) in Seoul, Südkorea für PM10. Die drei verwendeten Algorithmen unterscheiden sich dabei hauptsächlich in den Details. Light Gradient Boosting Machines, dicht gefolgt von XGBoost, liefern die besten Ergebnisse. Als Prädiktoren werden hauptsächlich Wetterdaten verwendet. Die Wetterdaten werden mittels Inverse Distance Weighting zu den (an andern Orten gelegenen) Messcontainern für die Luftschadstoffe gebracht. Dafür wird ein 30m Grid verwendet. Obwohl *hauptsächlich Wetterdaten* genutzt werden, liefern die Modelle schon akzeptable Ergebnisse (R^2 von 0.84 bzw. 0.83). Seoul scheint dabei von sehr starker Luftverschmutzung geprägt zu sein, es wird “Yellow Dust” gemessen. Dieser geht als wichtigste Variable aus dem Modell hervor.

- Zhong et al. (2021)
  - DOI: https://doi.org/10.1093/nsr/nwaa307

  - Zhong et al. trainieren Gradient Boosting Machines (genauer: LightGBM, ein ML Modell) mit lokalen meteorologischen Daten, um PM2.5 in China vorherzusagen und eine Karte auch für Gebiete ohne Messstationen bereitzustellen. Die zeitliche Auflösung ist stündlich bis jährlich, die räumliche Auflösung um Größenordnungen ungenauer als in Berlin geplant (0.25°). Es wird gezeigt, dass sich das System sehr gut eignet um Lücken in satellitengestützten PM2.5 (*Aerosol optical depth*) zu schließen.

- Goulier et al. (2020)
  - DOI: https://doi.org/10.3390/ijerph17062025
  - In dem Artikel werden Artificial Neural Networks (Kategorie: KI) genutzt um ein Modell für eine Messstation in Münster für zehn verschiedene Luftschadstoffe (unter anderem NO2, PM2.5, PM10) zu lernen. Die zeitliche Auflösung ist stündlich, es wird aber keine Vorhersage durchgeführt, sondern der jeweils aktuelle Wert mit dem Modell bestimmt ("nowcasting"). Die Stationen, die meteorologische Daten und Hintergrundbelastung liefern, sind jeweils einige Kilometer entfernt. Ein interessanter Aspekt ist, dass neben Meteorologie und der Anzahl der Fahrzeuge der *Lärmpegel* in unterschiedlichen Frequenzen eine wichtige Variable für die Vorhersage von NO2 ist.

- Li et al. (2020a)
  - DOI: https://doi.org/10.1016/j.jclepro.2020.121975
  - In dem Artikel werden sechs verschiedene ML Algorithmen verglichen, um PM2.5 und NOx in Hongkong auf Straßenebene an drei verschiedenen Messstationen stündlich vorherzusagen. Unter den Algorithmen die für die Vorhersage am besten funktionieren sind Boosted regression trees und XGBoost, die beide auf Gradient Boosting basieren. Emissionen werden mit einer einfachen deterministischen Beziehung abgeschätzt und in das Modell gegeben. Ein für Berlin interessantes Nebenergebnis der Studie ist auch, dass in Bezug auf das Verkehrsaufkommen für das Modell keine große Rolle spielte, wie groß der für das summierte Verkehrsaufkommen einbezogene Radius (im Rahmen von 50m bis 500 m) war.

- Schneider et al. (2020)
    - DOI: https://doi.org/10.3390/rs12223803
    - In dem Artikel wird PM2.5 in Großbritannien in einer täglichen Auflösung modelliert. Es werden Random Forests genutzt (ein ML Algorithmus), der auf Entscheidungsbäumen basiert. Ein interessantes Merkmal, welches für das Modell erfolgreich genutzt wird, sind räumlich benachbarte Stationsdaten, die mit einem zeitlichen Lag in dem Modell zur Vorhersage genutzt werden. Zudem wird auf ein *deterministisches Modell (EMEP4UK)* aufgebaut. Dieses stellt die mit Abstand wichtigste Variable für das statistische Modell dar. Das unterstreicht für Berlin, dass die geplante Integration der CAMS Daten einen großen Schritt zur Verbesserung unseres Berliner Modellsystems darstellen könnte.

- Sharma et al. (2020)
    - DOI: https://doi.org/10.1016/j.scitotenv.2019.135934
    - Sharma et al. bauen ein Modellsystem, das jeweils für die nächste Stunde PM2.5 und PM10 in Australien erfolgreich vorhersagt. Sie kombinieren dafür eine Empirical Mode Decomposition (eine Zerlegung des Signals in Schwingungen) mit extreme learning machines (ein AI Algorithmus). Die vorherige Zerlegung des Signals mittels Empirical Model Decomposition ist ein innovativer Ansatz, der unserer Einschätzung nach insbesondere für *kurzfristige* Prognosen (hier: der nächste Wert in einer Stunde) sehr gute Ergebnisse liefert.

- Stafoggia et al. (2020)
    - DOI: https://doi.org/10.3390/atmos11030239
    - In dem Artikel werden Random Forests (ein ML Algorithmus) genutzt um täglich PM2.5, PM10, NO2 und Ozon in 1km² räumlicher Auflösung in Schweden bereitzustellen. 50 Messstationen stehen zum Training zur Verfügung. Die CAMS Daten des Copernicus Konsortiums, die wir auch für Berlin nutzen möchten, gehen aus dem Modell als eine wichtige Variable hervor. Auch wichtig sind zwei meteorologische Parameter, die wir bisher für die Luftschadstoffprognose in Berlin nicht vorgesehen hatten

(*Wolkenbedeckung, Luftdruck*). Beide importieren wir nun zusätzlich im Rahmen der Daten der DWD Wettervorhersage.

- Di et al. (2019)
  - DOI: https://doi.org/10.1016/j.envint.2019.104909
  - Di et al. nutzen ein Modellsystem um PM2.5 in den USA auf zwei räumlichen Auflösungen (1km² und 100m x 100m) bereitzustellen. Das Modellsystem ist ein Ensemble aus 3 Algorithmen (2 ML, 1 KI): Gradient Boosting, Random Forest und Neuronales Netze. Ein Ensemble aus mehreren verschieden Algorithmen liefert typischerweise bessere Ergebnisse, als ein Algorithmus alleine.

- Chen et al. (2019)
  - DOI: https://doi.org/10.1016/j.atmosenv.2019.01.027
  - Chen et al. nutzen XGBoost um fehlende PM2.5 Werte in China mit einer sinnvollen Prognose zu füllen (räumliche Auflösung 3 km²). Es gibt eine ganze Reihe ähnlicher Literatur in China, die statistische Modelle anwenden, um PM2.5 Karten in China mit unterschiedlichen Algorithmen und räumlichen Auflösungen zur erstellen, z.B.
    - Chen et al. (2021)
      - https://doi.org/10.1016/j.scitotenv.2020.144724
    - Guo et al. 2021
      - https://doi.org/10.1016/j.scitotenv.2021.146288
    - Wei et al. (2021)
      - https://doi.org/10.1016/j.rse.2020.112136
    - Wei et al. (2020)
      - https://doi.org/10.5194/acp-20-3273-2020
    - Li et al. (2020b)
      - https://doi.org/10.1016/j.isprsjprs.2020.06.019
    - Zhang et al. (2021)
      - https://doi.org/10.1016/j.chemosphere.2020.128801

- Li et al. (2017)
  - DOI: https://doi.org/10.1002/2017GL075710
  - Li et al. nutzen sogenannte Deep Belief networks (Kategorie: AI), um stündlich PM2.5 in China vorherzusagen. Sie vergleichen die Modellperformanz mit und ohne die Integration raum-zeitlicher Eigenschaften in das Modell. Die Integration *raum-zeitlicher Eigenschaften* kann dabei den Vorhersagefehler deutlich reduzieren.

- Paas et al (2017)
  - DOI: https://doi.org/10.3390/environments4020026
  - In dem Artikel werden PM2.5 und PM10 und die Konzentration von Partikelanzahlen mittels Artificial Neural Networks (Kategorie: AI) an zwei Messstationen in Aachen und Münster vorhergesagt. Die zeitliche Auflösung beträgt 10 Minuten. Eine Neuheit ist die Nutzung von *akustischen Daten* in verschiedenen Frequenzen für die Modellierung. Der Ansatz funktioniert unterschiedlich gut an den beiden Messstationen der unterschiedlichen Städte, da für Feinstaub wichtige Eingangsgrößen wie Emissionen lokaler Haushalte nicht berücksichtigt werden.

- Beleen et al. (2013)
  - DOI: https://doi.org/10.1016/j.atmosenv.2013.02.037
  - Im Rahmen des ESCAPE Projektes (European Study of Cohorts for Air Pollution Effects) werden Effekte der Luftqualität auf die Gesundheit untersucht. Dafür müssen die Messwerte (14 tägig) durch NO2 und NOX Passivsammler zu den Wohnorten der Teilnehmer gebracht werden. Das passiert in 36 teilnehmenden europäischen Städten durch sogenannte *Land Use Regression* (LUR) Modelle, die in Eeftens et al. (2012) entwickelt wurde. Für NO2 spielen insbesondere die *Verkehrsdichte und andere verkehrsbezogene Parameter* eine zentrale Rolle.

- Eeftens et al. (2012)
  - DOI: https://doi.org/10.1021/es301948k

- Eeftens et al. entwickeln Land Use Regression (LUR) Modelle für das o.g. ESCAPE Projekt. Ein LUR-Modell ist dabei eine klassische multiple lineare Regression in die Prädiktoren schrittweise aufgenommen werden, falls sie zu einer Verbesserung des korrigierten R^2 führen und die Richtung des Effekts für Experten plausibel ist. Die Güte des Modells wird mit einer sogenannten leave-one-out cross validation schlussendlich überprüft. Die Entwicklung des Modells in dem Artikel erfolgt für PM2.5, PM10 und andere Feinstaub zugehörige Variablen. Die Anwendung auf NO2 erfolgt in Beleen et al. (2013) und entwickelt sich zu einem Standard in dem Gebiet (z.B. Dons et al., 2015[1], Stafoggia et al., 2014[2]). Im Kontext des geplanten Modells in Berlin ist interessant, dass die letztendlich vom LUR Modell *selektierten Variablen sich stark je nach Stadt unterscheiden*.

[1] https://doi.org/10.1016/j.scitotenv.2014.01.025
[2] https://doi.org/10.1289/ehp.1307301

3. Fazit

Ein auf Gradient Boosting Machines beruhendes Modellsystem findet in der Literatur zu statistischen Modellen für die Luftschadstoffprognose häufig Anwendung - mit oft sehr guten Ergebnissen. In dem aktuellen Review von Anchan et al (2022)[3] zur Vorhersage von PM2.5 wird erwähnt, dass *XGBoost in 4 von 5 Methoden-Vergleichen überlegen* war und ML Techniken generell vorteilhaft sind, wenn es um eine (zeit-)effiziente Anwendung mit größtmöglicher Genauigkeit geht. Aus diesem Grund möchten wir daran festhalten, ein Modellsystem, das auf Gradient Boosting beruht, für Berlin zu implementieren.

In Bezug auf klassische Statistik, finden insbesondere Land Use Regression (LUR) Modelle und Inverse Distance Modelling erfolgreich Anwendung, wenn es darum geht, Daten, die an unterschiedlichen Orten gemessen wurden, zusammen zu bringen. Das ist für Berlin zum Beispiel nötig, wenn Beobachtungen der Verkehrsdetektoren nicht immer direkt neben den Messcontainern der Luftschadstoffe stattfinden. Hierfür werden wir einen *LUR-Ansatz* nutzen, um den an der Verkehrdetektoren beobachteten Verkehr anhand der Verkehrsmengenkarte 2019, der städtebaulichen Dichte und der StEP Straßenklassifikation aus dem Detailnetz von Berlin auf den zu erwartenden Verkehr an den Messstationen zu skalieren.

Aus dem Artikel von Staffogia et al. (2020) geht hervor, dass meteorologische Parameter, die wir bisher nicht für die Vorhersage der Luftqualität in Berlin vorgesehen hatten *(Wolkenbedeckung und Luftdruck*), wichtige Prädiktoren sein könnten. Wir integrieren deshalb beide Variablen zusätzlich in unsere Datenbank. Sie sind im Rahmen der Wettervorhersage des DWD verfügbar.

Zudem gibt es in der Literatur Hinweise darauf, dass die *Integration von CAMS Daten* für Berlin einen großen Schritt zur Verbesserung des Modellsystems darstellen kann. Wenn die Integration eines existierenden deterministischen Modells stattfindet, so spielt dieser Prädiktor in der Regel eine wichtige Rolle in dem Modell und trägt zur Verbesserung der Prognosequalität bei (Schneider et al, 2020).

[3] https://doi.org/10.1007/978-981-16-3342-3_6

Für die Zukunft könnten *geostationäre Satelliten* eine Rolle spielen. Diese werden stündlich hochaufgelöste Beobachtungen von PM und NO2 liefern (Bsp. Sentinel-4,, Review von Sorek-Hammer et al., 2020[4]). Der Launch der Sentinel-4 Satelliten ist für 2023 und 2030 geplant. Sorek-Hammer et al. (2022)[5] demonstrieren in einem im April 2022 erschienenen Artikel für vier Städte (New York, Los Angeles, Vancouver, London) mit welcher räumlichen Auflösung und Genauigkeit eine Übersetzung der Information von Satellitenbildern in Luftschadstoffkonzentrationen in Zukunft möglich sein könnte.

[4] https://doi.org/10.1016/j.envint.2020.106057
[5] https://doi.org/10.3390/atmos13050696