Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Luftqualität in Deutschland

Semesterarbeit im Modul

B34 Einführung Data Science

Gustav Gehricke, Marco Michaelis, Dennis-Simon Kuna

17. Juli 2025

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Codeverzeichnis

Kapitel 1

Einleitung

Die Luftqualität in urbanen und ländlichen Gebieten stellt weltweit ein entscheidendes Umwelt- und Gesundheitsthema dar, dessen Bedeutung durch zunehmende Urbanisierung und industrielle Aktivität kontinuierlich wächst [1]. Feinstaub, insbesondere $PM_{2.5}$, ist hierbei ein relevanter Schadstoff, der maßgeblich durch Verkehr, Industrie und Heizverhalten beeinflusst wird und zu erhöhten Konzentrationen in Städten führen kann [2, S. 1–2]. Die genaue Erfassung und Analyse dieser Konzentrationen sowie ihrer Einflussfaktoren sind für Umweltanalysen und die Formulierung effektiver verkehrspolitischer Maßnahmen zur Luftreinhaltung von zentraler Bedeutung. Die vorliegende Arbeit widmet sich der Anwendung von Data-Science-Methoden zur Untersuchung der Luftqualität in Deutschland und verfolgt das Ziel, einen Beitrag zur Umweltforschung zu leisten.

Die Feinstaubkonzentration wird von verschiedenen Faktoren wie meteorologischen Bedingungen (z.B. Niederschlag, Wind) und anthropogenen Emissionsquellen beeinflusst [2, S. 4–6]. Für die Untersuchung sind standardisierte, zeitlich hochaufgelöste Messwerte essenziell, die über öffentliche APIs des Umweltbundesamtes (UBA) für Luftqualitätsdaten und des Deutschen Wetterdienstes (DWD) für Wetterdaten bezogen werden können [3], [4]. Zusätzlich bieten geographische Daten wie OpenStreetMap (OSM) die Möglichkeit, stadtstrukturelle Einflussgrößen auf die Luftqualität zu quantifizieren [5]. Ziel dieser Arbeit ist es, die vielschichtigen Einflussfaktoren auf die Luftqualität in Deutschland mittels datenwissenschaftlicher Methoden zu analysieren und zu quantifizieren. Daraus ergeben sich folgende zentrale Forschungsfragen und Hypothesen:

Niederschlagsereignisse und $PM_{2.5}$-Konzentration: Besteht ein Zusammenhang zwischen Niederschlagsereignissen, insbesondere der täglichen Niederschlagsmenge, und den darauf folgenden Veränderungen der $PM_{2.5}$-Konzentration in der Luft in Deutschland?

Urbane Strukturen und lokale Feinstaubbelastung: In welchem Ausmaß hängen geographische Merkmale aus OpenStreetMap (OSM) im Umkreis von 100m bis 500m mit der lokalen Feinstaubbelastung ($PM_{2.5}$) in Berlin zusammen, und weisen Stationen mit einer höheren Anzahl an Park- oder Wohngebietsflächen tendenziell geringere tägliche $PM_{2.5}$-Mittelwerte auf?

Kapitel 2

Datenbeschaffung

2.1 Luftqualitätsdatenbeschaffung über die BundesAPI

Die Datenerhebung für die vorliegende Arbeit basiert auf öffentlich zugänglichen Luftqualitätsdaten, die das Umweltbundesamt (UBA) über eine REST-basierte Programmierschnittstelle (API) bereitstellt. Diese API ermöglicht den Zugriff auf standardisierte Messwerte aus einem bundesweiten Netzwerk von Messstationen und bildet damit eine zentrale Grundlage für Umweltanalysen in Deutschland.

2.1.1 Datenquelle und API-Struktur

Die verwendete API des Umweltbundesamtes (UBA) bietet einen strukturierten Zugriff auf Luftgütedaten über die Basis-URL `https://www.umweltbundesamt.de/api/air_data/v3`. Zwei relevante Endpunkte wurden im Rahmen dieser Arbeit genutzt:

- `metajson`: Liefert Metadaten zu Stationen und Messkomponenten (z. B. NO_2, O_3, PM_{10}).
- `airqualityjson`: Gibt zeitlich aggregierte Messdaten zu Luftqualitätsindikatoren zurück, wobei jeweils eine Station pro Abfrage adressiert wird.

Die API liefert die Messwerte im JSON-Format und erlaubt die gezielte Abfrage für spezifische Zeiträume und Stationen. Ein beispielhafter API-Aufruf könnte wie folgt aussehen:

```
1 requests.get("https://www.umweltbundesamt.de/api/air_data/v3/airquality/
      json",
2     params={"date_from": "2016-01-01", "date_to": "2025-05-31", "station":
          1234})
```

Listing 2.1: *Beispielhafter API-Aufruf (Python)*

Dieses Format unterstützt eine flexible und skalierbare Verarbeitung von Umweltmessdaten, wie sie für wissenschaftliche Analysen erforderlich ist [3].

2.1.2 Technische Umsetzung der Datenakquise

Für die automatisierte Datenbeschaffung wurde ein Python-basiertes Crawler-Skript implementiert. Es ruft zunächst über den `meta/json`-Endpunkt die vollständigen Metadaten der Messstationen sowie der Messkomponenten ab. Basierend auf diesen Informationen wird anschließend für jede

verfügbare Station ein API-Request an den `airquality/json`-Endpunkt durchgeführt, um sämtliche Luftgütedaten für den Zeitraum vom 01.01.2016 bis zum 31.05.2025 zu extrahieren.
Die Antwortdaten werden stationsweise verarbeitet und in eine strukturierte Form überführt, wobei neben Zeitstempeln auch Messwerte, Indexklassifikationen sowie geographische Metainformationen in das finale Datenset einfließen. Das Ergebnis wird sukzessive in eine CSV-Datei (`airquality_data_full.csv`) geschrieben, die später als Datengrundlage für Analysen und Visualisierungen dient. Die Datenstruktur orientiert sich dabei an wissenschaftlichen Anforderungen an Reproduzierbarkeit und Nachvollziehbarkeit.

2.1.3 Beschreibung wichtiger Felder im Datensatz

Der Datensatz umfasst zahlreiche relevante Felder, von denen die wichtigsten nachfolgend erläutert werden:

- **Station ID**: Eindeutige Identifikationsnummer der Messstation.
- **Station City und Urbanization Type**: Standort und Art des Gebiets (z. B. städtisch, vorstädtisch, ländlich), was die räumliche Kontextualisierung ermöglicht.
- **Station Longitude und Station Latitude**: Geographische Koordinaten zur exakten Verortung der Messstation.
- **Start Datetime und End Datetime**: Zeitraum, für den die Messwerte gültig sind.
- **Component Name und Component ID**: Bezeichnung und ID des gemessenen Schadstoffs (z. B. Feinstaub ($PM_{2.5}$), Stickstoffdioxid (NO_2)).
- **Value**: Gemessener numerischer Wert der jeweiligen Schadstoffkonzentration.
- **Airquality Index (Component)**: Luftqualitätsindex der einzelnen Schadstoffe, der eine schnelle Einschätzung der Luftqualität erlaubt.
- **Data incomplete**: Flag zur Kennzeichnung, ob alle Komponenten vollständig sind; beeinflusst die Vollständigkeit des LQI.

2.1.4 Datenqualität und Umfang

Ein zentraler Aspekt der Datenbeschaffung war die Vollständigkeit der Messreihen. Die API kennzeichnet unvollständige Zeitintervalle über ein explizites Flag, das im Datensatz erhalten bleibt. Diese Information kann im weiteren Verlauf zur Datenfilterung und Plausibilitätsprüfung herangezogen werden.
Zudem erlaubt die Kombination von Stationsmetadaten (inklusive Urbanisierungstyp und exakten Geokoordinaten) mit spezifischen Komponentenmesswerten (z. B. Feinstaub oder Ozon) die fundierte Einordnung der Luftqualität in unterschiedlichen geographischen und infrastrukturellen Kontexten.

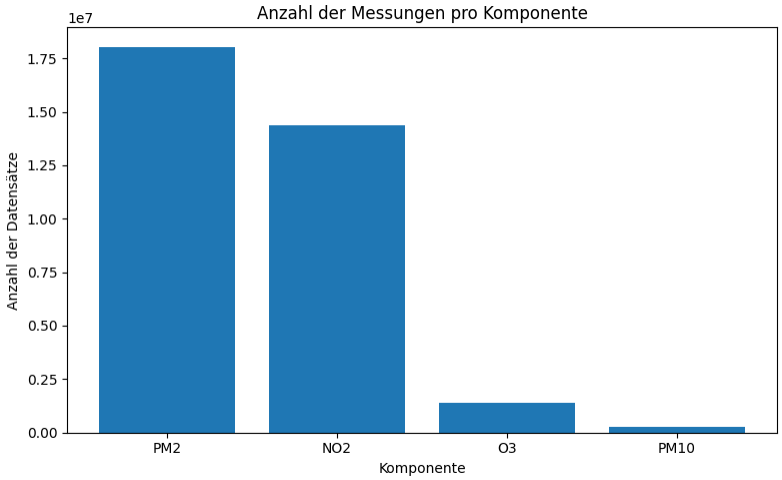


Abbildung 2.1: *Anzahl der Messwerte pro Komponente*

Abbildung 2.1 zeigt die absolute Verteilung der Messdatensätze pro Schadstoffkomponente. Es wird deutlich, dass Feinstaub $PM_{2.5}$ am häufigsten gemessen wurde, gefolgt von NO_2. Diese Dominanz weist auf die besondere Relevanz dieser Komponenten im nationalen Luftüberwachungskontext hin. Komponenten wie Ozon (O_3) oder PM_{10} wurden hingegen seltener aufgezeichnet, was entweder auf eine geringere Umweltpriorität oder eine selektivere Messabdeckung hindeuten könnte. Diese Ungleichverteilung sollte bei der Interpretation und Modellierung von Luftqualitätsdaten berücksichtigt werden.
Aufgrund der deutlichen Überrepräsentation von $PM_{2.5}$-Messungen sowie der herausgehobenen Bedeutung von $PM_{2.5}$ im Rahmen der Luftqualitätsbewertung erfolgt die weiterführende Analyse in dieser Arbeit primär auf Basis der $PM_{2.5}$-Werte [2, S. 1–2].

2.2 Wetterdatenbeschaffung über die DWD-API

Um mögliche Hypothesen eines Zusammenhangs zwischen Wetterereignissen und Veränderungen von Luftqualitätsdaten zu untersuchen, wurden stündliche Wetterdaten des Deutschen Wetterdienstes (DWD) herangezogen [6]. Das Vorgehen gliedert sich in folgende Schritte:

2.2.1 Datenquelle und Auflösung

Die Daten stammen aus dem Climate Data Center (CDC) des DWD, das Klimadaten für Deutschland bereitstellt. Zur Verfügung stehen sieben Auflösungen – u. a. 10-Minuten, stündlich, täglich und monatlich – wobei jede Auflösung unterschiedliche Parameter abdeckt [6]. Da die Luftqualitätsdaten in stündlicher Auflösung vorliegen, wurde für alle meteorologischen Größen die stündliche („hourly") Auflösung gewählt.

2.2.2 Metadatenakquise und Stationsauswahl

Über die frei zugängliche FTP-Schnittstelle des DWD wurden zunächst alle Metadaten zu den stündlichen Messstationen aus den Unterverzeichnissen der einzelnen Klimaparameter ausgelesen [7]. Ein typischer Metadatensatz enthält:

```
Stations_id;von_datum;bis_datum;Stationshoehe;geoBreite;geoLaenge;
Stationsname;Bundesland;Abgabe
```

(Beispiel: `00003;19500401;20110401;202;50.7827;6.0941;`
`Aachen;Nordrhein-Westfalen;Frei`).

Insgesamt ergab sich dadurch ein Pool von 2341 Stationen deutschlandweit, die mindestens einen der interessierenden Parameter erfassten.

2.2.3 Parameterauswahl

Um eine breite Basis für spätere Hypothesentests zu schaffen, wurden folgende stündlichen Parameter abgefragt:

Parameter	Beschreibung
temperature_air_mean_2m	Lufttemperatur in 2 m, °C
cloud_cover_total	Gesamtbedeckungsgrad, Achtel
humidity_absolute	Absolute Feuchte, g/m^3
precipitation_height	Niederschlagshöhe, mm
sunshine_duration	Sonnenscheindauer, Minuten
visibility_range	Sichtweite, m
wind_speed	Windgeschwindigkeit, m/s

Tabelle 2.1: *Ausgewählte Parameter und ihre Beschreibung*

Diese Parameter wurden mittels der Python-Bibliothek `wetterdienst` (Version 0.108.0) in einem einzigen Request je Monat und allen relevanten Stations-IDs abgefragt [8].

2.2.4 Datenabruf und -verarbeitung

Zeitraum: 01.01.2024–01.01.2025 (für einzelne Hypothesen wurden in Ausnahmefällen abweichende Zeiträume genutzt).

Batch-Downloads: Pro Monat wurde eine CSV-Datei erzeugt (jeweils ∼200 MB, ∼2,2 Mio. Einträge).

Dateiformat: Jeder Eintrag enthält die Felder:

`station_id,resolution,dataset,parameter,date,value,`
`quality,latitude,longitude`

(Beispiel: `00011,hourly,wind,wind_speed,2024-04-01T00:00:00+00:00,4.2,`
`10.0,47.9736,8.5205`).

Konsolidierung: Mit einem einfachen Python-Skript wurden alle Monats-CSVs zeilenweise aneinandergereiht, sodass eine Gesamttabelle mit ∼28 Mio. Einträgen (2,4 GB) entstand [9].

2.2.5 Qualitäts- und Zeitstempel-Handling

Quality-Flags: Das Feld `quality` (QUALITAETS_NIVEAU) beschreibt automatisierte und manuelle Prüfverfahren (u. a. Konsistenz- und statistische Tests) – es wurde im Vorfeld entschieden, sämtliche Messwerte ohne zusätzliche Filterung zu verwenden, da die Qualitätskontrolle des DWD bereits den Großteil offensichtlicher Ausreißer entfernt [4].

Zeitzonen: Die DWD-Timestamps liegen in UTC („+00:00"). Eine Umrechnung in Mitteleuropäische Sommerzeit (MESZ, UTC+2) erfolgte dort, wo die Zuordnung zu Luftqualitätsdaten in lokaler Zeit nötig war; andernfalls blieben die Zeitstempel in UTC.

2.2.6 Unsicherheiten

Die DWD-Messnetze folgen heute WMO-Standards, womit lokale Effekte minimiert werden. Historisch betrachtet variieren jedoch Prüfroutinen und Standardisierung (z. B. vor 1990 unterschiedliche Vorschriften in Ost-/Westdeutschland). Daher sollte für jede Hypothese geprüft werden, inwieweit lokale Besonderheiten oder Lücken in den Stationsreihen die Ergebnisse beeinflussen können [4].

Kapitel 3

Niederschlagsereignissen und PM2.5-Konzentration

3.1 Einleitung und Forschungsfrage

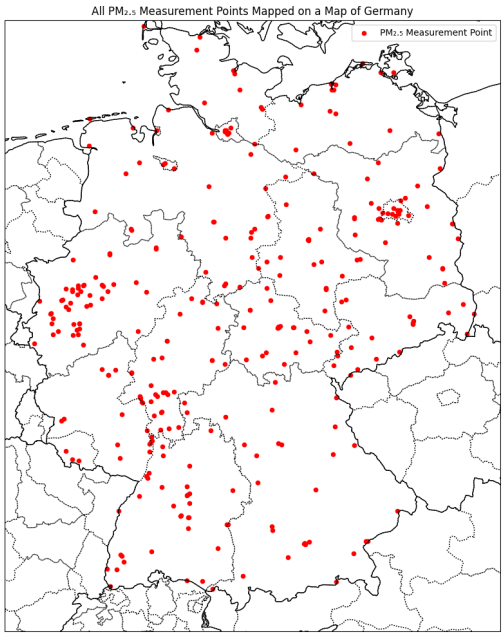
Die vorliegende Analyse untersucht den potentiellen Zusammenhang zwischen Niederschlagsereignissen und den darauf folgenden Veränderungen der $PM_{2.5}$-Konzentration in der Luft im Raum Deutschland. Die zentrale Hypothese lautet: „Es besteht ein Zusammenhang zwischen Niederschlagsereignissen – insbesondere der täglichen Niederschlagsmenge – und den darauf folgenden Veränderungen der $PM_{2.5}$-Konzentration in der Luft."
Das Ziel dieser Untersuchung ist es, mittels statistischer Verfahren zu überprüfen, ob und in welchem Ausmaß Niederschlagsereignisse die Luftqualität in Deutschland beeinflussen. Theoretisch wird davon ausgegangen, dass Niederschlag durch Auswascheffekte (Wet Deposition) zu einer Reduktion der Feinstaubkonzentration in der Atmosphäre führen kann [10, S. 249]. Die Analyse basiert auf den in Kapitel 2 beschriebenen Datensätzen der Luftqualitätsmessungen des Umweltbundesamtes und meteorologischen Daten des Deutschen Wetterdienstes für das Jahr 2024.

3.2 Datenpräparation und Vorverarbeitung

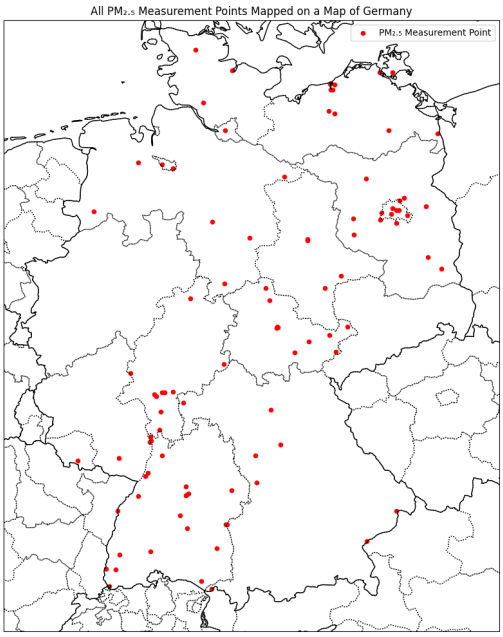
3.2.1 Aufbereitung der Luftqualitätsdaten

Die Präparation der $PM_{2.5}$-Messdaten erfolgte durch ein systematisches Vorgehen zur Gewährleistung der Datenqualität und Vollständigkeit. Zunächst wurden die aus Abschnitt 2.1 beschriebenen Daten bestehend aus 3,5 Millionen Einträgen eingelesen. Anschließend erfolgte eine Filterung auf die für die Hypothese relevante Komponente `PM2` (entspricht $PM_{2.5}$), deren Auswahl bereits im Unterabschnitt 2.1.4 detailliert begründet wurde.
Ein kritischer Aspekt der Datenqualitätssicherung war die Überprüfung der Vollständigkeit der Messreihen. Da die Luftqualitätsdaten stündlich erfasst werden und das Untersuchungsjahr 2024 ein Schaltjahr darstellt, wurde für jede Messstation die Vollständigkeit von $24 \cdot 366 = 8.784$ Einträgen pro Station überprüft. Stationen mit unvollständigen Datensätzen wurden systematisch aus der Analyse ausgeschlossen, um die statistische Belastbarkeit der Ergebnisse zu gewährleisten.

Diese Qualitätskontrolle führte zu einer erheblichen Reduktion der Anzahl der Stationen von ursprünglich 318 auf 97 Stationen mit vollständigen Jahresreihen. Die räumliche Verteilung der Stationen vor und nach der Bereinigung ist in Abbildung 3.1a und Abbildung 3.1b dargestellt.



(a) *Räumliche Verteilung aller ursprünglichen $PM_{2.5}$-Messstationen in Deutschland vor der Datenbereinigung*



(b) *Räumliche Verteilung der $PM_{2.5}$-Messstationen nach Ausschluss unvollständiger Datensätze*

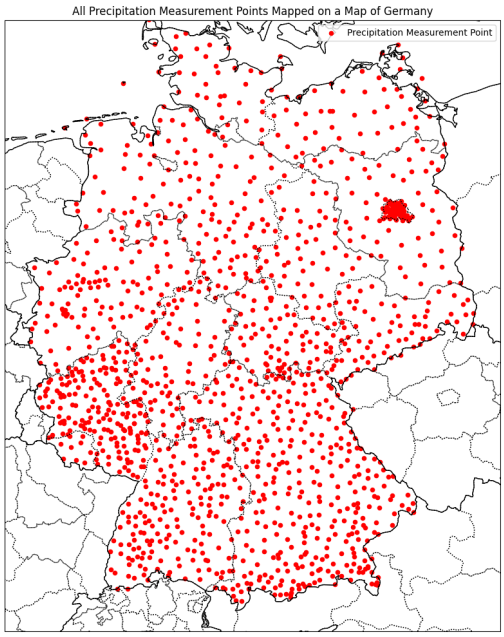
Die Abbildungen verdeutlichen, dass durch die Qualitätskontrolle eine Ausdünnung der Messtationen erfolgte, wodurch einige Regionen Deutschlands weniger dicht abgedeckt werden. Diese Einschränkung der räumlichen Repräsentativität muss bei der Interpretation der Ergebnisse berücksichtigt werden, da sich die Hypothese auf gesamt Deutschland bezieht.
Zur Effizienzsteigerung der nachfolgenden Analysen wurden als nächstes irrelevante Datenattribute entfernt. Von den in Unterabschnitt 2.1.3 detailliert aufgeführten Attributen verblieben die folgenden für die Analyse relevanten Spalten: `Station Longitude`, `Station Latitude`, `Start Datetime` und `Value`. Diese wurden zur einheitlichen Nomenklatur in `longitude`, `latitude`, `datetime` und `value` umbenannt.

Das Ergebnis dieser Präparation war ein optimierter Datensatz mit circa einer Million Einträgen.

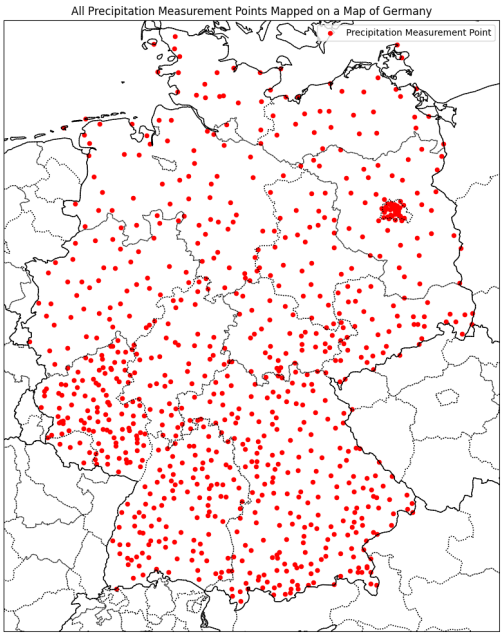
3.2.2 Aufbereitung der Niederschlagsdaten

Die Vorverarbeitung der meteorologischen Daten folgte einem analogen Vorgehen wie bei den Luftqualitätsdaten. Der umfangreiche Wetterdatensatz mit 28 Millionen Einträgen wurde zunächst eingelesen und entsprechend der Hypothese auf die Komponente `precipitation` (Niederschlagshöhe) gefiltert. Wie in Unterabschnitt 2.2.3 erläutert, werden die Niederschlagswerte in Millimetern (mm) angegeben.

Die Vollständigkeitsprüfung der Niederschlagsdaten über das Jahr 2024 ergab eine Reduktion von ursprünglich 1.361 auf 834 Stationen mit kompletten Jahresreihen. Die räumliche Verteilung der Niederschlagsmessstationen vor und nach der Bereinigung zeigen Abbildung 3.2a und Abbildung 3.2b.



(a) *Räumliche Verteilung aller ursprünglichen Niederschlagsmessstationen in Deutschland vor der Datenbereinigung*



(b) *Räumliche Verteilung der Niederschlagsmessstationen nach Ausschluss unvollständiger Datensätze*

Trotz der Reduktion der Anzahl der Stationen blieb die räumliche Verteilung der Niederschlagsmessstationen aufgrund der höheren Ausgangsdichte weitgehend repräsentativ.
Nach der Entfernung irrelevanter Attribute verblieben die Spalten `longitude`, `latitude`, `date` und `value`. Desweiteren war eine Anpassung des Zeitformats der Niederschlagsdaten nötig. Dieses wurde von der ISO 8601-Notation (2024-01-01T00:00:00+00:00) in das Datetime-Format (2024-01-01 00:00:00) konvertiert, um die Kompatibilität mit den Luftqualitätsdaten zu gewährleisten. Nach der Umbenennung ergab sich die einheitliche Spaltenstruktur: `longitude`, `latitude`, `datetime`, `value`.

Das Resultat dieser Präparation war ein Datensatz mit etwa 7,5 Millionen Einträgen.

3.3 Räumliche Zuordnung und Datenzusammenführung

3.3.1 Nearest-Neighbor-Zuordnung

Die Zusammenführung der Luftqualitäts- und Niederschlagsdaten erforderte eine räumliche Zuordnung der Messstationen, da diese nicht identische Standorte aufweisen. Für diese Aufgabe wurde ein Nearest-Neighbor-Algorithmus genutzt, der für jede $PM_{2.5}$-Messstation die geografisch nächstgelegene Niederschlagsmessstation identifiziert [11, S. 235–258].

Die technische Umsetzung erfolgte mittels der scikit-learn-Bibliothek unter Verwendung des BallTree-Algorithmus [12]. Nach der Konvertierung aller Koordinaten von Grad- in Bogenmaß wurde ein BallTree der Niederschlagsstation-Koordinaten erstellt. Der gewählte BallTree-Algorithmus eignet sich besonders für geografische Anwendungen wie die „Nearest-Neighbor" Suche [13, S. 472]. In unserem Fall wurde zur Berechnung der Distanz die Haversine-Formel verwendet.

Die Haversine-Formel berechnet die kürzeste Entfernung zwischen zwei Punkten auf der Erdoberfläche unter Annahme einer perfekten Kugel [14]. Obwohl diese Vereinfachung bei größeren Distanzen zu Ungenauigkeiten führen kann, ist sie für die räumlichen Dimensionen innerhalb Deutschlands hinreichend präzise und bietet den Vorteil hoher Recheneffizienz bei großen Datenmengen [15, S. 8143–8147].

Die Entscheidung für die Zuordnung von Niederschlagsstationen zu $PM_{2.5}$-Stationen basierte auf der geringeren Anzahl der Luftqualitätsmessstationen (97) im Vergleich zu den Niederschlagsstationen (834).

Abbildung 3.3 visualisiert die erfolgreiche räumliche Zuordnung zwischen den Mess-Stationstypen.

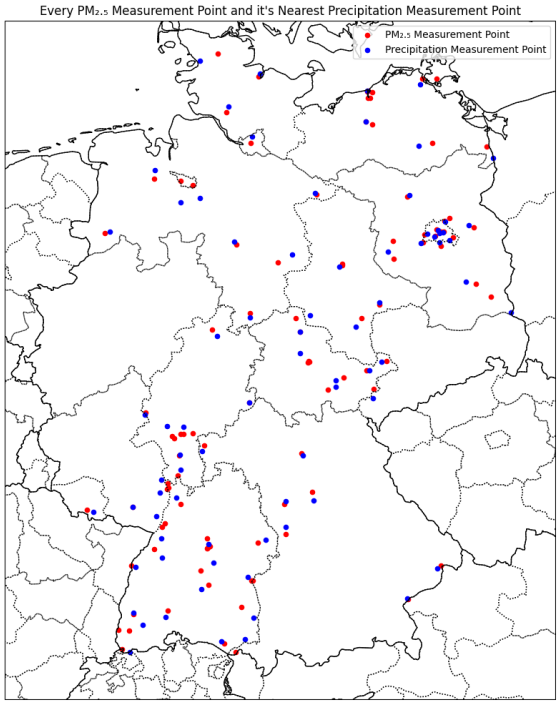


Abbildung 3.3: *Räumliche Darstellung der $PM_{2.5}$-Messstationen mit ihren zugeordneten nächstgelegenen Niederschlagsmessstationen*

3.3.2 Datenfusion und Aggregation

Die Zusammenführung der beiden Datensätze erfolgte mittels eines Inner-Joins unter Verwendung der pandas-merge-Funktion [16]. Die Verknüpfung basierte auf den Schlüsselattributen `prec_index` (Niederschlagsstation-Index) und `datetime`, wodurch zeitlich und räumlich korrespondierende Messwertepaare identifiziert wurden.

Der resultierende kombinierte Datensatz wies die Struktur `longitude`, `latitude`, `datetime`, `pm25`, `prec_index`, `prec` auf. Zur Reduzierung der zeitlichen Variabilität und zur Fokussierung auf die hypothesenrelevanten Trends wurden tägliche Aggregationen durchgeführt. Für $PM_{2.5}$-Werte wurde der tägliche Mittelwert (daily mean) berechnet, während für Niederschlagswerte die tägliche Summe (daily sum) ermittelt wurde.

Diese Aggregation resultierte in einem finalen Analysedatensatz mit den Attributen: `latitude`, `longitude`, `prec_index`, `date`, `daily_mean_pm25`, `daily_total_prec`.

3.4 Explorative Datenanalyse und Visualisierung

3.4.1 Zeitreihenanalyse

Die erste Analyse der zeitlichen Entwicklung erfolgte durch die Darstellung der aggregierten Tageswerte über das gesamte Jahr 2024.

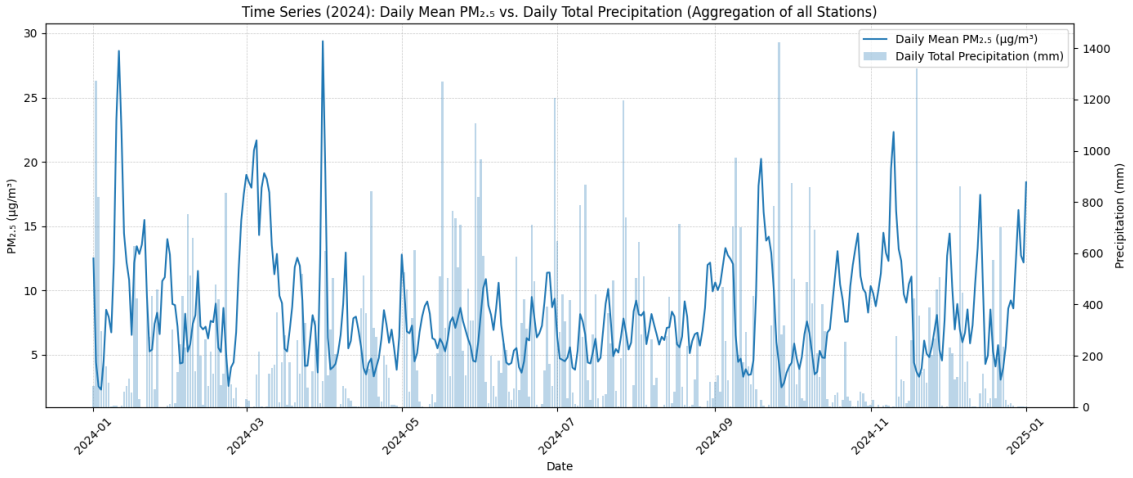


Abbildung 3.4: *Zeitreihendiagramm der täglichen $PM_{2.5}$-Mittelwerte und Niederschlagssummen für alle Stationen aggregiert über das Jahr 2024*

Die gesamtdeutsche Aggregation aller Stationen (Abbildung 3.4) zeigt deutliche saisonale Muster und lässt insbesondere in den Monaten Januar, März und November erkennbare inverse Beziehungen zwischen längeren niederschlagsarmen Perioden und erhöhten $PM_{2.5}$-Konzentrationen erkennen.
Eine Einzelstationsbetrachtung hat gezeigt, dass die räumliche Aggregation für die Erkennbarkeit der hypothetischen Zusammenhänge von Bedeutung war. Während auf Einzelstationsebene die Trends weniger ausgeprägt sind, werden sie durch die deutschlandweite Aggregation verstärkt und damit besser identifizierbar.

3.5 Statistische Korrelationsanalyse

3.5.1 Pearson- und Spearman-Korrelation

Zur quantitativen Bewertung des visuell vermuteten Zusammenhangs wurden sowohl der Pearson-Korrelationskoeffizient als auch der Spearman-Rangkorrelationskoeffizient berechnet. Die Berechnungen erfolgten mittels der scipy-Bibliothek auf Basis der aggregierten Tageswerte für `daily_mean_pm25` und `daily_total_prec` [17]. Die Ergebnisse der Korrelationsanalyse zeigen:

- Pearson-Korrelationskoeffizient: $r = -0,145$ ($p < 0,001$)
- Spearman-Rangkorrelationskoeffizient: $\rho = -0,246$ ($p < 0,001$)

Die Interpretation dieser Befunde erfolgt entsprechend den statistischen Konventionen aus [18, S. 1389]:
Der Pearson-Koeffizient weist auf einen schwachen negativen linearen Zusammenhang zwischen $PM_{2.5}$-Konzentration und Niederschlagshöhe hin. Mit steigender Niederschlagsmenge tendiert die

$PM_{2.5}$-Konzentration zur Abnahme. Die Korrelation ist statistisch signifikant ($p < 0,001$), jedoch ist die Effektstärke als gering einzustufen.
Der Spearman-Koeffizient zeigt einen schwachen bis mäßigen negativen monotonen Zusammenhang auf, der deutlich stärker ausgeprägt ist als die lineare Pearson-Korrelation ($-0,246$ vs. $-0,145$). Dies deutet darauf hin, dass der Zusammenhang zwischen Niederschlag und $PM_{2.5}$-Konzentration nicht linear verläuft. Möglicherweise zeigen bereits geringe Niederschlagsmengen überproportionale Auswascheffekte, während bei höheren Niederschlagsmengen der Effekt weniger stark zunimmt.
Diese Erkenntnisse bestätigen, dass an niederschlagsreichen Tagen tendenziell eine bessere Luftqualität (geringere $PM_{2.5}$-Werte) messbar ist. Der Effekt ist statistisch nachweisbar, jedoch nicht stark ausgeprägt, was darauf hindeutet, dass andere meteorologische Faktoren wie Wind, Temperatur oder Emissionsquellen einen größeren Einfluss auf die $PM_{2.5}$-Konzentration ausüben könnten.

3.5.2 Kontingenztafelanalyse

Zur vertiefenden Analyse wurde eine Kontingenztafelanalyse durchgeführt, die „[...] eine systematische Darstellung der Merkmalsausprägungskombinationen zweier kategorieller Merkmale [...]“ [19, S. 49] ermöglicht. Da die ursprünglichen Messdaten – wie in Quelle [19, S. 21] beschrieben – ein metrisches Skalenniveau (Verhältnisskala) aufweisen, erfolgte eine Kategorisierung beider Variablen zur Erstellung kategorieller Merkmale [20, S. 16].

Kategorisierung der $PM_{2.5}$-Werte: Die Kategorisierung basierte auf den Erkenntnissen einer WHO-Studie des Jahres 2022 [21]. Eine Analyse der Verteilungscharakteristika der $PM_{2.5}$-Daten bestätigte die aus dem WHO-Artikel hervorgehenden – für Mitteleuropa typischen – Konzentrationsbereiche von 0–30 $\mu g\,m^{-3}$ (99. Quantil: 28,92 $\mu g\,m^{-3}$, 50. Quantil: 6,96 $\mu g\,m^{-3}$).
Daraus resultierte folgende Kategorisierung:

- 0-5 $\mu g\,m^{-3}$: „gut”
- 5-15 $\mu g\,m^{-3}$: „mäßig”
- >15 $\mu g\,m^{-3}$: „schlecht”

Kategorisierung der Niederschlagswerte: Die Kategorisierung der Niederschlagsdaten erfolgte auf Basis der empirischen Verteilung (10. Quantil: 0,0 mm, 50. Quantil: 0,1 mm, 90. Quantil: 7,0 mm):

- 0-0,1 mm: „trocken”
- 0,1-7 mm: „mäßig”
- >7 mm: „stark”

Kontingenztafel:

		Niederschlag		
		trocken	mäßig	stark
$PM_{2.5}$	gut	440	4692	1515
	mäßig	903	7280	1926
	schlecht	134	633	95

Tabelle 3.1: *Kontingenztafel: $PM_{2.5}$-Kategorien vs. Niederschlag-Kategorien*

3.5.3 Chi-Quadrat-Test und Kontingenzkoeffizienten

Die statistische Bewertung der Kontingenztafel erfolgte mittels der Berechnung von Chi-Quadrat, welche einen Wert von $\chi^2 = 147,566$ ($p < 0,001$) ergab [17]. Da der Chi-Quadrat-Wert bei völliger Unabhängigkeit null beträgt, deutet das Ergebnis auf einen statistisch relevanten Zusammenhang zwischen den Variablen hin [19, S. 52].
Zur besseren Interpretierbarkeit wurden zusätzlich der Kontingenzkoeffizient und der korrigierte Kontingenzkoeffizient berechnet:

Kontingenzkoeffizient nach Pearson: $C = \sqrt{\frac{\chi^2}{n+\chi^2}} = 0,091$ [19, S. 52]

Korrigierter Kontingenzkoeffizient: $C_{korr} = \frac{C}{C_{max}} = 0,112$ mit $C_{max} = \sqrt{\frac{min(k,m)-1}{min(k,m)}}$ und $k = Zeilenzahl$, $m = Spaltenzahl$ [19, S. 53]

Der korrigierte Kontingenzkoeffizient von 0,112 bestätigt einen schwachen, aber messbaren Zusammenhang zwischen den kategoriellen Variablen [19, S. 52]. Die Analyse der Kontingenztafel zeigt, dass bei starken Niederschlägen (> 7 mm) der Anteil schlechter Luftqualität (hohe $PM_{2.5}$-Werte) deutlich sinkt, während moderate Niederschläge einen weniger ausgeprägten Effekt auf die Luftqualität zeigen als starke Niederschläge.

3.6 Fazit der Hypothesenprüfung

Die durchgeführte statistische Analyse bestätigt die aufgestellte Hypothese eines Zusammenhangs zwischen Niederschlagsereignissen und $PM_{2.5}$-Konzentrationen in der Luft im Raum Deutschland. Sowohl die korrelativen Verfahren als auch die Kontingenztafelanalyse weisen statistisch relevante, wenngleich schwache bis mäßige negative Beziehungen zwischen Niederschlagsmenge und Feinstaubkonzentration nach.
Die Ergebnisse stützen die theoretische Annahme von Auswascheffekten durch Niederschlag (Wet Deposition) [10, S. 249]. Besonders bemerkenswert ist der Befund, dass der Zusammenhang nichtlinear verläuft, was durch die stärkere Spearman- gegenüber der Pearson-Korrelation indiziert wird. Dies deutet darauf hin, dass bereits geringe Niederschlagsmengen einen überproportionalen Reinigungseffekt der Atmosphäre bewirken können.
Die Kontingenztafelanalyse unterstreicht, dass insbesondere starke Niederschlagsereignisse zu einer Reduktion schlechter Luftqualitätsepisoden führen. Allerdings zeigen die moderate Effektstärken auch, dass Niederschlag nur einen von mehreren Einflussfaktoren auf die $PM_{2.5}$-Konzentration darstellt. Andere meteorologische Parameter sowie anthropogene Emissionsquellen spielen vermutlich eine gleichwertige oder größere Rolle bei der Bestimmung der Luftqualität.
Die räumliche Einschränkung durch die Reduktion der Messstationen auf vollständige Datensätze stellt eine Limitation der Studie dar, die bei der Generalisierung der Ergebnisse auf gesamt Deutschland berücksichtigt werden muss.

Kapitel 4

Urbane Strukturen und lokale Feinstaubbelastung

4.1 Einleitung

Untersucht wird, ob und inwieweit geographische Merkmale aus OpenStreetMap (OSM) mit der lokalen Feinstaubbelastung ($PM_{2.5}$) in Berlin im Jahr 2024 zusammenhängen. Im Fokus stehen OSM-Objekte wie Parkflächen, Wohnstraßen oder Wasserflächen [5], die als Indikatoren für stadtstrukturelle Eigenschaften dienen [22, S. 5750]. Die Analyse und Modellierung räumlich verteilter Daten ist ein zentrales Werkzeug der Geowissenschaften [23, S. V - Vorwort]. $PM_{2.5}$ gilt als eines der relevantesten städtischen Luftprobleme [22, 24, 25], da Verkehr, Industrie und Heizverhalten erhöhte Konzentrationen verursachen [25, S. 7][24, S. 18]. Zwar wurde der EU-Grenzwert seit 2015 und auch 2022 eingehalten, doch der WHO-Jahresmittelwert ($5\,\mu g/m^3$) wurde an fast allen (99,5 %) der rund 200 Stationen überschritten, der Kurzzeit-Grenzwert an allen [25, S. 13 – 14]. OSM-Daten ermöglichen eine niederschwellige Erfassung potenzieller Einflussgrößen wie Bebauungsdichte, Verkehrsaufkommen oder Grünanteil und werden zunehmend zur ergänzenden Analyse der Luftqualität eingesetzt [22, S. 5750]. Zur Analyse werden zwei Regressionsmodelle, eine lineare Regression und ein Random Forest, trainiert und verglichen. Ziel ist es, zu bewerten, inwieweit stadtstrukturelle Merkmale als Prädiktoren für tägliche PM2.5-Konzentrationen geeignet sind.

4.2 Forschungsfrage und Hypothese

Hypothese (H1): Stationen mit höherer Anzahl an OSM-Objekten wie Park- oder Wohngebietsflächen (z. B. *leisure=park*, *highway=residential*) im Umkreis von 100 m bis 500 m weisen tendenziell geringere tägliche $PM_{2.5}$-Mittelwerte auf.

Nullhypothese (H0): Es besteht kein systematischer Zusammenhang zwischen OSM-Merkmalen im Umfeld und den $PM_{2.5}$-Werten.

Die Modelle werden hinsichtlich erklärter Varianz, Regressionskoeffizienten und Feature-Wichtigkeit analysiert.

4.3 Datenbeschreibung

Die Luftqualitätsdaten stammen vom Umweltbundesamt und umfassen stündliche Messwerte für $PM_{2.5}$ im Zeitraum vom 01.01.2024 bis 01.01.2025. Betrachtet werden ausschließlich Messstationen in Berlin. Nach einer automatisierten Filterung und Strukturprüfung in Python enthält der Datensatz 52.710 Zeilen mit stationären Informationen (z. B. Koordinaten, Urbanisierungsgrad) sowie $PM_{2.5}$-Messwerten in $\mu g/m^3$[26]. Für die Analyse wurde der Feinstaubparameter $PM_{2.5}$ (`Component Name = PM2`) betrachtet. Die Zielgröße ist der stündliche $PM_{2.5}$-Wert (`Value`). Zur einheitlichen räumlichen Basis wurden alle nicht-berliner Stationen entfernt. Anschließend erfolgte die Aggregation auf Tagesebene durch Mittelwertbildung je Station und Kalendertag, wie in [27] dokumentiert, um eine robustere Zielgröße für die Modellierung zu erhalten.

4.4 Explorative Datenanalyse (EDA)

Zur Beschreibung der $PM_{2.5}$-Werte wurde eine Verteilung der Tagesmittel erstellt. Die zugehörige Visualisierung (Histogramm, Monatsboxplot und Zeitreihe) wurde mit Python umgesetzt [28]. Das Histogramm zeigt eine rechtsschiefe Verteilung mit häufigen Werten zwischen 5 und 15 $\mu g/m^3$ und einzelnen Ausreißern über 60 $\mu g/m^3$ (Abbildung 4.1). Diese Verteilung der $PM_{2.5}$-Werte ist zu erwarten, da Datensätze oft überwiegend aus niedrigen $PM_{2.5}$-Werten bestehen, während nur wenige sehr hohe Werte auftreten [22, S. 5752]. Dies entspricht der Definition einer rechtsschiefen Verteilung, die eine große Häufigkeit an niedrigen und eine geringe Häufigkeit an hohen Werten aufweist [23, S. 44].

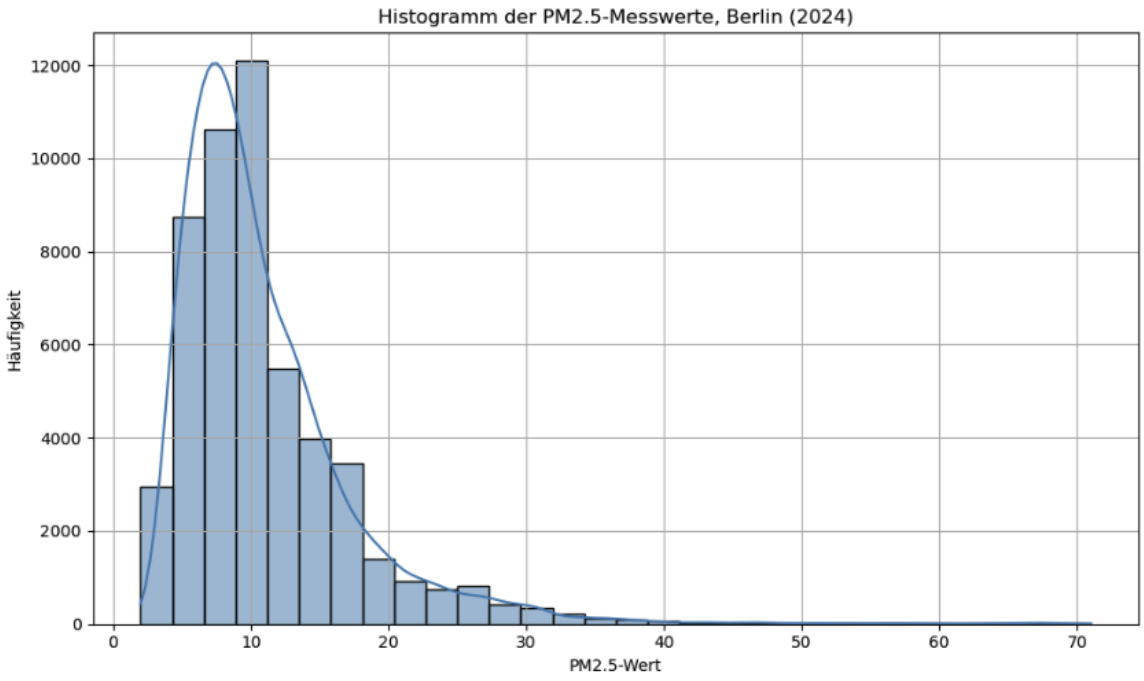


Abbildung 4.1: *Histogramm der $PM_{2.5}$-Messwerte, Berlin (2024)*

Ein Boxplot nach Monaten (Abbildung 4.2) verdeutlicht saisonale Schwankungen: Höhere Medianwerte und Streuungen treten im Winter auf, niedrigere Werte im Sommer. Dieses Muster lässt sich durch jahreszeitliche Faktoren wie Heizverhalten und Inversionswetterlagen erklären, wie auch in [29, S. 4] aufgezeigt.

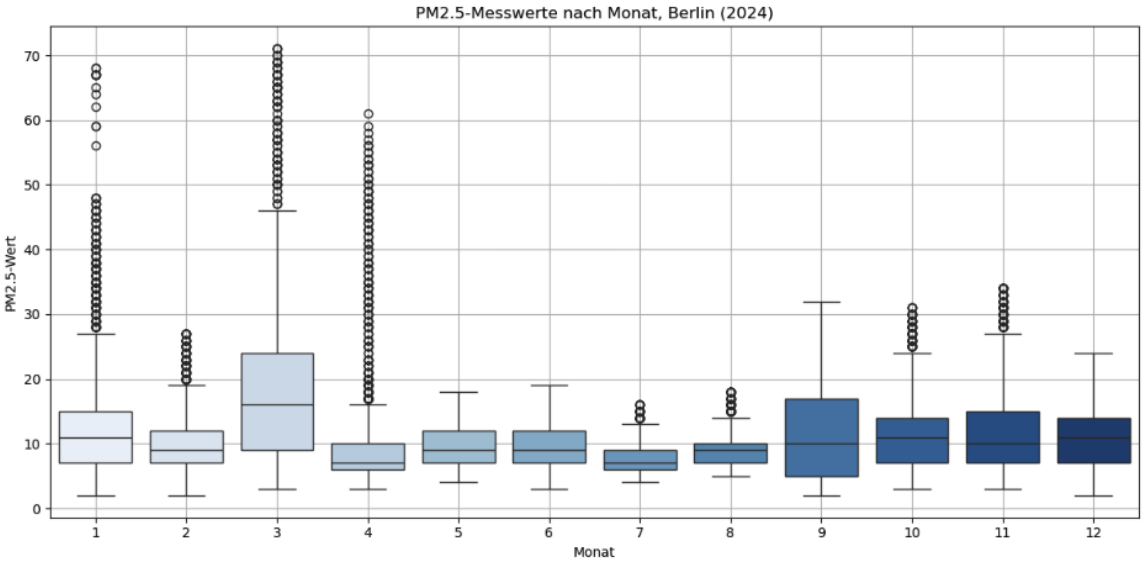


Abbildung 4.2: *$PM_{2.5}$-Messwerte nach Monat, Berlin (2024)*

Eine exemplarische Zeitreihe (Abbildung 4.3) bestätigt die tagesgenaue Dynamik mit Belastungsspitzen in den kühleren Monaten. Insgesamt deutet die EDA auf starke saisonale Schwankungen und potenzielle externe Einflussgrößen wie Wetter und Verkehr hin, die in dieser Arbeit jedoch bewusst nicht modelliert werden. Der meteorologische Einfluss wird auch bei [30, S. 9] bestätigt.

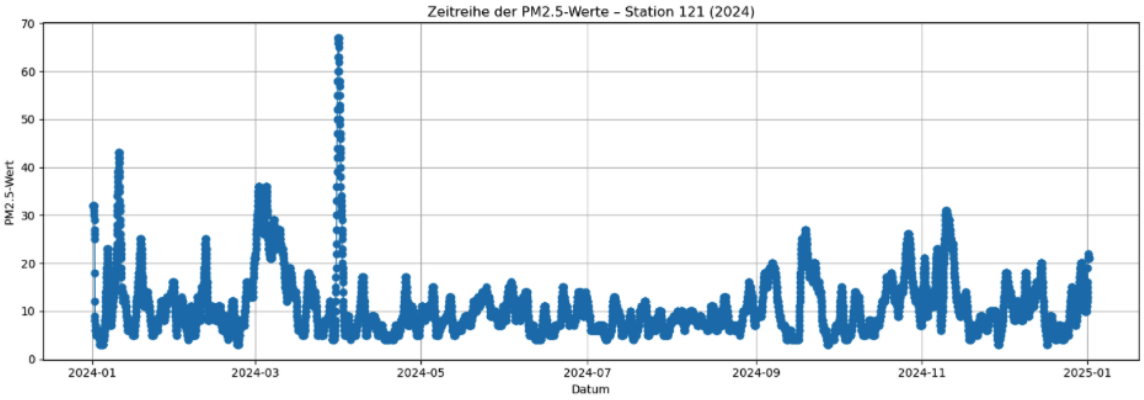


Abbildung 4.3: *Zeitreihe $PM_{2.5}$ für eine Beispielstation (2024)*

4.5 Datenbereinigung und Aggregation

Zur Glättung stündlicher Schwankungen wurden die $PM_{2.5}$-Messwerte auf Tagesmittel je Station und Kalendertag aggregiert, ähnlich wie bei [24, S. 12]. Grundlage ist der arithmetische Mittelwert der stündlichen Einzelwerte. Die aggregierten Daten wurden anschließend auf fehlende Werte, Duplikate und Werte außerhalb des Analysezeitraums geprüft [31] [32], wie bei [23, S. 38]beschrieben. Im Rahmen der Datenprüfung wurden folgende Kriterien berücksichtigt:

- **Fehlende Werte:** Keine NaNs bei Datum, Wert oder Stationsangabe.
- **Ungültige Zeitstempel:** Sechs Zeilen mit Daten außerhalb 2024 wurden entfernt.
- **Doppelte Einträge:** Keine Duplikate je Station und Tag.
- **Extremwerte:** Keine negativen oder unrealistisch hohen $PM_{2.5}$-Werte.

Der bereinigte Datensatz bildet die Grundlage für die geographische Anreicherung mit OSM-Merkmalen. Dabei wird für jede Station ein eindeutiger Tageswert der Zielgröße erzeugt.

4.6 Methodik

4.6.1 Zielvariable

Die Zielgröße ist der Tagesmittelwert der $PM_{2.5}$-Messwerte pro Station im Jahr 2024. Grundlage ist die Mittelung stündlicher Einzelwerte pro Kalendertag. Die zugrunde liegenden Messstationen wurden als eindeutige Kombinationen aus ID und Koordinaten aus dem Tagesmittel-Datensatz extrahiert [33].

4.6.2 Feature Engineering

Für jede Berliner Messstation wurden mithilfe der Overpass-API ausgewählte OSM-Merkmale im Umkreis von 100 m, 250 m und 500 m gezählt. Die betrachteten Kategorien umfassen:

- *leisure=park* (Grünflächen)
- *highway=residential* (Wohnstraßen)
- *landuse=industrial* (Industrieflächen)
- *railway=rail* (Bahninfrastruktur)
- *natural=water* (Gewässer)

Die Zählungen erfolgten automatisiert mithilfe der Abfragesprache Overpass-QL über die Python-Bibliothek `overpy`, mit der relevante OpenStreetMap-Objekte gezielt und standortbezogen extrahiert wurden [34]. Die Ergebnisse wurden anschließend in einer Feature-Matrix gespeichert. Jede Station besitzt dadurch 15 numerische Merkmale (5 Kategorien $\times$ 3 Radien).

4.6.3 Datenzusammenführung

Die OSM-Merkmale wurden über die Station-ID mit den aggregierten $PM_{2.5}$-Werten zusammengeführt. Die Verknüpfung der Daten erfolgte über einen Merge in Python [35]. Die finale Modellierungsgrundlage besteht aus:

- 15 standardisierten OSM-Features als erklärende Variablen (X)
- Tagesmittel $PM_{2.5}$ als Zielvariable (y)

4.7 Modellierung

Die folgende Modellierung basiert auf standardisierten OSM-Features (X) zur Vorhersage des $PM_{2.5}$-Tagesmittelwertes (y). Die Modellmatrix wurde durch Auswahl radiusbasierter OSM-Features erstellt [36]. Zwei Modellansätze werden verglichen: eine lineare Regression [23, S. 58 – 61] und ein Random Forest [37, S. 6].

4.7.1 Lineares Regressionsmodell

Die erklärenden OSM-Merkmale wurden standardisiert (z-Transformation) und in ein lineares Regressionsmodell überführt (Abbildung 4.4) [38]. Die Standardisierung von Merkmalen ist ein verbreiteter Schritt in der Datenvorverarbeitung. Sie transformiert die Merkmalswerte so, dass der Mittelwert der Daten Null und die Standardabweichung Eins ist. Dies ermöglicht einen leichteren und verständlicheren Vergleich zwischen Merkmalen, da jedes Merkmal gleichmäßig zum Analyseprozess beiträgt. Die z-Transformation wird explizit als Methode zur Steigerung der Modelleffektivität erwähnt [39, S. 3].

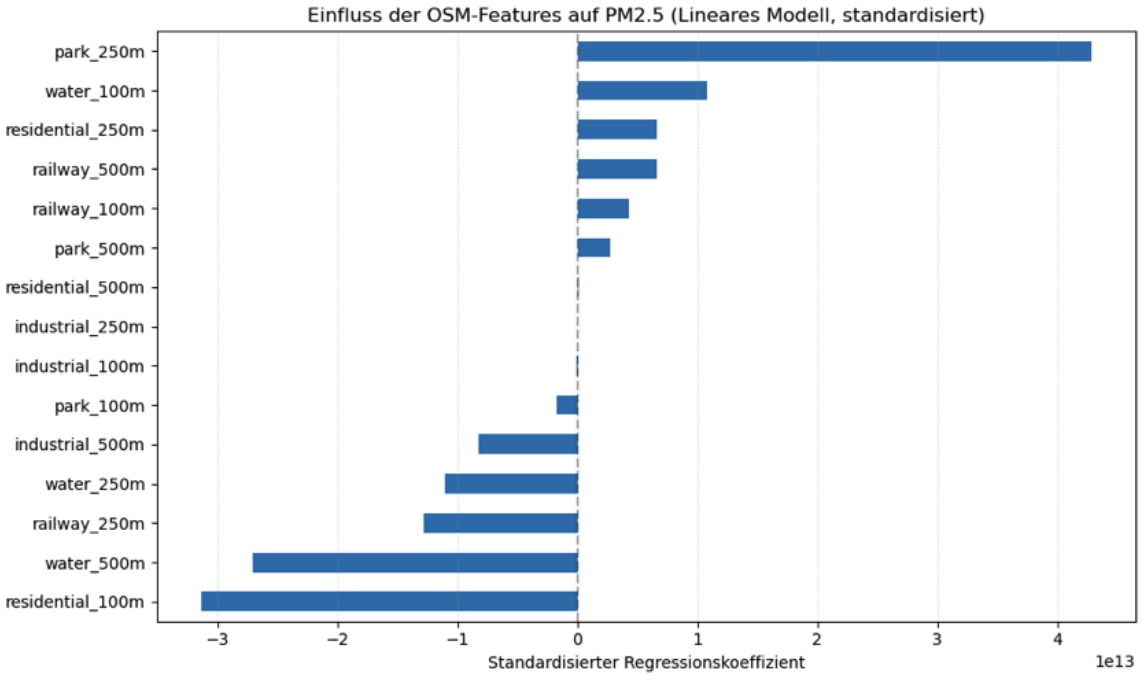


Abbildung 4.4: *Einfluss der OSM-Features auf $PM_{2.5}$ (Lineares Modell, standardisiert)*

Erklärte Varianz: $R^2 \approx 0{,}070$ **Root Mean Squared Error (RMSE):** $5{,}86\,\mu g/m^3$

Die erklärte Varianz (R^2, auch Determinationskoeffizient genannt) beschreibt, welcher Anteil der Gesamtvarianz der Zielgröße durch die erklärenden Variablen abgebildet wird. In diesem Fall erreicht das lineare Regressionsmodell einen Wert von etwa 0,070, was deutlich unterhalb der Schwelle liegt, ab der von einem erklärungsstarken Modell gesprochen werden kann [40, S. 113 – 114]. Damit bleibt der Großteil der Variation in den $PM_{2.5}$-Tageswerten unerklärt. Zur Bewertung der durchschnittlichen Prognoseabweichung dient der Root Mean Squared Error (RMSE), der in diesem Modell bei $5{,}86\,\mu g/m^3$ liegt [40, S. 420]. In Relation zur typischen Bandbreite der $PM_{2.5}$-Werte (vgl.

Abschnitt 4.4 EDA), die häufig zwischen 5 und 30 µg/m^3 liegt, weist dieser Wert auf eine begrenzte Vorhersagegenauigkeit hin. Einige Regressionskoeffizienten fallen durch ungewöhnlich hohe Beträge auf, insbesondere bei seltenen oder stark korrelierten OSM-Features wie *residential_100m* oder *park_250m*. Dies, sowie die teilweise unerwarteten Vorzeichen, weisen auf potenzielle numerische Instabilitäten durch Multikollinearität hin [40, S. 183 – 184]. Dies geschieht potenziell, weil das Modell Schwierigkeiten hat, die individuellen Effekte stark korrelierter Variablen präzise zu schätzen, was zu unzuverlässigen und instabilen Koeffizienten führt [40, S. 196].

4.7.2 Korrelationsanalyse der OSM-Merkmale

Die Korrelationsmatrix der standardisierten OSM-Features zeigt teils sehr hohe lineare Zusammenhänge zwischen ähnlichen Merkmalen in unterschiedlichen Radien (Abbildung 4.5) und bestätigt den Verdacht auf Multikollinearität. So korrelieren beispielsweise *residential_100m* und *residential_250m* mit $r = 0{,}97$, auch zwischen inhaltlich weniger verwandten Merkmalen treten stellenweise große Korrelationen auf. Berechnet wurde die Pearson-Korrelation, ein gängiges Maß für lineare Zusammenhänge [23, S. 54 – 55]. Solche starken Korrelationen können lineare Regressionsmodelle beeinträchtigen, da sie zu numerisch instabilen oder übersteigerten Koeffizienten führen können [30, S. 5] [40, S. 252]. Für zukünftige Analysen erscheint eine Merkmalsreduktion sinnvoll [41, S. 141 – 143]. Die Visualisierung erfolgte per Heatmap in Python [42].

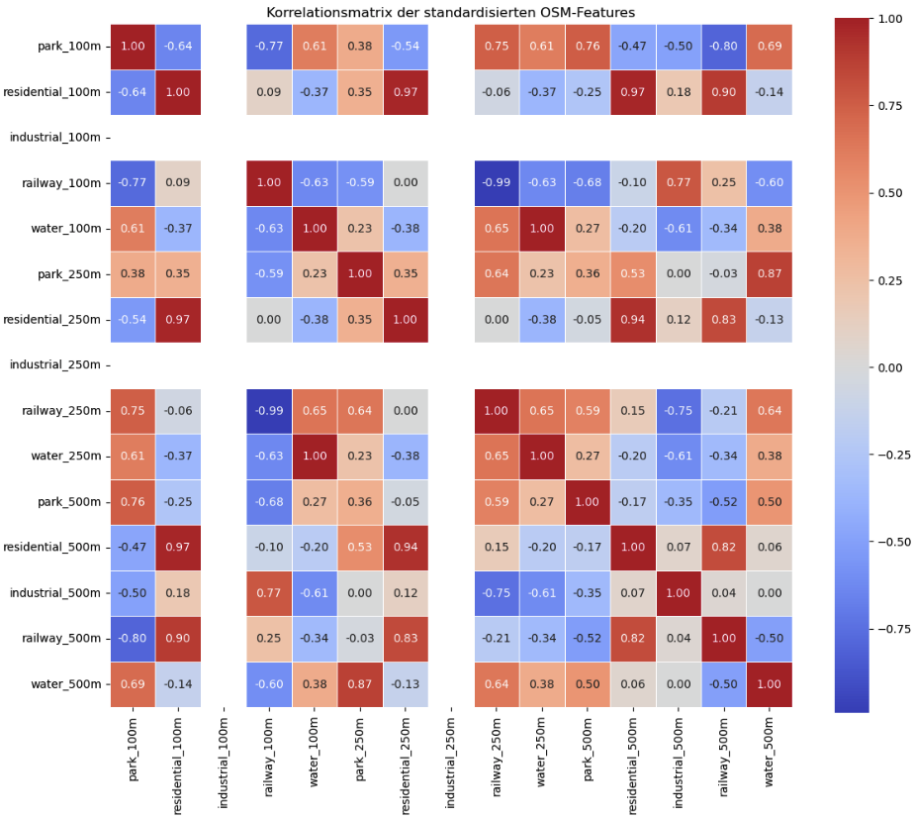


Abbildung 4.5: *Korrelationsmatrix der standardisierten OSM-Merkmale*

4.7.3 Random Forest Regressionsmodell

Zur Abbildung nichtlinearer Zusammenhänge wurde ein Random-Forest-Modell mit 100 Entscheidungsbäumen trainiert (Abbildung 4.6). Dabei handelt es sich um ein baumbasiertes Ensembleverfahren, das durch Aggregation mehrerer Entscheidungsbäume robuste Vorhersagen erzeugt [43, S. 1]. Es erlaubt eine differenzierte Bewertung der Merkmalsbedeutung anhand sogenannter Feature Importances [44, S. 9 – 11].

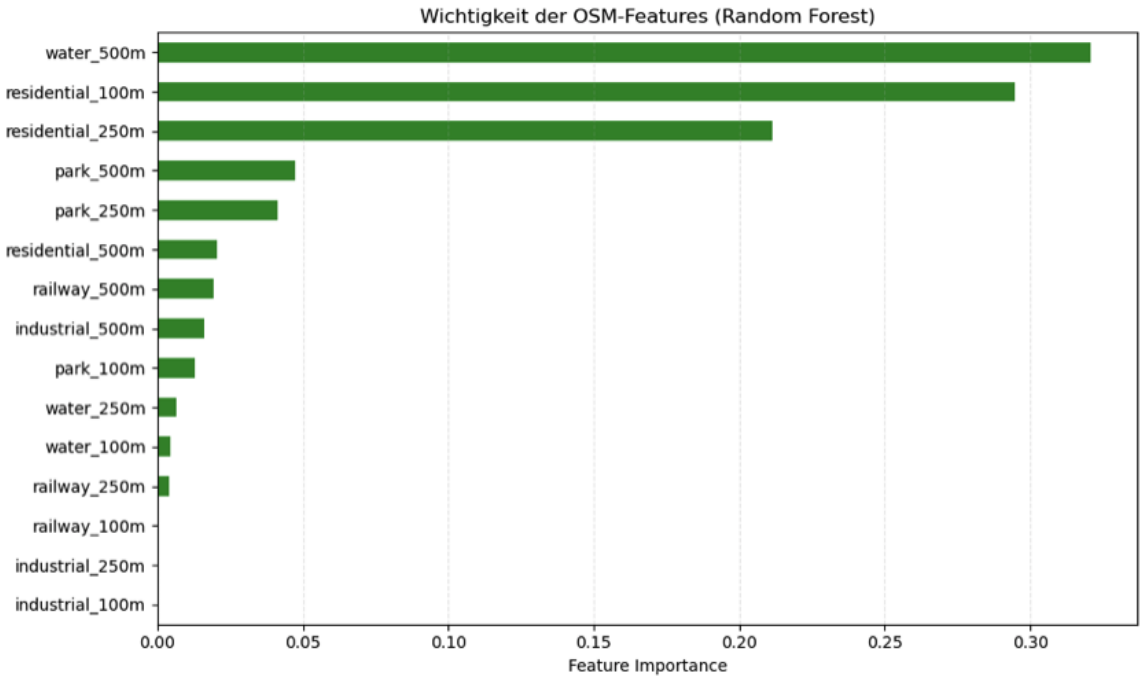


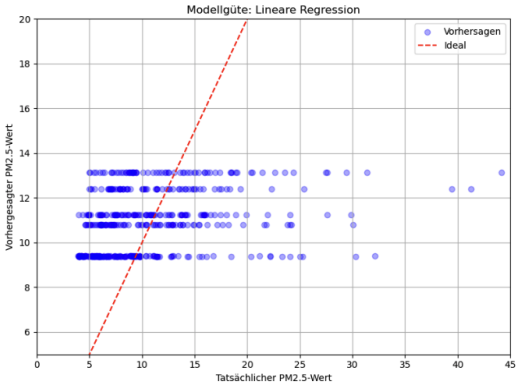
Abbildung 4.6: *Einfluss der OSM-Features auf $PM_{2.5}$ (Wichtigkeit der OSM-Features (Random Forest)*

Erklärte Varianz: $R^2 \approx 0{,}072$ **Root Mean Squared Error (RMSE):** $5{,}86\,\mu g/m^3$

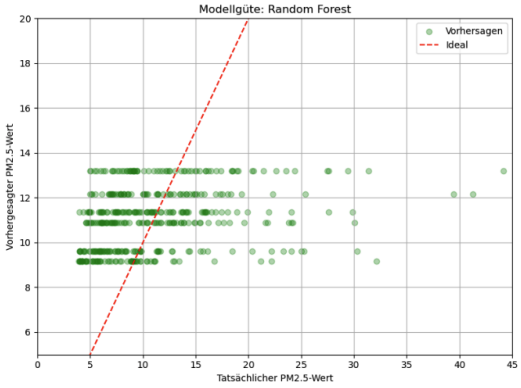
Am stärksten zur Vorhersage trugen Wasserflächen (500 m) sowie Wohnstraßen (100 – 250 m) bei. Die Merkmalsrelevanz ist plausibel und konsistent mit stadtstrukturellen Erwartungen [25, S. 7]. Das Random-Forest-Modell wurde mit dem Regressor aus dem `scikit-learn`-Paket (`sklearn.ensemble`) trainiert [45]. Die erklärte Varianz ($R^2 \approx 0{,}072$) liegt nur geringfügig über dem Wert des linearen Regressionsmodells ($R^2 \approx 0{,}070$) (vgl. Abschnitt 4.7.1 Lineares Regressionsmodell) und lässt ebenfalls auf eine nur begrenzte Modellgüte schließen. Der RMSE-Wert unterscheidet sich mit $5{,}86\,\mu g/m^3$ nicht vom linearen Modell, was darauf hindeutet, dass der Random Forest zwar in der Lage ist, nichtlineare Zusammenhänge abzubilden, diese im vorliegenden Datensatz jedoch keinen substanziellen Erklärungsmehrwert liefern.

4.7.4 Visualisierung der Modellgüte

Die folgenden Abbildungen zeigen den Zusammenhang zwischen vorhergesagten und tatsächlichen $PM_{2.5}$-Werten (Abbildung 4.7a, Abbildung 4.7b). Beide Modelle liefern nur eine begrenzte Vorhersagebandbreite (ca. 9.5 – 13.5 µg/m^3), während die tatsächlichen Werte deutlich breiter streuen. Dies deutet bei beiden Ansätzen auf Underfitting hin [41, S. 484].



(a) *Modellgüte: Tatsächlich vs. vorhergesagt (Lineare Regression)*



(b) *Modellgüte: Tatsächlich vs. vorhergesagt (Random Forest)*

4.7.5 Hyperparameterwahl und Modellkonfiguration

Das Random-Forest-Modell wurde mit Standardparametern (*n_ estimators=100*, *max_ depth=None*) trainiert. Eine gezielte Hyperparameter-Optimierung (z. B. via GridSearchCV) wurde bewusst nicht durchgeführt, um die Vergleichbarkeit der beiden Modelltypen zu wahren. Für künftige Analysen scheint diese aber vielversprechend [37, S. 13].

4.8 Fazit

Es wurde untersucht, ob geographische Merkmale aus OpenStreetMap (OSM) Rückschlüsse auf die Feinstaubbelastung ($PM_{2.5}$) in Berlin im Jahr 2024 zulassen. Dazu wurden Tagesmittelwerte von $PM_{2.5}$ mit OSM-Objektzählungen im Umkreis von 100 – 500 m um Messstationen kombiniert und in zwei Regressionsmodellen evaluiert. Beide Ansätze, lineare Regression und Random Forest, erzielten eine sehr geringe erklärte Varianz ($R^2 \approx 0{,}07$). Während das lineare Modell mit numerischer Instabilität und stark schwankenden Koeffizienten kämpfte, lieferte der Random Forest plausiblere Merkmalsbewertungen, insbesondere für Wasserflächen und Wohngebiete im weiteren Umfeld. Die Streudiagramme beider Modelle zeigen Anzeichen von Underfitting. Die Vorhersagen bleiben in einem engen Bereich, während die tatsächlichen $PM_{2.5}$-Werte stark streuen. Dies deutet darauf hin, dass OSM-Strukturdaten allein nicht ausreichen, um die $PM_{2.5}$-Belastung zuverlässig vorherzusagen. Zudem geben OSM-Zählwerte keine Auskunft über Flächenausdehnung oder Nutzungsintensität. Zentrale Einflussgrößen wie Wetter, Verkehr oder Emissionsquellen blieben unberücksichtigt. Die Analyse verdeutlicht damit die Grenzen rein geodatenbasierter Umweltmodellierung, zeigt jedoch auch das Potenzial in der Verknüpfung mit meteorologischen und verhaltensbezogenen Kontextdaten. Für zukünftige Arbeiten sind genauere Metriken wie Flächenanteile, Dichtekennzahlen, Wetter oder feingranulare Distanzgewichte vielversprechend [46, S. 13].

Kapitel 5

Fazit

Die vorliegende Arbeit hat die Luftqualität in Deutschland mittels datenwissenschaftlicher Methoden analysiert, um vielschichtige Einflussfaktoren zu quantifizieren. Es wurde ein statistisch relevanter, negativer Zusammenhang zwischen Niederschlag und $PM_{2.5}$-Konzentration nachgewiesen, der auf Auswascheffekte hindeutet, wenngleich mit schwacher bis mäßiger Effektstärke und nicht-linearem Verlauf (vgl. Kapitel 3).
Herausforderungen ergaben sich insbesondere aus der Datenqualität und -vollständigkeit, die zu einer erheblichen Reduktion der analysierbaren Messstationen führte und somit die räumliche Repräsentativität der Ergebnisse einschränkte (vgl. Unterabschnitt 3.2.1).
Die Untersuchung stadtstruktureller Merkmale auf Basis von OpenStreetMap-Daten ergab, dass einfache Objektzählungen im Umfeld von $PM_{2.5}$-Messstationen allein keine ausreichende Grundlage für belastbare Vorhersagen bieten. Beide getesteten Modelle, lineare Regression und Random Forest, erzielten nur eine sehr geringe erklärte Varianz ($R^2 \approx 0{,}07$) und zeigten Anzeichen von Underfitting (vgl. Abschnitt 4.7). Multikollinearität zwischen ähnlichen OSM-Merkmalen führte im linearen Modell zu instabilen Koeffizienten (vgl. Unterabschnitt 4.7.2). Die Ergebnisse verdeutlichen die methodischen Grenzen rein geodatenbasierter Ansätze, betonen jedoch zugleich das Potenzial in der Kombination mit meteorologischen und emissionsbezogenen Kontextdaten. Für zukünftige Analysen erscheinen flächengewichtete Merkmale, Dichtekennzahlen und Distanzmetriken vielversprechend (vgl. Abschnitt 4.8).

Literatur

[1] Deutsche Umwelthilfe e.V., *Neue Europäische Luftqualitätsrichtlinie – Umsetzung von Sauberer Luft in Deutschland*, Hintergrundpapier, Stand: 20. November 2024, 2024. Adresse: https://www.duh.de/fileadmin/user_upload/download/Pressemitteilungen/Verkehr/Luftreinhaltung/2024-11-20_DUH_Hintergrundpapier_Neue_EU-Luftqualit%C3%A4srichtlinie_final.pdf.

[2] H. Bhattarai, A. P. Tai, M. Val Martin und D. H. Yung, „Responses of fine particulate matter (PM2.5) air quality to future climate, land use, and emission changes: Insights from modeling across shared socioeconomic pathways," *Science of The Total Environment*, Jg. 948, S. 174 611, 2024, ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2024.174611. Adresse: https://www.sciencedirect.com/science/article/pii/S0048969724047600.

[3] Umweltbundesamt, *API-Dokumentation Luftdaten*, https://www.umweltbundesamt.de, 2025. Adresse: https://www.umweltbundesamt.de/daten/luft/luftdaten/doc (besucht am 05. 06. 2025).

[4] D. Wetterdienst, *Stündliche Stationsmessungen der Lufttemperatur und Luftfeuchte für Deutschland*, Version v24.03, 2024. Adresse: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/air_temperature/BESCHREIBUNG_obsgermany_climate_hourly_air_temperature_de.pdf (besucht am 22. 06. 2025).

[5] OpenStreetMap contributors. „Map Features." en. OpenStreetMap-Wiki, laufend aktualisiert. (Juni 2024), Adresse: https://wiki.openstreetmap.org/wiki/Map_features (besucht am 18. 06. 2025).

[6] *Wetter und Klima - Deutscher Wetterdienst*, en. Adresse: https://www.dwd.de/EN/ourservices/cdc/cdc_ueberblick-klimadaten_en.html (besucht am 18. 06. 2025).

[7] *Index of /climate_environment/CDC*, ftp, Deutscher Wetterdienst. Adresse: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/ (besucht am 18. 06. 2025).

[8] B. Gutzmann und A. Motl, *wetterdienst*, Python, Juni 2025. DOI: 10.5281/ZENODO.3960624. Adresse: https://wetterdienst.readthedocs.io/en/latest/usage/python-api/ (besucht am 18. 06. 2025).

[9] G. Gehricke, *Airquality*, en. Adresse: https://github.com/gxstxxv/Airquality (besucht am 18. 06. 2025).

[10] P. Zannetti, „Dry and Wet Deposition," in *Air Pollution Modeling: Theories, Computational Methods and Available Software*. Boston, MA: Springer US, 1990, S. 249–262, ISBN: 978-1-4757-4465-1. DOI: 10.1007/978-1-4757-4465-1_10. Adresse: https://doi.org/10.1007/978-1-4757-4465-1_10.

[11] R. Nisbet, J. Elder und G. Miner, „Chapter 11 - Classification," in *Handbook of Statistical Analysis and Data Mining Applications*, R. Nisbet, J. Elder und G. Miner, Hrsg., Boston: Academic Press, 2009, S. 235–258, ISBN: 978-0-12-374765-5. DOI: https://doi.org/10.1016/B978-0-12-374765-5.00011-5. Adresse: https://www.sciencedirect.com/science/article/pii/B9780123747655000115.

[12] *sklearn.neighbors - BallTree*, en, documentation, Version v1.7.0. Adresse: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.html (besucht am 18. 06. 2025).

[13] L. Zhang, G. Wang, L. Peng, W. Peng und J. Zhang, „Applying pareto frontier theory and ball tree algorithms to optimize growth boundaries for sustainable mountain cities," *Journal of Urban Management*, Jg. 14, Nr. 2, S. 468–484, 2025, ISSN: 2226-5856. DOI: https://doi.org/10.1016/j.jum.2024.11.015. Adresse: https://www.sciencedirect.com/science/article/pii/S2226585624001614.

[14] S. Kettle, *Distance on a sphere: The Haversine Formula*, en, Okt. 2017. Adresse: https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/ba-p/902128 (besucht am 18. 06. 2025).

[15] A. Baskar und M. Anthony Xavior, „A facility location model for marine applications," *Materials Today: Proceedings*, Jg. 46, S. 8143–8147, 2021, 3rd International Conference on Materials, Manufacturing and Modelling, ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.03.107. Adresse: https://www.sciencedirect.com/science/article/pii/S221478532102126X.

[16] *General functions - pandas.merge*, en, documentation. Adresse: https://pandas.pydata.org/docs/reference/api/pandas.merge.html (besucht am 18. 06. 2025).

[17] *Statistical functions - scipy.stats*, en, documentation. Adresse: https://docs.scipy.org/doc/scipy/reference/stats.html (besucht am 18. 06. 2025).

[18] R.-D. Hilgers, N. Heussen und S. Stanzel, „Korrelationskoeffizient nach Pearson," in *Lexikon der Medizinischen Laboratoriumsdiagnostik*, A. M. Gressner und T. Arndt, Hrsg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, S. 1389–1389, ISBN: 978-3-662-48986-4. DOI: 10.1007/978-3-662-48986-4_1763. Adresse: https://doi.org/10.1007/978-3-662-48986-4_1763.

[19] T. Benesch, „Statistische Maßzahlen für den Zusammenhang," in *Schlüsselkonzepte zur Statistik: die wichtigsten Methoden, Verteilungen, Tests anschaulich erklärt*. Heidelberg: Spektrum Akademischer Verlag, 2013, S. 49–80, ISBN: 978-3-8274-2772-4. DOI: 10.1007/978-3-8274-2772-4_3. Adresse: https://doi.org/10.1007/978-3-8274-2772-4_3.

[20] K. Völkl und C. Korb, „Variablen und Skalenniveaus," in *Deskriptive Statistik: Eine Einführung für Politikwissenschaftlerinnen und Politikwissenschaftler*. Wiesbaden: Springer Fachmedien Wiesbaden, 2018, S. 7–28, ISBN: 978-3-658-10675-1. DOI: 10.1007/978-3-658-10675-1_2. Adresse: https://doi.org/10.1007/978-3-658-10675-1_2.

[21] S. J. Pai, T. S. Carter, C. L. Heald und J. H. Kroll, „Updated World Health Organization Air Quality Guidelines Highlight the Importance of Non-anthropogenic PM2.5," *Environmental Science & Technology Letters*, Jg. 9, Nr. 6, S. 501–506, Juni 2022, ISSN: 2328-8930. DOI: 10.1021/acs.estlett.2c00203. Adresse: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9202349/.

[22] A. Porcheddu, V. Kolehmainen, T. Lähivaara und A. Lipponen, „Post-process correction improves the accuracy of satellite $PM_{2.5}$ retrievals," en, *Atmospheric Measurement Techniques*, Jg. 17, Nr. 19, S. 5747–5764, Mai 2024. DOI: 10.5194/amt-17-5747-2024. Adresse: https://amt.copernicus.org/articles/17/5747/2024/ (besucht am 18. 06. 2025).

[23] J. Benndorf, *Angewandte Geodatenanalyse und -Modellierung, Eine Einführung in die Geostatistik für Geowissenschaftler und Geoingenieure* (erfolgreich studieren), de. Wiesbaden: Springer Vieweg Wiesbaden, Aug. 2023, S. XII, 357, ISBN: 978-3-658-39980-1. DOI: 10.1007/978-3-658-39981-8. Adresse: https://doi.org/10.1007/978-3-658-39981-8 (besucht am 18. 06. 2025).

[24] S. K. Chaudhry, S. N. Tripathi, T. V. R. Reddy u. a., „Influence of seasonal variation on spatial distribution of $PM_{2.5}$ concentration using low-cost sensors," en, *Environmental Monitoring and Assessment*, Jg. 196, Nr. 12, 1234, Nov. 2024. DOI: 10.1007/s10661-024-13377-5. Adresse: https://doi.org/10.1007/s10661-024-13377-5 (besucht am 18. 06. 2025).

[25] S. Kessinger, A. Minkos, U. Dauert u. a., „Luftqualität 2022, Vorläufige Auswertung," de, Umweltbundesamt, Hintergrundpapier 32, Feb. 2023, S. 32. Adresse: https://www.umweltbundesamt.de/publikationen/luftqualitaet-2022 (besucht am 18. 06. 2025).

[26] M. Michaelis, *import_pm25.py – Datenimport und Filterung für Berlin*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/import_pm25.py (besucht am 23. 06. 2025).

[27] M. Michaelis, *aggregation_pm25.py – Aggregation der PM2.5-Daten auf Tagesmittel*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/aggregation_pm25.py (besucht am 23. 06. 2025).

[28] M. Michaelis, *eda_pm25.py – Explorative Datenanalyse der PM2.5-Werte*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/eda_pm25.py (besucht am 23. 06. 2025).

[29] H. Yang, Q. Peng, J. Zhou, G. Song und X. Gong, „The unidirectional causality influence of factors on $PM_{2.5}$ in Shenyang city of China," en, *Scientific Reports*, Jg. 10, Nr. 1, 8403, Mai 2020. DOI: 10.1038/s41598-020-65391-5. Adresse: https://doi.org/10.1038/s41598-020-65391-5 (besucht am 18. 06. 2025).

[30] G. T. H. Nguyen, H. Hoang-Cong und L. T. La, „Statistical Analysis for Understanding $PM_{2.5}$ Air Quality and the Impacts of COVID-19 Social Distancing in Several Provinces and Cities in Vietnam," en, *Water, Air, & Soil Pollution*, Jg. 234, Nr. 2, 85, Jan. 2023. DOI: 10.1007/s11270-023-06113-1. Adresse: https://doi.org/10.1007/s11270-023-06113-1 (besucht am 18. 06. 2025).

[31] M. Michaelis, *prüfen_pm25.py – Validierung von Datum, Duplikaten und Extremwerten*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/pruefen_pm25.py (besucht am 23. 06. 2025).

[32] M. Michaelis, *bereinigung_datum.py – Entfernen nicht plausibler Zeitstempel*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/bereinigung_datum.py (besucht am 23. 06. 2025).

[33] M. Michaelis, *stationen_pm25.py – Extraktion eindeutiger Berliner Messstationen*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/stationen_pm25.py (besucht am 23. 06. 2025).

[34] M. Michaelis, *osm_abfrage.py – Overpass-API-Zugriff zur OSM-Feature-Erfassung*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/osm_abfrage.py (besucht am 23. 06. 2025).

[35] M. Michaelis, *merge_ osm_ pm25.py – Zusammenführung von PM2.5-Daten mit OSM-Features*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/merge_osm_pm25.py (besucht am 23. 06. 2025).

[36] M. Michaelis, *modellvorbereitung.py – Auswahl und Vorbereitung der Modellvariablen*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/modellvorbereitung.py (besucht am 23. 06. 2025).

[37] Y. Özüpak, F. Alpsalaz und E. Aslan, „Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction," en, *Water, Air, & Soil Pollution*, Jg. 236, Nr. 7, 464, Mai 2025. DOI: 10.1007/s11270-025-08122-8. Adresse: https://doi.org/10.1007/s11270-025-08122-8 (besucht am 18. 06. 2025).

[38] M. Michaelis, *linear_ model_ pm25.py – Lineares Regressionsmodell für PM2.5*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/linear_model.py (besucht am 23. 06. 2025).

[39] K. Wiltos, „Impact of Feature Standardization on Classification Process Using PCA and SVM Algorithms," en, in *Information and Software Technologies*, A. Lopata, D. Gudonienė, R. Butkienė und J. Čeponis, Hrsg., Cham: Springer Nature Switzerland, 2025, S. 3–13, ISBN: 978-3-031-84263-4. DOI: 10.1007/978-3-031-84263-4_1. Adresse: https://doi.org/10.1007/978-3-031-84263-4_1 (besucht am 18. 06. 2025).

[40] H. Schneider, „Nachweis und Behandlung von Multikollinearität," de, in *Methodik der empirischen Forschung*, S. Albers, D. Klapper, U. Konradt, A. Walter und J. Wolf, Hrsg. Wiesbaden: Gabler, 2007, S. 183–198, ISBN: 978-3-8349-9121-8. DOI: 10.1007/978-3-8349-9121-8_13. Adresse: https://doi.org/10.1007/978-3-8349-9121-8_13 (besucht am 18. 06. 2025).

[41] C. Aliferis und G. Simon, „Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, G. J. Simon und C. Aliferis, Hrsg. Cham: Springer International Publishing, 2024, S. 477–524, ISBN: 978-3-031-39355-6. DOI: 10.1007/978-3-031-39355-6_10. Adresse: https://doi.org/10.1007/978-3-031-39355-6_10.

[42] M. Michaelis, *korrelation_ osm.py – Korrelationsanalyse der OSM-Merkmale*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/korrelation_osm.py (besucht am 23. 06. 2025).

[43] L. Breiman, „Random Forests," *Machine Learning*, Jg. 45, Nr. 1, S. 5–32, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. Adresse: https://doi.org/10.1023/A:1010933404324.

[44] U. Laa und F. Leisch, „Klassisches maschinelles Lernen," de, in *Moderne Verfahren der Angewandten Statistik*, J. Gertheiss, M. Schmid und M. Spindler, Hrsg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, S. 1–28, ISBN: 978-3-662-63496-7. DOI: 10.1007/978-3-662-63496-7_6-2. Adresse: https://doi.org/10.1007/978-3-662-63496-7_6-2 (besucht am 18. 06. 2025).

[45] M. Michaelis, *rf_ model.py – Random Forest Regressor zur PM2.5-Vorhersage*, en. Adresse: https://github.com/msqr95/DataScience/blob/main/rf_model.py (besucht am 23. 06. 2025).

[46] S.-R. Mehra, „Städtische Atmosphäre und Stadtklima," in *Stadtbauphysik: Grundlagen klima- und umweltgerechter Städte*. Wiesbaden: Springer Fachmedien Wiesbaden, 2021, S. 71–149, ISBN: 978-3-658-30449-2. DOI: 10.1007/978-3-658-30449-2_5. Adresse: https://doi.org/10.1007/978-3-658-30449-2_5.