

ABSCHLUSSBERICHT

Luftschadstoffprognose in Berlin
mit Machine-Learning-Methoden

Senatsverwaltung
für Mobilität, Verkehr,
Klimaschutz und Umwelt

BERLIN



IMPRESSUM

HERAUSGEBERIN

Senatsverwaltung
für Mobilität, Verkehr, Klimaschutz und Umwelt
Öffentlichkeitsarbeit
Am Köllnischen Park 3, 10179 Berlin
www.berlin.de/sen/mvku

AUFTRAGGEBERIN

Senatsverwaltung
für Mobilität, Verkehr, Klimaschutz und Umwelt
Referat Immissionsschutz

AUFTRAGNEHMERIN

INWT Statistics GmbH
Hauptstraße 8
Meisenbach Höfe, Aufgang 3a
10827 Berlin

TITELBILD

[deposithphotos/firefox](https://deposithphotos.com/firefox)

STAND

2023

INHALTSVERZEICHNIS

Einleitung	4
Begriffsklärung: „Modell“	5
Datengrundlage	6
Datenquellen	6
Datenimport, -speicherung und -validierung	7
Methodisches Vorgehen	8
Out-of-sample-Modellevaluation	8
Modellierungs-Ansatz: XGBoost	8
Von der Mess-Historie zu Prognosen für Straßenzüge	8
Modell zur Luftschadstoffprognose	9
Features	9
Zweistufiges Vorgehen.....	10
Hilfsmodell zur Verkehrsprognose	10
Kennwerte zur Beurteilung der Modellgüte.....	12
Modellgüte	13
Luftschadstoffe - zeitliche Modellgüte im Live-Betrieb	13
Luftschadstoffe - räumliche Evaluation durch Kreuzvalidierung	16
Luftschadstoffe - räumliche Evaluation durch den Abgleich mit Passivsammlerdaten	17
Verkehrsmodell - zeitliche und räumliche Modellgüte.....	20
Verkehrsmodell - räumliche Evaluation anhand von Verkehrszählungen.....	21
Erkenntnisse aus der Modellierung.....	24
Luftschadstoffmodell.....	24
Verkehrsmodell	26
Fazit.....	27
Verfügbarkeit von Ergebnissen und Informationen für die Öffentlichkeit.....	28

EINLEITUNG

Im Jahr 2022 entwickelte die INWT Statistics GmbH für die Berliner Senatsverwaltung für Mobilität, Verkehr, Klimaschutz und Umwelt ein Modellsystem, das mit Hilfe statistischer und Machine-Learning-Methoden eine Luftschadstoffprognose für Berlin berechnet.

Dabei werden stündliche Prognosen für Stickstoffdioxid (NO₂) und Feinstaub (PM_{2,5}, PM₁₀) für vier Tage in die Zukunft erstellt. Die Prognosen sind an den Schadstoff-Messcontainern, an den großen Straßenzügen und auf einem 50 × 50 Meter-Gitter in ganz Berlin sowie aggregiert auf den so genannten „Lebensweltlich orientierten Räumen (LOR)“ verfügbar.

In diesem Bericht werden die verwendeten Daten und das methodische Vorgehen vorgestellt.

Luftschadstoffe haben eine direkte gesundheitsschädigende Wirkung auf den Menschen. Sie verursachen beispielsweise Atemwegsprobleme und Herz-Kreislauf-Erkrankungen. Deswegen möchte die Berliner Senatsverwaltung für Mobilität, Verkehr, Klimaschutz und Umwelt der Bevölkerung zuverlässige Informationen zu Luftschadstoffkonzentrationen und zu Grenzwertüberschreitungen zur Verfügung stellen. Durch die Modellprognosen bekommen die Bewohnerinnen und Bewohner die Möglichkeit, Gebiete mit hoher Luftschadstoffbelastung zu meiden oder beispielsweise ihre Wohnung an der Hauptverkehrsstraße gezielt außerhalb der Rush Hour zu lüften. Außerdem ermöglicht es der Senatsverwaltung, wenige Tage im Voraus abzuschätzen, wie viel Verkehr gesundheitsverträglich ist. Dies kann in Zukunft als Grundlage für verkehrslenkende Maßnahmen dienen, um die Schadstoffbelastung zu senken.

Wichtig hierfür ist eine zuverlässige, stündlich aufgelöste Vorhersage der Konzentration von Stickstoffdioxid (NO₂) und Feinstaub PM_{2,5} und PM₁₀ in den nächsten vier Tagen in ganz Berlin. Dabei sind vor allem große Straßenzüge von Interesse, aber auch beliebige andere Orte im Berliner Stadtgebiet. Somit wird sowohl eine zeitliche als auch eine räumliche Prognose benötigt: Es werden nicht nur Schätzungen für die Zukunft benötigt, sondern auch für Orte, an denen bisher noch nie Luftschadstoffmessungen durchgeführt wurden.

Dieser Bericht beschreibt den Algorithmus, der im Jahr 2022 in einer ersten Version fertiggestellt wurde, um Luftschadstoffe mit Statistik beziehungsweise Machine Learning zu prognostizieren. Die Methoden der Statistik und des Machine Learning überschneiden sich stark, weshalb die Begriffe hier austauschbar verwendet werden. Streng genommen ist die hier verwendete Modellierungs-Methode eher dem Machine Learning zuzuschreiben, der Begriff „Statistik“ ist jedoch weit verbreitet.

Nach einer kurzen Vorbemerkung zum Begriff „Modell“ werden die Datengrundlage, das methodische Vorgehen und die Ergebnisse der Modellierung (Prognosegenauigkeit und gefundene Zusammenhänge) beschrieben.

BEGRIFFSKLÄRUNG: „MODELL“

In diesem Bericht wird oft von Modellen gesprochen.

Je nach Domäne kann dieser Begriff ganz verschiedene Bedeutungen haben. Beispielsweise können Luftschadstoffe auch mit sogenannten Chemietransportmodellen prognostiziert werden, welche chemische und physikalische Regeln mit Informationen über Emissionen und beispielsweise dem Wetter kombinieren. Analog dazu werden in der Verkehrsmodellierung verschiedene Daten und Annahmen (zum Beispiel Bevölkerungszahl in einem Gebiet; Wahl des Verkehrsmittels; Anteil der Personen, die zur Arbeit fahren) kombiniert, um abzuleiten, wie viel Verkehr an einer bestimmten Stelle zu erwarten ist. In der Wissenschaftstheorie oder für Modelleisenbahn-Fans hat der Begriff „Modell“ wiederum ganz andere Bedeutungen.

In unserem Kontext sind mit „Modell“ stets statistische beziehungsweise Machine-Learning-Modelle gemeint. In diesem Kontext ist das Modell ein Algorithmus, der aus vorhandenen Daten Zusammenhänge lernt und diese auf neue Daten anwendet, um eine Prognose zu erstellen. Zusammenhänge werden hier nicht in Form von Regeln an das Modell übergeben, sondern aus den Daten extrahiert. Um zu überprüfen, ob die Datenbasis geeignet ist und die gelernten Zusammenhänge plausibel sind, werden die Daten zunächst einer gründlichen Datenvalidierung unterzogen und die vom Modell gefundenen Zusammenhänge im Detail untersucht.

DATENGRUNDLAGE

In diesem Abschnitt wird beschrieben, welche Daten als Basis für das Modell dienen und wie diese importiert und für die Modellierung vorbereitet werden.

Datenquellen

Um ein statistisches Modell zu trainieren, das eine zeitliche sowie räumliche Prognose der Luftschadstoffe ermöglicht, sind grundsätzlich drei Gruppen von Daten notwendig:

- beobachtete Luftschadstoffwerte aus der Vergangenheit
- Features¹ mit zeitlicher Variabilität, zum Beispiel Verkehrsaufkommen, Temperatur, Windgeschwindigkeit, Uhrzeit
- Features mit räumlicher Variabilität, zum Beispiel Landnutzung (Wohnfläche, Grünfläche, ...) und – ebenfalls – Verkehrsaufkommen

Die folgende Liste führt alle verwendeten Daten auf (Links zu den Quellen weiter unten in der Tabelle):

- Luftschadstoffmessungen an den Containern des Berliner Luftgüte-Messnetzes
- MOSMIX-Wettervorhersagen des Deutschen Wetterdienstes (DWD):
 - beobachtete Werte für vergangene Zeitpunkte
 - Wetterprognosen für die nächsten 5 Tage
 - beides in räumliche Auflösung: $0.05^\circ \times 0.05^\circ$ Länge-Breite-Gitter
- Ensemble-Prognosen der Luftschadstoffe des Copernicus-Konsortiums (Copernicus Atmosphere Monitoring Service, CAMS):
 - letzten drei Jahre: Prognosen von CAMS-Europe (jeweils 5 Tage in die Zukunft; räumliche Auflösung: $0.1^\circ \times 0.1^\circ$ Länge-Breite-Gitter)
 - Zeitpunkte, die weiter als drei Jahre vor dem Beginn des Datenimports liegen (circa Juni 2019): Prognosen von CAMS-Global (jeweils 4 Tage in die Zukunft; räumliche Auflösung: circa $0.35^\circ \times 0.35^\circ$ Länge-Breite-Gitter)
- Verkehrsdaten: Daten des Messquerschnitts aus der Verkehrsdetektion in Berlin
- Ferienzeiten (von schulferien.org) und Feiertage
- Stadtstrukturdaten (räumlich auf 50×50 Meter-Zellen aufgelöst):
 - Baumbestände (in Parks, an Straßen)
 - Flächennutzung (zum Beispiel Wald, Wasser, Wohnbebauung, Infrastruktur)
 - Emissionen (räumlich, nicht zeitlich aufgelöst)
 - Bezirke und lebensweltlich orientierte Räume (LOR)
 - Bebauungsdichte und -höhe
 - Details zu den Luftschadstoff-Messstationen
 - Straßendetailnetz
 - Verkehrsmengenkarte
- Auf Basis von Datum und Uhrzeit berechnete Hilfsfeatures wie Wochentag, Stunde des Tages, Tag im Jahr

Im späteren Modelltraining und der Prognoseerstellung wird mit einem 50×50 Meter-Gitter gearbeitet. Bei Datenquellen mit geringerer Auflösung (zum Beispiel CAMS) wird für jede 50×50 Meter-Zellen jeweils der am nächsten gelegene verfügbare Datenpunkt verwendet.

¹ Features werden auch oft als „erklärende Variablen“ oder „unabhängige Variablen“ bezeichnet, im Gegensatz zur Zielgröße, die „Zielvariable“ oder „abhängige Variable“ genannt wird.

Die folgende Tabelle listet die Quellen der Daten (abgerufen am 05. Dezember 2023) auf:

Daten	Quelle
Luftschadstoffmessungen	https://luftdaten.berlin.de/lqi
Wettervorhersagen des DWD	https://brightsky.dev/docs
CAMS-Prognosen	https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-forecasts?tab=form und https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-atmospheric-composition-forecasts?tab=form
Verkehrsdaten	https://api.viz.berlin.de/daten/verkehrsdetektion
Ferienzeiten	https://www.schulferien.org
Stadtstrukturdaten	<ul style="list-style-type: none"> – Flächennutzung: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k06_02_1nutz_vegbestand2020@senstadt – Emissionen: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=wmsk_03_12_2emissionen@senstadt – Bezirke und LORs: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_alkis_bezirke@senstadt und https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_lor_2021@senstadt – Bebauungsdichte und -höhe: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k06_09_01gfz2015@senstadt und https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_06_10_1gebhoehen@senstadt – Details zu den Luftschadstoff-Messstationen: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=messpunkte@senstadt – Straßennetz: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_vms_detailnetz_wms_spatial@senstadt – Verkehrsmengenkarte: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_vmen-gen2019@senstadt

Datenimport, -speicherung und -validierung

Die Daten werden je nach Quelle zwischen stündlich (Wetterdaten, Luftschadstoffmessungen) bis einmalig (Stadtstrukturdaten) importiert. Die regelmäßigen Datenimporte werden über sogenannte Jobs mit der Software Kubernetes (<https://kubernetes.io>) gesteuert.

Im Rahmen des Datenimports erfolgen mehrere Verarbeitungsschritte, zum Beispiel die Aggregation der sehr hochauflösten Stadtstrukturdaten auf das 50 × 50 Meter-Gitter und das Füllen kleiner Datenlücken mittels statistischer Methoden. Die Daten werden in einer Datenbank abgelegt und stehen dort für die weitere Verwendung zur Verfügung. Hierfür wird die Software Clickhouse (<https://clickhouse.com>) verwendet.

Außerdem wurden alle Daten zu Beginn einer ausführlichen Datenvalidierung unterzogen, um Besonderheiten oder gegebenenfalls Fehler in den Daten zu identifizieren. Dabei wurden keine größeren Probleme festgestellt, sodass keine Datenquelle von vorn herein ausgeschlossen werden musste. Zusätzlich zur manuellen Datenvalidierung erfolgen bei jedem Datenimport automatische Überprüfungen, zum Beispiel um die Vollständigkeit zu überprüfen. Bei Problemen wird INWT durch einen Benachrichtigungsdienst informiert.

METHODISCHES VORGEHEN

In diesem Abschnitt wird beschrieben, wie mit Machine-Learning-Methoden ein Modell entwickelt wurde, welches die Luftschadstoffe vorhersagt. Dazu werden zunächst einige Grundbegriffe erläutert. Anschließend wird das Modell zur Luftschadstoffprognose konkret beschrieben. Außerdem wird das Hilfsmodell zur Verkehrsprognose beschrieben, welches die prognostizierte Verkehrsmenge als Eingangsfeature für das Luftschadstoffmodell liefert. Im letzten Unterkapitel werden als Vorbereitung auf die Ergebnisbetrachtung die Kennwerte zur Beurteilung der Modellgüte eingeführt.

Out-of-sample-Modellevaluation

Um mit statistischen Modellierungsansätzen Prognosen zu erstellen, werden zunächst die Zusammenhänge auf sogenannten Trainingsdaten erlernt. Das bezeichnet man als Modelltraining. Die Trainingsdaten bestehen in diesem Projekt aus den historischen Messwerten der Messcontainer und **weiteren oben erläuterten Quellen** wie der Wetterprognose. Beispielsweise könnte ein Modell aus den Daten lernen, dass die Luftschadstoffbelastung bei stärkerem Wind tendenziell geringer ist.

Um im Voraus zu ermitteln, wie gut das Modell im Betrieb funktionieren wird, muss das Modell auf zuvor ungesehenen Daten evaluiert werden, sogenannten Testdaten. Dazu kann man etwa den letzten Monat oder das letzte Jahr der Daten aus dem Modelltraining ausschließen, für diesen Zeitraum Prognosen erstellen und dann mit den beobachteten Werten vergleichen. Da die Berechnung der Modellgüte nicht auf derselben Daten-Stichprobe erfolgt wie das Modelltraining, spricht man dann von einer Out-of-Sample-Validierung. Alle in diesem Dokument berichteten Modellgüte-Werte basieren auf Out-of-Sample-Berechnungen.

Modellierungs-Ansatz: XGBoost

Zur Prognose der Luftschadstoffe wird ein sogenannter Gradient-Boosting-Ansatz verwendet. Dazu wird eine Code-Bibliothek namens XGBoost genutzt (<https://xgboost.readthedocs.io/en/stable>). Es handelt sich hierbei um die Kombination sogenannter Entscheidungs-bäume. Ein Entscheidungsbaum ist ein Lernalgorithmus, der sowohl für Klassifikations- als auch für Regressionsfragen verwendet werden kann. Die Grundidee von XGBoost ist, dass nicht ein großer Entscheidungsbaum konstruiert wird, der die Daten komplett erklären kann. Stattdessen werden viele kleine „schwache“ Entscheidungs-bäume kombiniert, um am Ende eine präzise Vorhersage zu erhalten. XGBoost kann nichtlineare Effekte und komplexe Interaktionen aus den Daten extrahieren und für die Prognose verwenden.

Von der Mess-Historie zu Prognosen für Straßenzüge

Das Modell zur Luftschadstoffprognose wird mit den Daten an den Luftschadstoff-Messcontainern trainiert. Damit sind Prognosen an den Messcontainern der erste Prognose-Schritt. Von größerem Interesse sind jedoch Prognosen an allen größeren Straßen Berlins, auch wenn dort keine Messcontainer stehen. Dazu werden als Zwischenschritt zunächst Prognosen auf einem 50 × 50 Meter-Gitter über ganz Berlin erstellt. Die räumliche Generalisierung ist dank räumlicher Informationen wie der Landnutzung und dem Verkehr in der Gitterzelle möglich. Von den Gitter-Prognosen werden Prognosen für die Straßenabschnitte abgeleitet, indem jeweils der Mittelwert aus allen Gitterzellen gebildet wird, die den Straßenabschnitt überdecken (siehe Abbildung 1).



Abbildung 1: Vorgehen bei der Luftschadstoffprognose. Das Modell wird zunächst auf Basis historischer Messzeitreihen an den Luftschadstoff-Messstationen trainiert. Zusätzlich zur der Prognose des weiteren Verlaufs an den Messstationen werden auch Prognosen auf einem 50 × 50 Meter-Gitter in ganz Berlin erstellt. Von diesen werden durch Aggregation die Prognosen für die einzelnen Straßenzüge abgeleitet.

Modell zur Luftschadstoffprognose

Zur Prognose der Luftschadstoffe wird pro Luftschadstoff ein separates XGBoost-Modell trainiert. Das Modelltraining erfolgt einmal monatlich, damit die Modelle sich an möglicherweise veränderte Wirkzusammenhänge anpassen. Nach dem Training wird das Modellobjekt in einer Datenbank abgelegt. Die Prognosen werden zweimal täglich (Gitter und Straßen) beziehungsweise stündlich (an den Messcontainern) auf Basis der neuesten verfügbaren Daten und dem vortrainierten Modellobjekt erstellt.

Features

Die folgenden Features gehen ins Modell ein:

- x- und y-Koordinate der betrachteten Gitterzelle des 50 × 50 Meter-Gitters über Berlin
- Jahr
- Wochentag
- Tag des Jahres
- Stunde
- - Sommer- versus Winterzeit
- Ferien: Sommerferien, andere Ferien oder Feiertag
- Menge der StEP-Straßenklassen² in Metern in der betrachteten Gitterzelle
- Wetter:
 - Windrichtung
 - Windgeschwindigkeit
 - Niederschlag
 - Temperatur
 - Bewölkung
 - Sonnenscheindauer
 - Luftdruck
- Landnutzung in der betrachteten Gitterzelle
 - Wasserfläche
 - Infrastruktur, zum Beispiel Gewerbe- und Industrienutzung, Ver- und Entsorgung
 - Wohnnutzung
 - Grünfläche, zum Beispiel Park, Kleingartenanlage, Ackerland, Friedhof
 - Verkehrsfläche (ohne Straßen), Plätze, vegetationsfreie Brachfläche
 - Mischnutzung
- Bebauungsdichte und -höhe
- Verkehrsmenge pro 24 Stunden in der betrachteten Gitterzelle

Zur Prognose an den Messstationen selbst werden außerdem vorangegangene Messwerte (sogenannte Lags) berücksichtigt.

² Straßenklassifizierung gemäß Berliner Stadtentwicklungsplan Mobilität und Verkehr.

Zweistufiges Vorgehen

In einem anderen Teil des Projekts wird simuliert, wie stark sich die Luftschadstoffbelastung durch eine Reduktion des Verkehrs verringern lässt. Dazu muss der Einfluss des Verkehrs von anderen damit korrelierten Features, wie beispielsweise der Uhrzeit, isoliert werden. Zu diesem Zweck werden zwei aufeinander aufbauende Modelle berechnet: In der ersten Stufe werden die Luftschadstoffe mit einem Teilmodell prognostiziert, welches nur Verkehrsfeatures, zeitlich konstante Features sowie Features, die mit dem Verkehr nicht stark korreliert sind, enthält. Dadurch wird sichergestellt, dass die gesamte Variation der Luftschadstoffbelastung, die durch den Verkehr erklärt werden kann, auch durch diesen erklärt wird. Im zweiten Schritt wird ein zweites Modell trainiert, das alle Features enthält und einen möglichst großen Anteil der noch verbleibenden Abweichungen (Residuen) erklärt (Abbildung 2). Die Prognosen, die aus der Kombination beider Modelle entstehen, sind den Prognosen aus einem einstufigen Modell mit allen Features sehr ähnlich, auch in Bezug auf die Prognosegüte.



Abbildung 2: Zweistufiges Vorgehen bei der Luftschadstoffprognose, um den Einfluss des Verkehrs von anderen Features mit zeitlicher Variation zu trennen (Details im Text).

Hilfsmodell zur Verkehrsprognose

Um mit dem statistischen Modell eine Prognose für die Luftschadstoffbelastung der nächsten vier Tage zu erstellen, müssen alle Features auch für die vier nächsten Tage verfügbar sein. Für zeitlich konstante Features wie die Landnutzung ist dies trivial. Für Wetter-Features wird wie erwähnt eine Wetterprognose verwendet. Auch die CAMS-Prognose ist für einen Zeithorizont von vier Tagen verfügbar. Der in den nächsten vier Tagen erwartete Verkehr ist jedoch zunächst unbekannt. Deswegen wurde ein Hilfsmodell entwickelt, das vorab Verkehrsprognosen für die kommenden vier Tage erstellt. Diese Prognosen dienen wiederum als Eingangsfeature für das eigentliche Luftschadstoffmodell.

Der Ansatz zur Verkehrsprognose ähnelt dem zur Luftschadstoffprognose sehr stark: Mit Hilfe eines XGBoost-Modells werden Fahrzeuganzahl und -geschwindigkeit für die nächsten 4 Tage auf dem 50 × 50 Meter-Gitter in ganz Berlin vorhergesagt. Ein Unterschied ist jedoch, dass die Daten aus den Verkehrsdetektoren nicht stundenaktuell zur Verfügung stehen, sondern erst mit mehrwöchiger Verzögerung. Dadurch ist eine Berücksichtigung der Echtzeit-Verkehrslage nicht möglich. Allerdings sind die typischen Muster des Verkehrsgeschehens sehr stabil (von einzelnen Ausnahmen wie Baustellen oder Großveranstaltungen abgesehen). Somit sind stundenaktuelle Updates nicht notwendig, um eine zuverlässige Verkehrsprognose zu generieren. Deswegen wird das Verkehrsmodell lediglich alle drei Monate mit Hilfe der jeweils aktuellen Daten neu trainiert und dann die Prognosen für die kommenden drei Monate berechnet. Außerdem werden im Verkehrsmodell keine Wettervorhersagen inkludiert, die sich in einer Vorab-Untersuchung als irrelevant für die Prognose des Verkehrs erwiesen.

Aus den Verkehrsdetektor-Informationen kann das Modell die typischen Tages- und Wochenmuster des Verkehrsaufkommens sehr gut extrahieren. Allerdings stehen die Detektoren fast ausschließlich an großen Straßen. Dadurch wird es schwierig, einzig anhand dieser Informationen eine Verkehrsprognose für eine Gitterzelle mit kleinen oder wenigen Straßen zu machen. Hier hilft jedoch die Verkehrsmengenkarte weiter: Diese enthält Informationen über das generelle Verkehrsniveau in einem Gebiet, ermittelt durch manuelle Verkehrszählungen. Durch eine Kombination aus Verkehrsmengenkarte und Straßendetailnetz wird für jede Zelle ein Skalierungsfaktor berechnet, um die Roh-Prognosen aus dem Verkehrsmodell auf das passende Niveau einer Zelle zu heben oder zu senken. Beispiel: Im Extremfall, dass in einer Zelle gar keine Straße liegt, ist der Skalierungsfaktor 0, sodass die resultierende Prognose ebenfalls 0 ist. Da die Verkehrsmengenkarte jedoch die verkehrliche Situation für das Jahr 2019 abbildet, können stärkere räumliche Veränderungen der Verkehrsmengen über diesen Ansatz nicht berücksichtigt werden.

Zusammengefasst funktioniert die Verkehrsprognose also folgendermaßen: Der typische Verlauf wird anhand eines XGBoost-Modells aus den Detektor-Daten extrahiert und die Information über das generelle Verkehrsniveau wird aus dem Straßendetailnetz und der Verkehrsmengenkarte 2019 abgeleitet. Dann werden beide Informationen kombiniert (Abbildung 3).

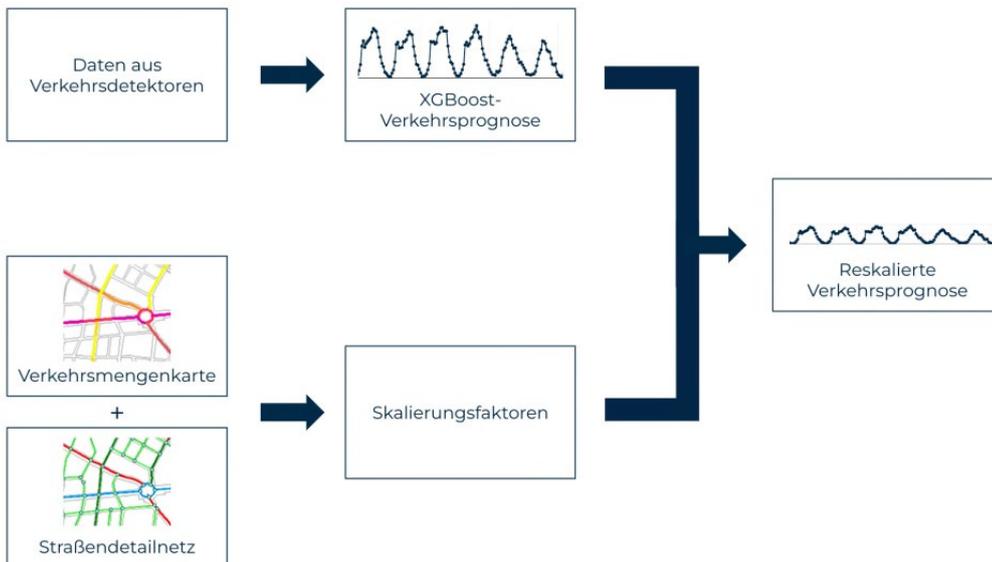


Abbildung 3: Hilfsmodell zur Verkehrsprognose. Anhand der Detektor-Daten wird mit XGBoost eine Prognose erstellt, welche vor allem die typischen Verläufe abbildet. Aus der Verkehrsmengenkarte 2019 und dem Straßendetailnetz wird ein Skalierungsfaktor abgeleitet, um das generelle Verkehrsniveau abzubilden. Dieser wird auf die XGBoost-Prognosen angewendet, um eine reskalierte Prognose zu erhalten, die sowohl das generelle Niveau als auch den typischen Verlauf berücksichtigt.

Die prognostizierten Geschwindigkeiten werden keinem weiteren Verarbeitungsschritt unterzogen.

Kennwerte zur Beurteilung der Modellgüte

Zur Beurteilung der Prognosegenauigkeit werden die folgenden Gütemaße verwendet:

- MAE: mean absolute error (mittlerer absoluter Fehler)
- RMSE: root mean squared error (Wurzel des mittleren quadratischen Fehlers)
- R^2 : R-Quadrat oder erklärte Varianz

MAE und RMSE lassen sich auf der Skala der Zielgröße interpretieren. Ein MAE von 10 bedeutet also: Die Vorhersage liegt im Mittel 10 Mikrogramm pro Kubikmeter neben dem wahren Wert. MAE und RMSE müssen also immer im Kontext des typischen Wertebereichs betrachtet werden: Ein RMSE von 5.000 Euro wäre bei der Prognose des Jahresumsatzes eines großen Unternehmens ausgesprochen gut, bei der Prognose von Monatsgehältern jedoch sehr schlecht.

Das R^2 – die erklärte Varianz – kann theoretisch zwischen 0 und 1 liegen. 1 bedeutet, dass das Modell die Daten zu 100 Prozent genau erklärt (was in der Praxis unrealistisch ist); 0 bedeutet, dass die Modellprognosen genau so gut sind, wie wenn man immer den Mittelwert vorhersagen würde. Rein rechnerisch kann das R^2 auch negative Werte annehmen (korrigiertes R^2). Dies ist dann der Fall, wenn die Vorhersagen des Modells schlechter sind, als wenn konstant der Mittelwert der Zielvariable vorhergesagt werden würde.

MODELLGÜTE

In diesem Abschnitt wird die Modellgüte, also die Genauigkeit der Prognosen, analysiert. Im **ersten Unterabschnitt** wird die zeitliche Modellgüte betrachtet, also die Genauigkeit der Prognosen an den Messcontainern in die Zukunft. Der **zweite Unterabschnitt** behandelt die räumliche Modellgüte, also die Genauigkeit von Vorhersagen an einem Ort ohne Messcontainer. Zusätzlich dazu erfolgt im **dritten Unterabschnitt** ein Abgleich mit Passivsammlerdaten. Im **vierten Unterabschnitt** wird kurz die Modellgüte des Verkehrsmodells berichtet. Eine weitere Validierung des Verkehrsmodells erfolgt im fünften Unterabschnitt durch einen Vergleich mit Verkehrszählungen.

Luftschadstoffe - zeitliche Modellgüte im Live-Betrieb

Die Prognosen werden seit Januar 2023 im Produktivbetrieb regelmäßig erstellt und gespeichert. Dadurch können sie nachträglich mit den beobachteten Messwerten verglichen werden. So lässt sich berechnen, wie präzise die Prognosen beispielsweise einen Tag oder zwei Tage im Voraus waren. Die berichteten Werte beziehen sich auf den Zeitraum vom 01. Januar bis 21. November 2023.

Abbildungen 4, 5 und 6 zeigen für alle drei Luftschadstoffe jeweils die drei oben eingeführten Modellgütemaße für die Prognosehorizonte von 1, 2, 3 und 4 Tagen. 1 Tag bedeutet dabei 0 bis unter 24 Stunden, 2 Tage bedeuten 24 bis unter 48 Stunden und so weiter. Mit steigendem Prognosehorizont sinkt die Prognosegenauigkeit erkennbar, aber nicht gravierend: MAE und RMSE werden größer, R^2 (die erklärte Varianz) nimmt ab.

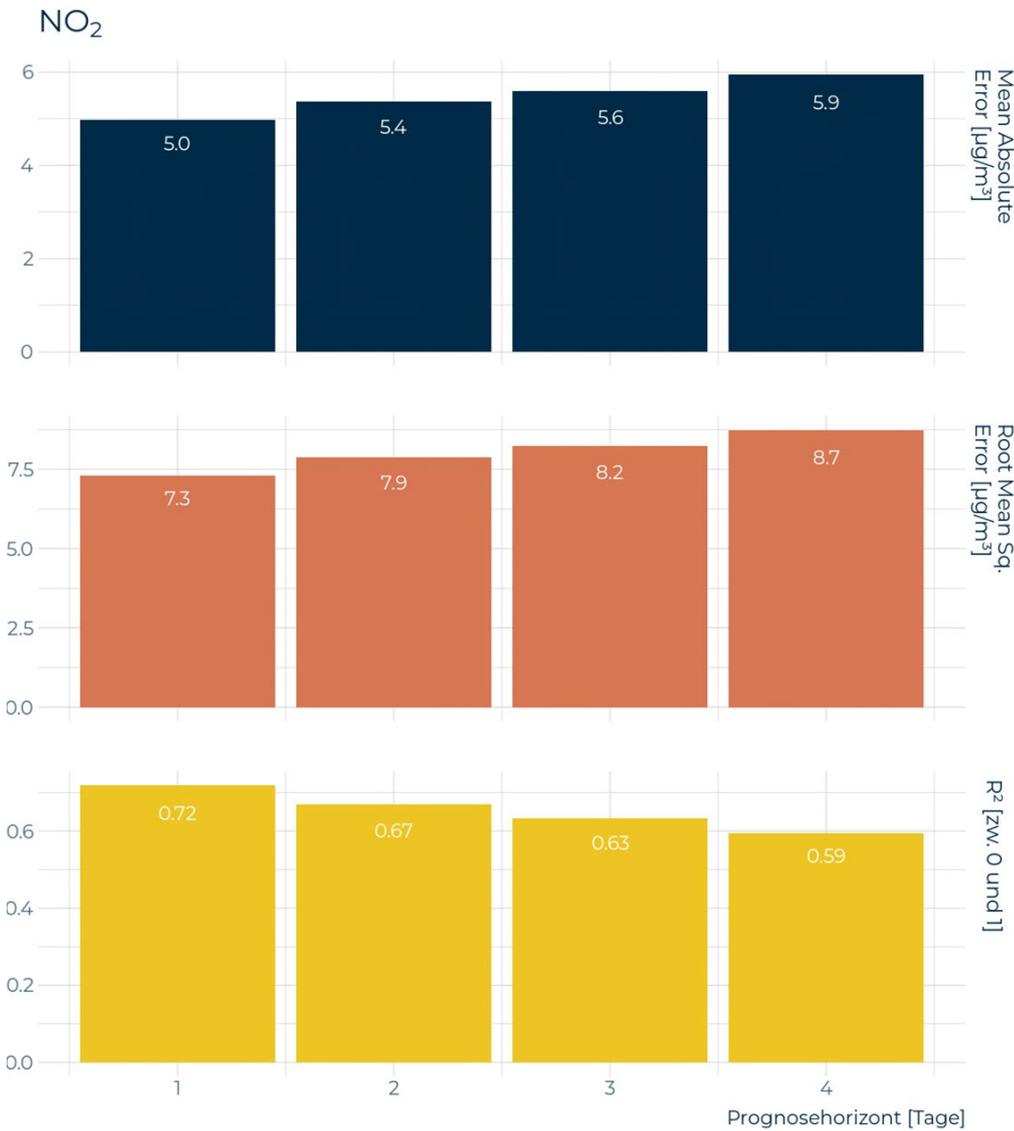


Abbildung 4: Modellgüte für NO₂.

Die beobachteten NO₂-Werte liegen im Mittel bei circa 16,5 Mikrogramm pro Kubikmeter, wobei NO₂ an Straßenmessstationen zum Teil doppelt so hoch ist wie an Stationen am Stadtrand. Die beobachteten mittleren absoluten Fehler (MAE) zwischen circa 5 und 6 Mikrogramm pro Kubikmeter sind damit als sehr gut zu beurteilen, auch im Kontext der typischerweise bei derartigen Fragestellungen erreichten Prognosegüten. Das R² beim Prognosehorizont von bis zu einem Tag ist mit circa 70 Prozent ebenfalls gut.

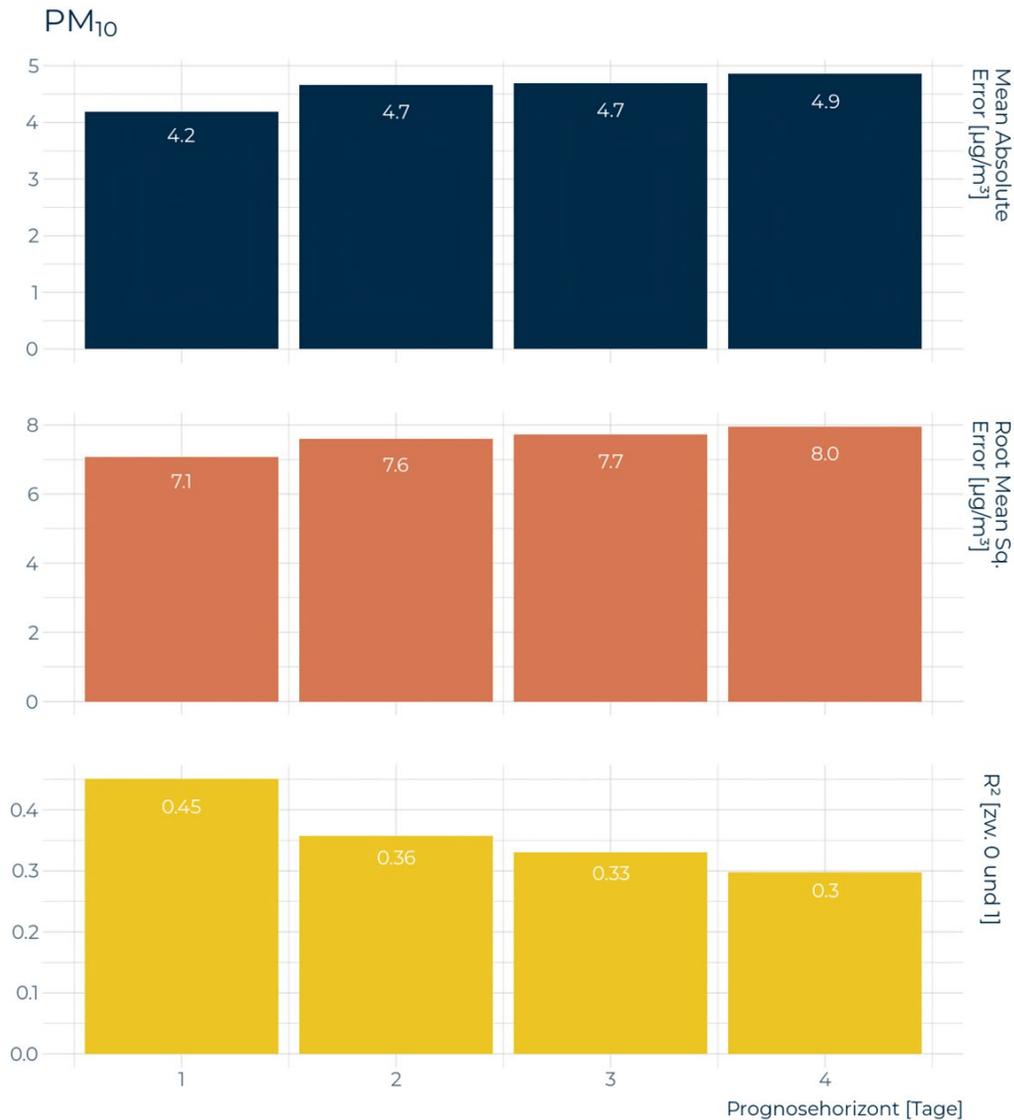


Abbildung 5: Modellgüte für PM₁₀.

Die beobachteten PM₁₀-Werte liegen im Mittel bei circa 17 Mikrogramm pro Kubikmeter bei einer wesentlich geringeren räumlichen Streuung als NO₂. Die beobachteten mittleren absoluten Fehler (MAE) zwischen circa 4 und 5 Mikrogramm pro Kubikmeter sind damit als sehr gut zu beurteilen. Das R² beim Prognosehorizont von bis zu einem Tag ist mit circa 45 Prozent deutlich geringer als für NO₂. Das liegt unter anderem daran, dass die Varianz von NO₂ viel stärker auf die - sehr regelmäßigen - Schwankungen des Verkehrs zurückzuführen ist, sodass sich für NO₂ die Varianz leichter erklären lässt. Das PM₁₀-Niveau hingegen wird stärker durch andere Emissionsquellen beeinflusst, die im Modell nicht abgebildet werden können, weil dafür keine passenden Datenquellen zur Verfügung stehen. Gerade Sonderereignisse wie Baustellen, die zu sehr hohen PM₁₀-Konzentrationen führen, können zum einen das Modell verfälschen, aber zum anderen aufgrund der fehlenden Information nicht abgebildet werden.

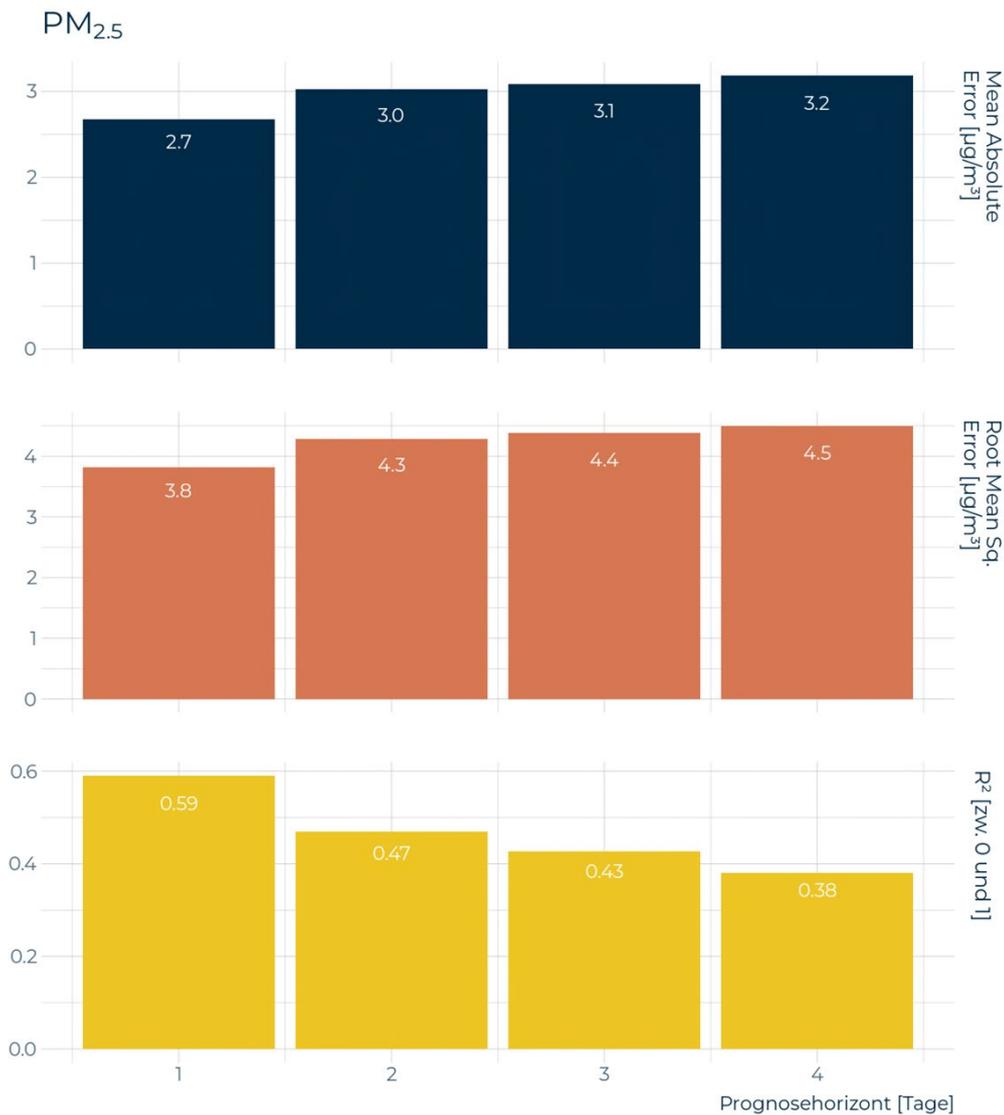


Abbildung 6: Modellgüte für PM_{2.5}.

Die beobachteten PM_{2.5}-Werte liegen im Mittel bei circa 10,5 Mikrogramm pro Kubikmeter und haben eine ähnliche, wenn auch noch geringere räumliche Varianz als PM₁₀. Die beobachteten mittleren absoluten Fehler (MAE) zwischen circa 3 und 4,5 Mikrogramm pro Kubikmeter sind damit als sehr gut zu beurteilen. Das R² beim Prognosehorizont von bis zu einem Tag ist mit circa 59 Prozent immer noch geringer als für NO₂, aber höher als für PM₁₀.

Luftschadstoffe - räumliche Evaluation durch Kreuzvalidierung

Die Prognosegüte bei der Generalisierung in den Raum wird durch eine Kreuzvalidierung ermittelt, die an die räumliche Struktur der Daten angepasst wurde: Dazu wird beim Modelltraining jeweils eine Messstation herausgelassen. Dann werden die Luftschadstoffwerte der herausgelassenen Station mit dem Modell vorhergesagt und mit den beobachteten Werten anhand der oben genannten Gütemaße verglichen. Dadurch wird künstlich so getan, als würde man Vorhersagen für eine Gitterzelle ohne Messstation machen. Dieser Vorgang wird für jede Station wiederholt und die Ergebnisse werden gemittelt.

Die folgende Tabelle enthält die räumliche Modellgüte für alle drei Luftschadstoffe.

		Gütemaß		
		MAE	RMSE	R ²
Luftschadstoff	NO ₂	7,6	11,5	0,67
	PM ₁₀	4,1	9,0	0,58
	PM _{2,5}	2,5	4,0	0,81

Auch bei der räumlichen Prognose stellt PM₁₀ die größte Herausforderung dar. Anders als bei der zeitlichen Prognose ist das Modell jedoch nicht für NO₂, sondern für PM_{2,5} am besten. Ein wesentlicher Grund für die niedrigere Genauigkeit bei der räumlichen Prognose von NO₂ ist die Messstation an der stark befahrenen Frankfurter Allee. Diese weist etwas andere Eigenschaften auf als andere Messstationen in Gebieten mit starkem Verkehr, unter anderem ist ein besserer Winddurchfluss möglich. Das schwächt den Zusammenhang zwischen Verkehrsaufkommen und NO₂-Belastung an dieser Straße ab. An dieser Messstation weicht die Zusammenhangsstruktur also von anderen Messstationen ab. Dies führt dazu, dass die Prognose von NO₂ an der Frankfurter Allee (MAE = 17,7) deutlich ungenauer ist als an allen anderen Straßen. Da die Feinstaubbelastung hingegen weniger vom Verkehr abhängt, beeinträchtigen die Ergebnisse an der Frankfurter Allee die Prognosegüte für PM₁₀ und PM_{2,5} kaum.

Auf die reale Anwendung übertragen, zeigt das, dass die räumliche Generalisierung vor allem dann funktioniert, wenn die neuen Standorte den Standorten der Messstationen relativ ähnlich sind, was die Wirkzusammenhänge betrifft. Denkbare Einflussfaktoren sind hier die Straßenbreite, die Bebauung oder die Ausrichtung der Straße (da der Wind in Deutschland am häufigsten aus Westen weht), die wesentlich bestimmen, wie stark Verkehr und Luftschadstoffbelastung zusammenhängen. Auch dem Modell unbekannt Emissionsquellen wie Industriestandorte oder Baustellentätigkeiten können dazu führen, dass die Prognosen an einem Standort ungenau sind.

Luftschadstoffe – räumliche Evaluation durch den Abgleich mit Passivsammlerdaten

Eine weitere Möglichkeit zur räumlichen Validierung des Modells für NO₂ liefern die Passivsammler: Sie messen in zweiwöchentlicher Auflösung die Luftschadstoffkonzentration an straßennahen Standorten in Berlin, an denen kein Messcontainer steht. Somit kann untersucht werden, ob die räumliche Prognose für diese Standorte mit den durch die Passivsammler gemessenen Werten übereinstimmt. Da die Passivsammlerdaten nicht für das Modelltraining verwendet werden, handelt es sich auch hier um eine Out-of-Sample-Validierung.

Für den Vergleich wurden die Jahresmittelwerte von 40 Passivsammler-Messungen für das Kalenderjahr 2022 gebildet. Außerdem wurden für alle Gitterzellen, in denen ein Passivsammler steht, die räumlichen Modellprognosen rückwirkend für 2022 erstellt und pro Zelle über das Jahr gemittelt.

Der Vergleich zeigt, dass in den meisten der 40 Zellen die Modellprognose gut mit den Messungen an den Passivsammlern übereinstimmt (Abbildung 7). In Zellen mit größeren Unterschätzungen liegt dies in der Regel an folgender Konstellation: Der Passivsammler steht in einer Gitterzelle, die in der Realität auch – gerade so – die benachbarte Straße überdeckt. Die Straße ist jedoch in den Daten nur als Linie repräsentiert, die zufällig nicht mehr die Gitterzelle des Passivsammlers kreuzt. In der Zelle des Passivsammlers sind dadurch für das Schadstoffmodell keine Straßen sichtbar, was zu einer niedrigeren Modellprognose führt (siehe Abbildung 8).

Passivsammler vs. Modellprognose

Abgleich der Jahresmittel

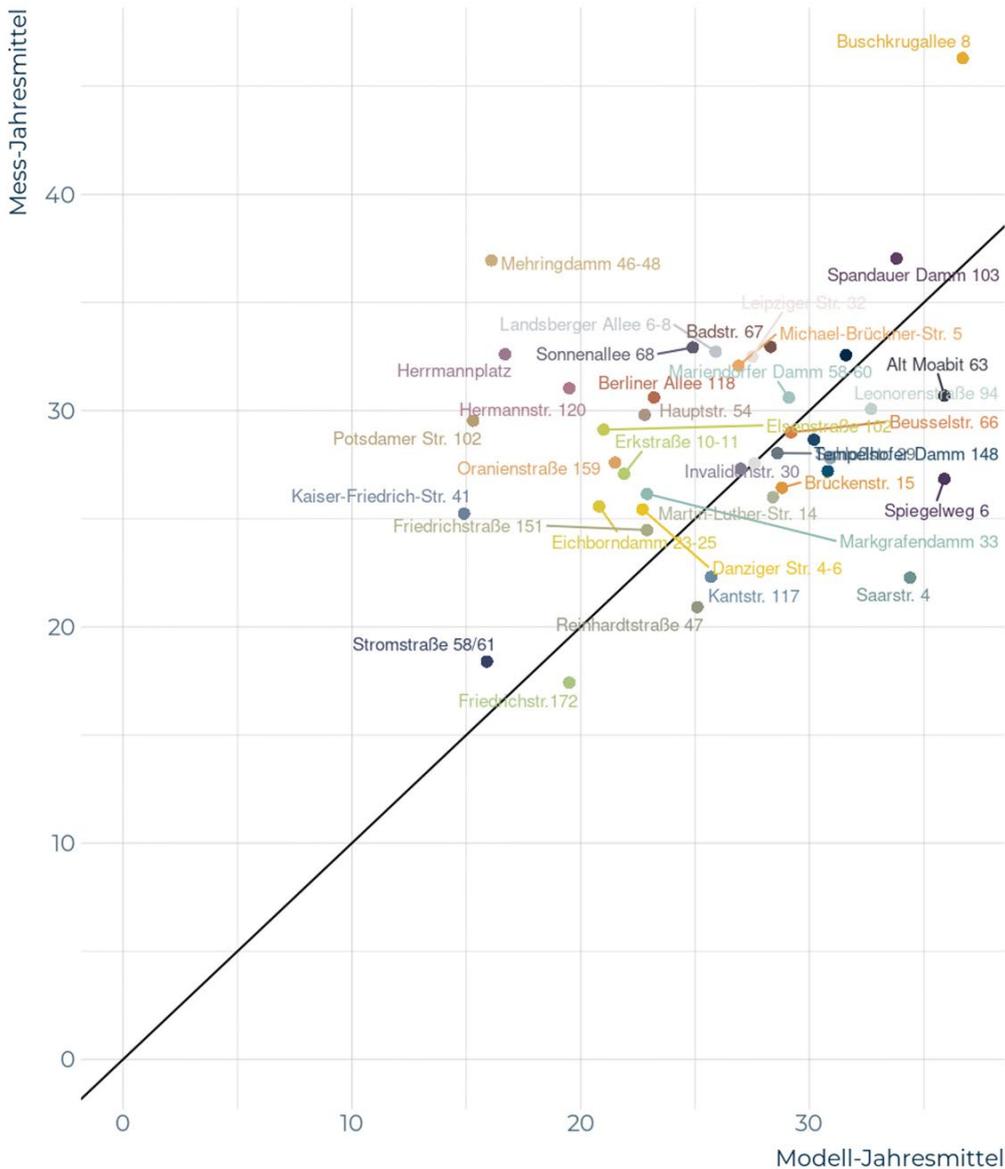


Abbildung 7: Jahresmittel der NO₂-Werte aus Modellprognosen und Passivsammlern im Vergleich. Bei perfekter Modellprognose liegt der Punkt auf der schwarzen Linie. Viele Modellprognosen liegen sehr nah an den Messungen der Passivsammler. Einige Werte werden vom Modell deutlich unterschätzt (oberer linker Bereich, zum Beispiel Mehringdamm).

Diese Konstellation besteht für den Mehringdamm, die Potsdamer Straße, die Kaiser-Friedrich-Straße, die Stromstraße und die Hermannstraße. Die Messungen am Hermannplatz und der Buschkrugallee sind aus anderen Gründen nicht repräsentativ, sodass sie zur Evaluation des Modells nur begrenzt herangezogen werden können.



Abbildung 8: Lage des Passivsammlers, der Gitterzelle und der benachbarten Straße. Links eine typische Situation: Passivsammler (violetter Punkt) und relevante Straße (blaue Linie) liegen in derselben Gitterzelle. Dadurch berücksichtigt das Modell die Straße in der Prognose und macht eine genaue Vorhersage. Rechts ein Beispiel für den Grenzfall, in dem die Zelle des Passivsammlers nicht die benachbarte Straße überdeckt. In diesem Fall ist die Modellprognose zu niedrig (dunkelblaue Färbung der Gitterzelle).

Bei einer Aggregation auf ganze Straßenzüge oder Gebiete fallen diese Ausnahmefälle kaum ins Gewicht. Sie zeigen jedoch, dass gerade für die Prognose einzelner Zellen oder kleinster Gebiete eine Berücksichtigung auch des umliegenden Verkehrs die Prognose verbessern könnte.

Abbildung 9 zeigt die Übereinstimmung der modellierten Daten mit den Passivsammler-Messungen, wenn die Modelldaten über den gesamten Straßenabschnitt zusammengefasst werden, das heißt, wenn die modellierten NO_2 -Werte aller Zellen, die den Straßenzug berühren, gemittelt werden. Die Messungen der Passivsammler spiegeln ebenfalls den repräsentativen Bereich eines gesamten Straßenzuges von mindestens 100 Metern, typischerweise von Kreuzung bis Kreuzung oder darüber hinaus, wider.

Die Übereinstimmung wird dadurch besser. Es ergibt sich ein R^2 von 0,6 bei einer Steigung von 0,65.

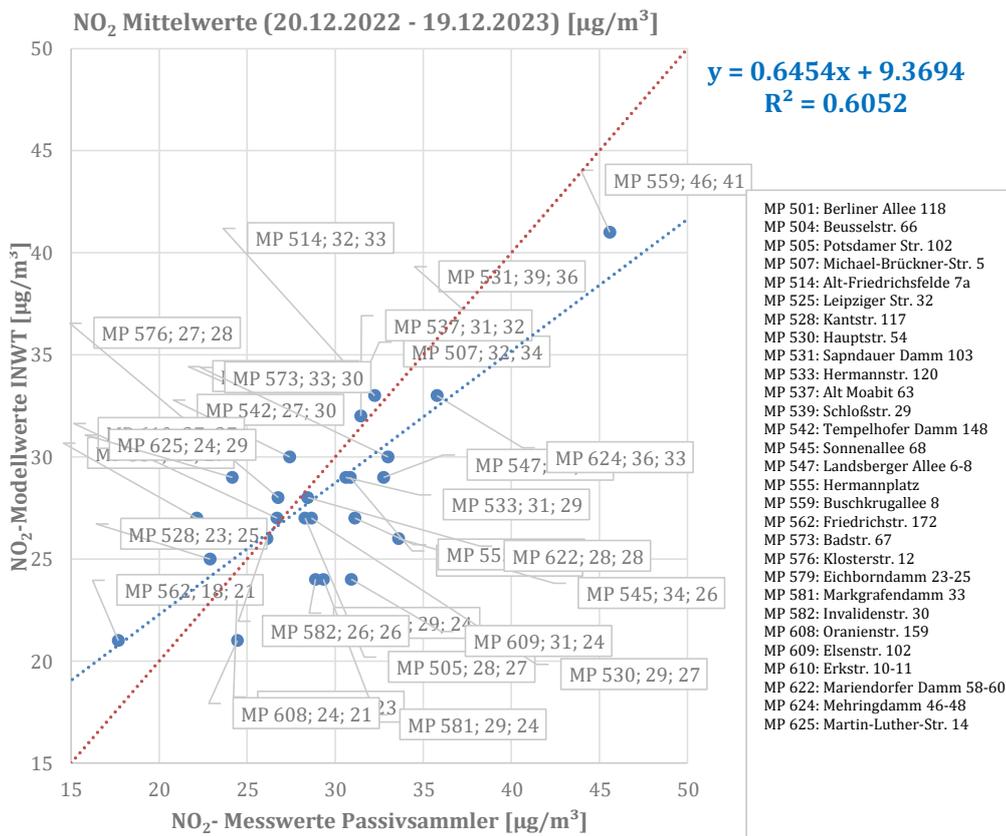


Abbildung 9: Jahresmittel der NO₂-Werte aus Modellprognosen, gemittelt über den gesamten Straßenzug von Kreuzung zu Kreuzung, und Passivsammlern im Vergleich. Bei perfekter Modellprognose liegt der Punkt auf der schwarzen Linie.

Verkehrsmodell - zeitliche und räumliche Modellgüte

Die folgende Tabelle zeigt die Modellgüte für das Verkehrsaufkommen und die Geschwindigkeit in Kilometer pro Stunde. Die zeitliche Modellgüte wird berechnet, indem aus dem Datenzeitraum am Ende 12 Wochen als Testdaten vorgehalten und das Modell auf dem davor liegenden Zeitraum trainiert wird. Die räumliche Modellgüte wird berechnet, indem die Verkehrsdetektoren zufällig in fünf Gruppen eingeteilt werden. Das Modell wird dann auf den Daten aus vier Gruppen trainiert und auf der fünften Gruppe evaluiert. Dies wird für jede Gruppe wiederholt und die Ergebnisse werden gemittelt.

	Gütemaß		
	MAE	RMSE	R ²
Verkehrsaufkommen (Mittelwert: 517 Fahrzeuge pro Stunde)			
zeitlich	113	187	0,75
räumlich	162	246	0,65
Geschwindigkeit (Mittelwert: 46 Kilometer pro Stunde)			
zeitlich	5,3	8,4	0,59
räumlich	7,0	10,0	0,15

Im Verhältnis zu den Mittelwerten ist die Modellgüte gut. Es fällt auf, dass die räumliche Generalisierung schwieriger ist als die zeitliche, insbesondere bei der Geschwindigkeit. Außerdem lässt sich das Verkehrsaufkommen genauer vorhersagen als die Geschwindigkeit, weil es stärker ausgeprägte typische Muster aufweist.

Verkehrsmodell - räumliche Evaluation anhand von Verkehrszählungen

Ein weiterer Ansatz zur Validierung des Verkehrsmodells ist der Abgleich mit Verkehrszählungen, die an ausgewählten Tagen und an ausgewählten Standorten durchgeführt werden.

Dazu stellte die Senatsverwaltung für Mobilität, Verkehr, Klimaschutz und Umwelt Daten zu Verkehrszählungen an 26 Standorten aus den Jahren 2022 und 2023 zur Verfügung. Die Zählungen dauerten jeweils einen Tag von 00:07 bis 19:00 Uhr. INWT berechnete rückwirkend die Prognosen des Verkehrsmodell für die entsprechenden Standorte und Zeiträume der Zählungen. Da sich die Verkehrszählungen auf Straßenzüge beziehen, wurden auch die Modellprognosen von Gitterzellen auf Straßenzüge aggregiert. Der Vergleich wurde mittels statistischer Kenngrößen durchgeführt. Dabei wurden sowohl die Tagessummen in dem Zeitraum der Zählungen betrachtet, als auch die Abbildung der Tagesgänge durch das Prognosemodell. Für den Abgleich wurden die anhand der Verkehrsmengenkarte skalierten Prognosen verwendet. Zwischen den Tagessummen aus dem Modell und den Verkehrszählungen zeigt sich insgesamt eine hohe Übereinstimmung (siehe Abbildung 10).

Verkehrszählung vs. Modellprognose

Abgleich der Tagessummen

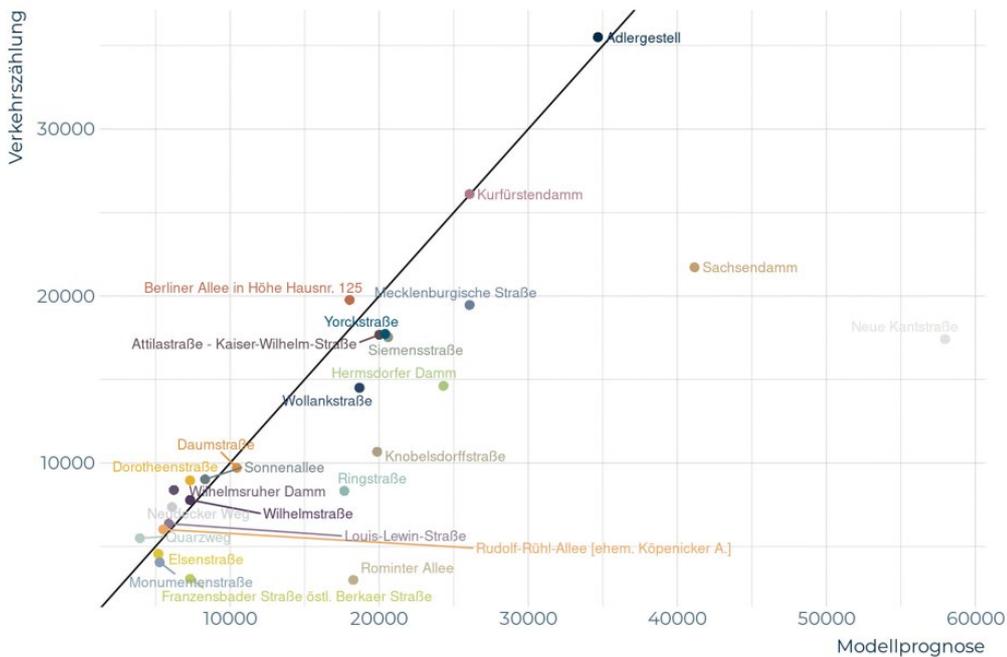


Abbildung 10: Tagessummen aus Modellprognosen und Verkehrszählungen im Vergleich. Bei perfekter Modellprognose liegt der Punkt auf der schwarzen Linie. Viele Modellprognosen liegen sehr nah an den Beobachtungen aus den Verkehrszählungen.

Besonders interessant ist jedoch die Betrachtung der weniger genauen Prognosen und der Ursachen dafür:

- Die Prognosen für die eigentlich schwach befahrenen Straßen Rominter Allee, Franzensbader Straße und Knobelsdorffstraße werden durch benachbarte größere Straßen nach oben gezogen, welche teilweise mit der entsprechenden kleineren Straße in derselben Gitterzelle liegen. Dies liegt am gewählten Vorgehen bei der Skalierung und Aggregation und ist im Kontext der Luftschadstoffprognose gewünscht, weil in diesem Fall größere Straßen von höherem Interesse sind als kleine. Für eine möglichst genaue Prognose des Verkehrsaufkommens müsste das Vorgehen angepasst werden.

- An der Neuen Kantstraße, am Sachsendamm und am Hermsdorfer Damm zeigt sich, dass die Verkehrszählungen deutlich von der Verkehrsmengenkarte (2019) abweichen. Da die Verkehrsmengenkarte zur Skalierung der Prognosen verwendet wird, weichen diese in der Folge stark von der neueren Zählung ab. Bei genauerer Betrachtung liegt die Ursache zwischen der Abweichung der Verkehrsmengenkarte und der Verkehrszählung daran, dass es Änderungen am Straßenabschnitt gab. Beispielsweise wurde bei der Neuen Kantstraße eine Fahrspur zu einer Radspur umgewidmet. Ein Update der Verkehrsmengenkarte würde die Abweichung zwischen der Verkehrsprognose und der Verkehrszählung lösen.
- Bei der Ringstraße lässt sich keine eindeutige Ursache identifizieren.

Die vom Modell prognostizierten Tagesgänge stimmen ebenfalls in den meisten Fällen mit den Verkehrszählungen überein. In 14 von 26 Fällen ist eine hohe Korrelation von über 0,7 zu beobachten. Nur in zwei Fällen ist die Korrelation kleiner als 0,2. Das Modell kann auch unterschiedliche Tagesgänge gut abbilden, beispielsweise den höheren Peak am Nachmittag am Kurfürstendamm (Abbildung 11) oder den über den Tag ansteigenden Trend an der Wilhelmstraße (Abbildung 12).

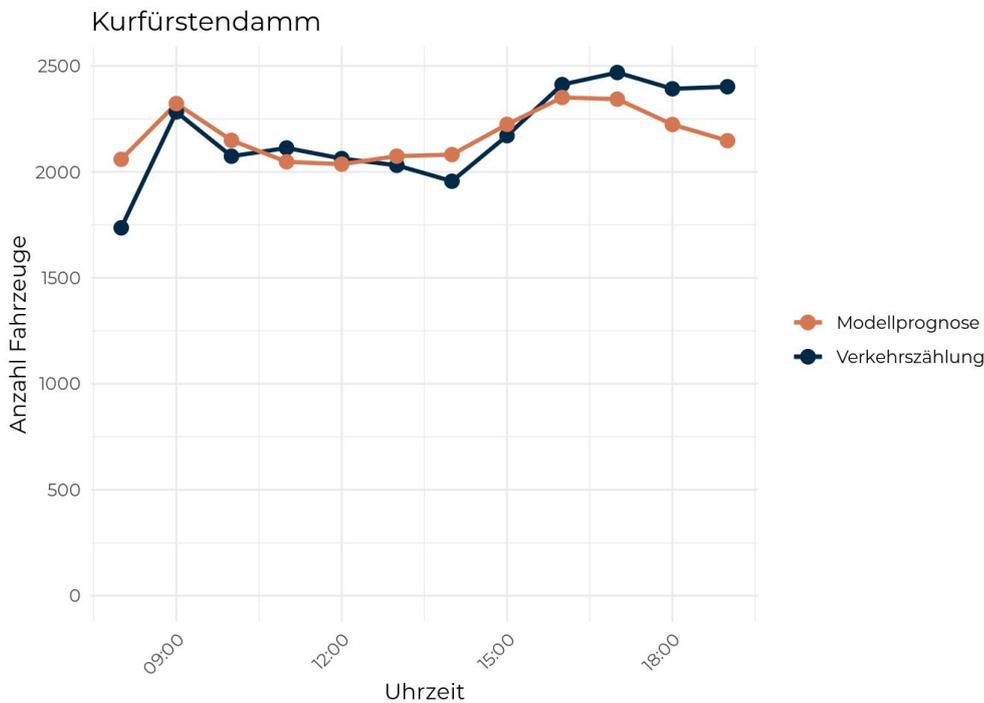


Abbildung 11: Die zwei durch die Rush Hour bedingten Peaks am Kurfürstendamm werden vom Modell prognostiziert, wobei der vorhergesagte Nachmittags-Peak in der Verkehrszählung länger andauert.

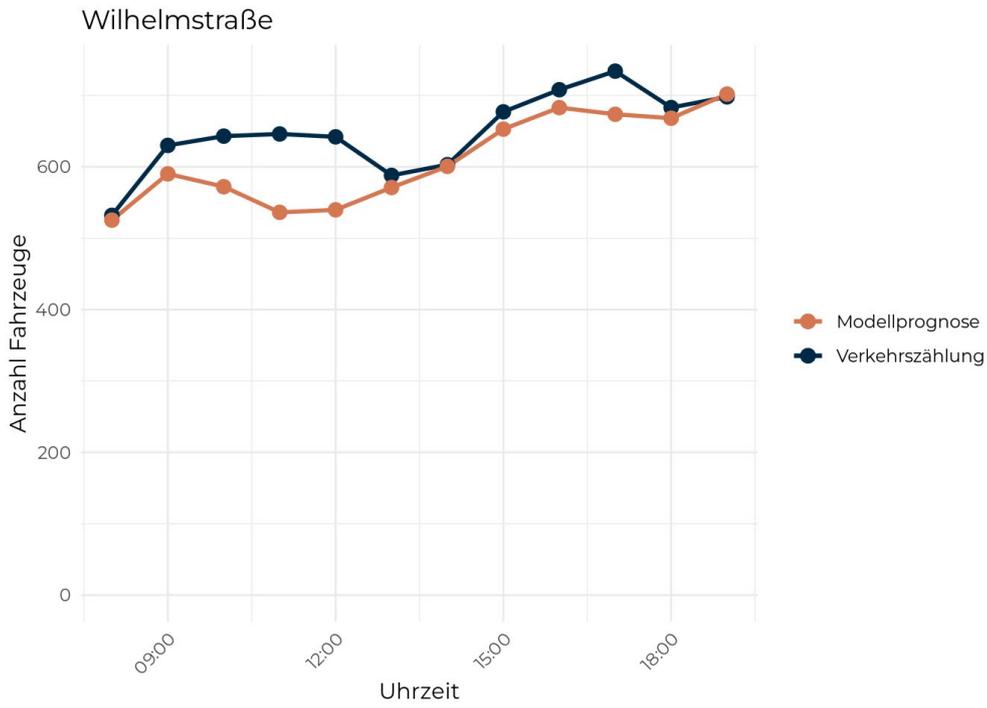


Abbildung 12: Das über den Tag stetig steigende Verkehrsaufkommen wird vom Modell korrekt prognostiziert.

ERKENNTNISSE AUS DER MODELLIERUNG

Nachdem im vorherigen Abschnitt die allgemeine Modellgüte betrachtet wurde, werden die von den Modellen gefundenen Wirkzusammenhänge im Folgenden genauer beleuchtet. Dies generiert nicht nur interessante Erkenntnisse, sondern ermöglicht auch eine Prüfung der Plausibilität der Modelle. Dazu wird untersucht, welche Features vom Modell als besonders wichtig identifiziert wurden und ob die gefundenen Wirkzusammenhänge mit vorhandenem Wissen im Einklang stehen.

Luftschadstoffmodell

Abbildungen 13, 14 und 15 zeigen den prozentualen Beitrag der zehn wichtigsten Features für die drei Luftschadstoffe zur Prognose an den Messstationen in den nächsten 24 Stunden (normiert am R^2 für einen Prognosehorizont von einem Tag). Es handelt sich um die kombinierten Beiträge aus beiden Modellen des **oben erläuterten zweistufigen Vorgehens**.

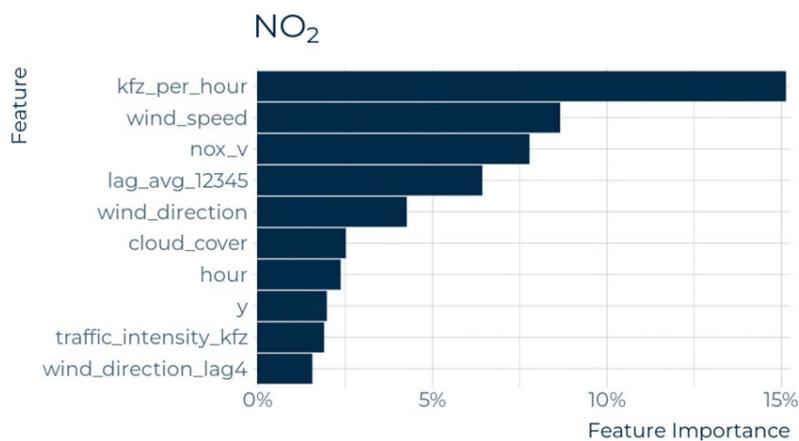


Abbildung 13: Die wichtigsten Features zur NO₂-Prognose: Besonders entscheidend für die NO₂-Belastung sind das Verkehrsaufkommen, die Windgeschwindigkeit und die räumlich (nicht zeitlich) aufgelöste Stickoxid-Belastung (Stickstoffmonoxid und -dioxid) durch Verkehr aus früheren Messungen.



Abbildung 14: Wichtigste Features zur PM₁₀-Prognose: Neben den Konzentrationen der vorangegangenen Stunden haben Windrichtung (aktuell und in den letzten Stunden), Windgeschwindigkeit und das Verkehrsaufkommen den größten Einfluss.

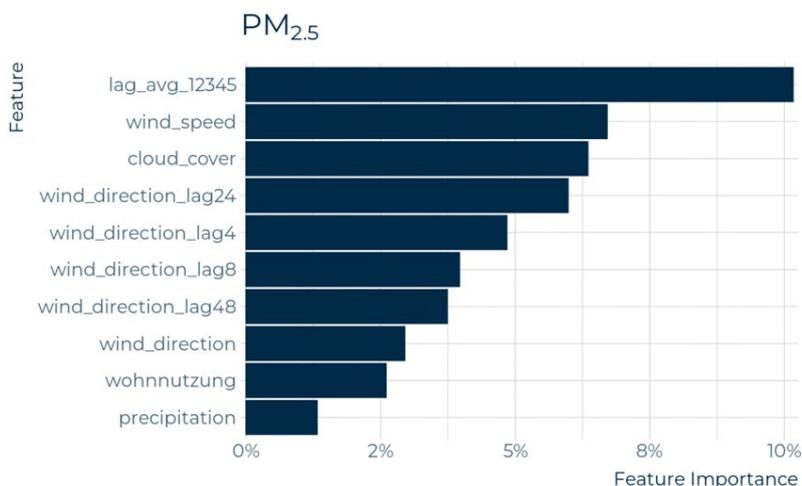


Abbildung 15: Wichtigste Features zur PM_{2,5}-Prognose: Neben den Konzentrationen der vorangegangenen Stunden haben Windrichtung (aktuell und in den letzten Stunden), Windgeschwindigkeit und Bewölkung den größten Einfluss.

Dabei zeigen sich unter anderem die folgenden Muster:

- Das Verkehrsaufkommen (kfz_per_hour) ist für die NO₂-Prognose äußerst wichtig; auch für PM₁₀ ist es ein wichtiges Feature, für PM_{2,5} nicht.
Die Windgeschwindigkeit (wind_speed) befindet sich für alle drei Luftschadstoffe unter den wichtigsten Features.
- Die durchschnittliche Belastung mit dem jeweiligen Luftschadstoff in den vorangegangenen fünf Stunden (lag_avg_12345) ist für die Prognose ebenfalls sehr relevant.
- Die Windrichtung – aktuell, sowie beispielsweise vor 4, 8 oder 24 Stunden – spielt vor allem für die Feinstaub-Prognose eine wichtige Rolle, trägt aber auch zur NO₂-Prognose substantiell bei.
- Die Bewölkung beeinflusst die Prognosen aller drei Luftschadstoffe wesentlich.
- Der Niederschlag ist nur für die Prognose von Feinstaub relevant.

Zur räumlichen Prognose (Gittermodell) sind im Wesentlichen dieselben Features relevant. Statt der – im Gitter unbekanntenen – Werte aus den vorangegangenen Stunden wird hier ein gewichtetes Mittel der Messstationen verwendet.

Diese Beobachtungen stehen im Einklang mit den Erwartungen. Die Relevanz der Features allein reicht jedoch zur Plausibilitätsprüfung nicht aus. Auch die Form der gefundenen Zusammenhänge ist von Interesse. Beispielsweise ist zwischen Windstärke und Luftschadstoffniveau ein negativer Zusammenhang zu erwarten, weil der Wind Luftschadstoffe weg transportiert und auf diese Weise die Luftqualität verbessert. Eine genaue Betrachtung der Zusammenhänge zeigt ebenfalls plausible Ergebnisse:

- Ein höheres Verkehrsaufkommen geht mit höherer Luftschadstoffbelastung einher, wobei der Zusammenhang für NO₂ am stärksten und für PM_{2,5} am schwächsten ausgeprägt ist.
- Eine höhere Windgeschwindigkeit geht mit geringerer Luftschadstoffbelastung einher. Es gibt hier einen Sättigungseffekt; ab einem gewissen Punkt reduziert mehr Wind die Luftschadstoffbelastung nicht mehr weiter. Bei Feinstaub ist sogar ein leichter Anstieg bei sehr hohen Geschwindigkeiten zu beobachten, möglicherweise durch Aufwirbelung von Sedimenten.
- Die NO₂-Belastung ist bei mittleren Geschwindigkeiten (circa 35 bis 40 Kilometer pro Stunde) am höchsten. Dies sind typische Geschwindigkeiten bei dichtem Verkehr in Zonen mit erlaubter Höchstgeschwindigkeit von 50 Kilometer pro Stunde oder mehr.
- Ostwind geht mit höherer Feinstaubbelastung einher, weil kontinentale Luftmassen östlich von Berlin höher belastet sind als maritime Luftmassen aus dem Westen.

Verkehrsmodell

Abbildung 16 zeigt die Relevanz der Features zur Prognose des Verkehrsaufkommens. Mit Abstand am wichtigsten ist die Uhrzeit (hour). Auch das generelle Verkehrsniveau (traffic_volume) spielt eine wichtige Rolle. Erst mit Abstand folgen der Wochentag (day_of_week), die Koordinaten x und y, verschiedene Features zur Landnutzung und die Menge verschiedener Straßenklassen in der betrachteten Zelle.

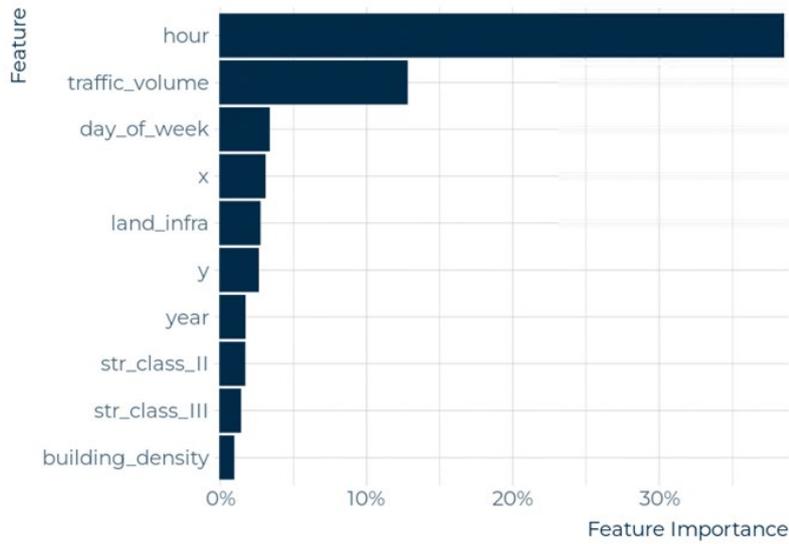


Abbildung 16: Wichtigste Features zur Prognose des Verkehrsaufkommens.

FAZIT

Insgesamt zeigt sich, dass der gewählte Machine-Learning-Ansatz sehr gut geeignet ist, um NO_2 -, PM_{10} - und $\text{PM}_{2,5}$ -Konzentrationen vorherzusagen. Sowohl eine Prognose in die Zukunft als auch die räumliche Generalisierung auf Standorte ohne Luftschadstoff-Messstation bieten in den meisten Fällen eine hohe Genauigkeit.

Bei der räumlichen Prognose ist die Modellgenauigkeit vor allem dann hoch, wenn die Zusammenhänge an den neuen Standorten ähnlich sind wie die an den Messstationen. Ein Gegenbeispiel ist hier die Frankfurter Allee, deren NO_2 -Belastung durch ihre speziellen Eigenschaften nicht gut durch ein Modell vorhergesagt lässt, das die Zusammenhänge an den anderen Messstationen gelernt hat.

Eine detaillierte Analyse zeigt außerdem, dass die XGBoost-Modelle die relevanten Features korrekt identifizieren und plausible Zusammenhänge in den Daten finden, die mit vorhandenem Domänenwissen im Einklang stehen.

VERFÜGBARKEIT VON ERGEBNISSEN UND INFORMATIONEN FÜR DIE ÖFFENTLICHKEIT

Alle Prognosen sind für die Öffentlichkeit über eine API verfügbar unter <https://api.fairq.inwt-statistics.de/docs>.

Der Programmcode kann unter <https://github.com/fairqberlin> eingesehen werden.

Ein Artikel mit technischen Details zur Automatisierung der Prognosen ist unter https://www.inwt-statistics.de/blog/business_case_luftschadstoffprognose verfügbar.

Eine Podcast-Folge über das Projekt wurde unter <https://inwt.podbean.com/e/12-use-case-luftschadstoffprognose-fur-berlin/> veröffentlicht.

Senatsverwaltung
für Mobilität, Verkehr,
Klimaschutz und Umwelt

BERLIN



Öffentlichkeitsarbeit
Am Kölnischen Park 3
10179 Berlin

www.berlin.de/sen/mvku

 x.com/senmvkuberlin

 instagram.com/senmvkuberlin

 youtube.com/@senmvkuberlin

 linkedin.com/company/senmvku

Berlin, 08/2024