INWT Statistics

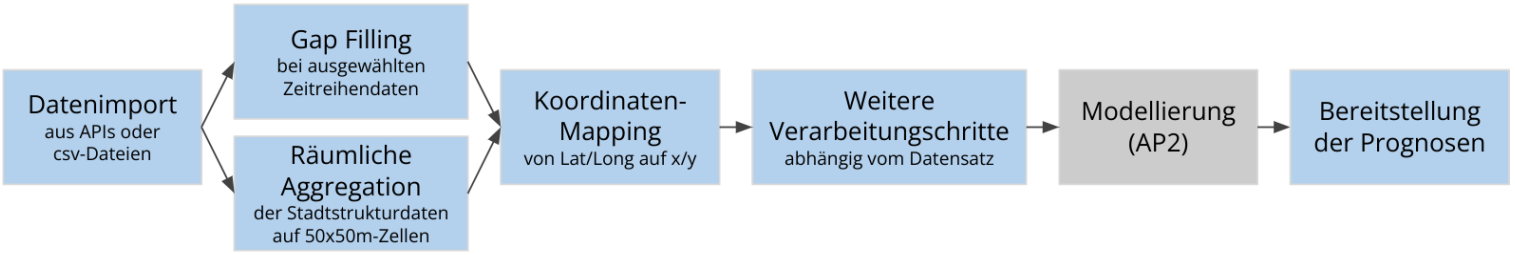
# AP3 - Datenflüsse

## Abschlussbericht

*In diesem Bericht wird erläutert, wie das Management der Datenflüsse im Rahmen des Projekts zur Luftschadstoffprognose umgesetzt wurde. Zunächst werden die Datenquellen und die Häufigkeit des Imports aufgeführt. Dann wird das sogenannte Gap Filling, das Auffüllen von Lücken in den Daten, beschrieben. Im zweiten Teil wird erläutert, an welchen Stellen die Umsetzung nach Absprache mit der SenUMVK anders erfolgte als ursprünglich geplant. Im dritten Abschnitt wird das Risikomanagement vorgestellt, also der Umgang mit Ausfällen bei Datenlieferanten sowie die Sicherstellung einer lückenlosen Prognose-Bereitstellung.*

# Datenimport und -verarbeitung

Die folgende Abbildung skizziert den allgemeinen Workflow des Datenflussmanagements. Die einzelnen Schritte - außer der Modellierung, welche Gegenstand von AP2 ist - werden in diesem Bericht genauer erläutert.




## Wo werden die Daten gespeichert?

Alle Daten werden nach Import und Aufbereitung in einer Datenbank gespeichert. Die Modellierung setzt auf diesen Daten auf; die Ergebnisse werden ihrerseits in einer Datenbank gespeichert. Die Datenbanken liegen auf Servern, die vom Cloud-Anbieter Hetzner gehostet werden. Es handelt sich um Clickhouse-Datenbanken, einem SQL-basierten Datenbanksystem, das besonders gut für große Datenmengen geeignet ist. Die SQL-Statements zum initialen Aufsetzen der Datenbanken werden gesammelt und mit dem Code hinterlegt.

Die Datenbank ist in drei sogenannte Schemata unterteilt:

- fairq-raw: Rohdaten, die nur wenigen Verarbeitungsschritten unterzogen wurden (z.B. Auswahl von Spalten)
- fairq-features: stärker aufbereitete Daten, die später in die Schadstoff-Modellierung eingehen (“Features”) bzw. Zwischenschritte, um die Daten für die Modellierung zu berechnen
- fairq-output: Modellergebnisse, die später nach außen zur Verfügung gestellt werden

Im Rahmen von AP3 sind ausschließlich die Schemata fairq-raw und fairq-features relevant.
Die Schemata fairq-features und fairq-output werden jeweils zweimal vorgehalten: Zur Entwicklung und für den Produktivbetrieb. Dadurch kann beispielsweise die Berechnung eines Features in der Entwicklungsumgebung angepasst und getestet werden, ohne den Produktivbetrieb zu beeinträchtigen.

Die Datenbanken sind aus Sicherheitsgründen nur aus dem INWT-VPN-Netz erreichbar. Perspektivisch sind regelmäßige Snapshots denkbar, die auf eine Plattform hochgeladen werden können, welche für die SenUMVK erreichbar ist. Wird von weiterer Stelle Zugriff auf die Inhalte der Datenbank benötigt (z.B. zur Visualisierung), so kann dies im Rahmen zusätzlicher Beratungsdienstleistungen über eine maßgeschneiderte API umgesetzt werden. Dies entspricht auch den Best Practices, welche den direkten Zugriff Dritter auf eine interne Datenbank nicht vorsehen.

## Welche Daten werden importiert?

Die folgenden Daten werden importiert:

- Luftschadstoffmessungen an den Containern des Berliner Luftgüte-Messnetzes
- MOSMIX-Wettervorhersagen des Deutschen Wetterdienstes (DWD):
    - Beobachtete Werte für abgeschlossene Zeitpunkte
    - Wetterprognosen für die nächsten 5 Tage
    - Räumliche Auflösung: 0.05°× 0.05° Latitude-Longitude-Gitter)
- Ensemble-Prognosen der Luftschadstoffe des Copernicus Konsortiums (Copernicus Atmosphere Monitoring Service, CAMS)
    - Letzte drei Jahre: Prognosen von CAMS-Europe (jeweils 5 Tage in die Zukunft; räumliche Auflösung: 0.1°× 0.1° Latitude-Longitude-Gitter)
    - Zeitpunkte, die weiter als 3 Jahre vor dem Beginn des Datenimports liegen (ca. Juni 2019): Prognosen von CAMS-Global (jeweils 4 Tage in die Zukunft; räumliche Auflösung: ca. 0.35°× 0.35° Latitude-Longitude-Gitter)
- Verkehrsdaten: Daten des Messquerschnittes aus der Verkehrsdetektion in Berlin
- Daten des Kraftfahrt-Bundesamtes (KBA) zur Flottenzusammensetzung
- Ferienzeiten (von schulferien.org) und Feiertage
- Stadtstrukturdaten (räumlich auf 50x50m$^2$-Zellen aufgelöst):

- Baumbestände (in Parks, an Straßen)
- Flächennutzung (z.B. Wald, Wasser, Wohnbebauung, Infrastruktur)
- Emissionen
- Bezirke und LORs
- Bebauungsdichte und -höhe
- Details zu den Schadstoff-Messstationen
- Straßennetz
- Verkehrsmengenkarte

Hinweis: In der späteren Modellschätzung und Prognoseerstellung wird mit einem 50x50m²-Gitter gearbeitet. Bei Datenquellen mit geringerer Auflösung (z.B. CAMS) wird für die 50x50m²-Zellen jeweils der am nächsten gelegene verfügbare Datenpunkt verwendet.

Die folgende Tabelle listet die Quellen der Daten 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 |
| KBA-Daten | https://www.kba.de/DE/Statistik/Produktkatalog/produkte/Fahrzeuge/fz1_b_uebersicht.html?nn=1146130 |
| Ferienzeiten | https://www.schulferien.org |
| Stadtstrukturdaten | - Baumbestände: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_wfs_baumbestand@senstadt und |

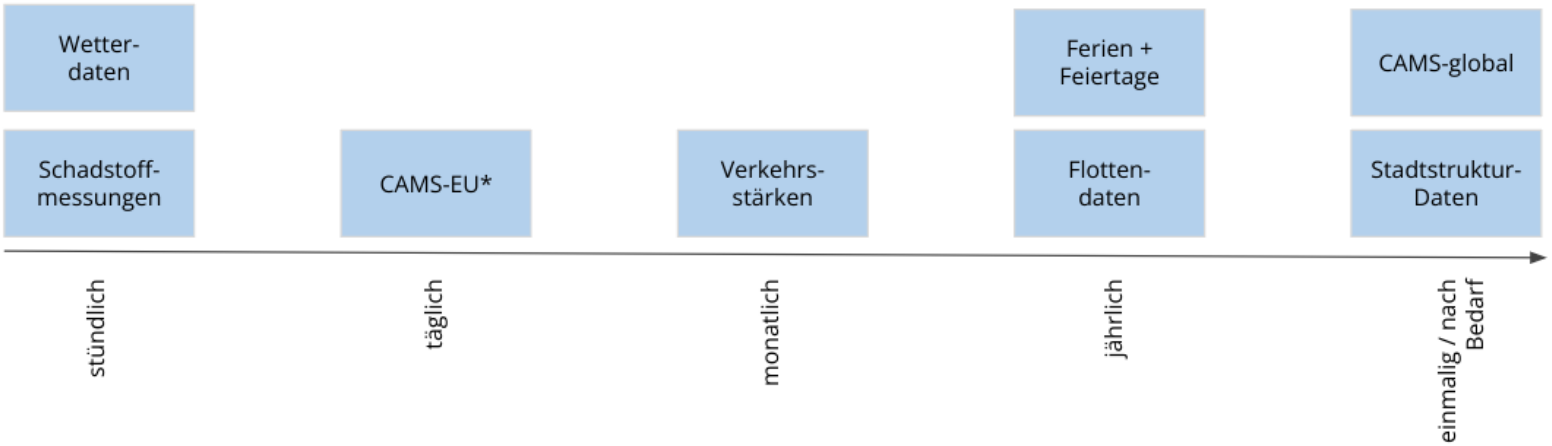
| | |
|---|---|
| | https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_wfs_baumbestand@senstadt<br>- Flächennutzung: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k06_02_1nutz_vegbestand2020@senstadt<br>- Emissionen: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=wmsk_03_12_2emissionen@senstadt<br>- 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<br>- 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<br>- Details zu den Schadstoff-Messstationen: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=messpunkte@senstadt<br>- Straßennetz: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_vms_detailnetz_wms_spatial@senstadt<br>- Verkehrsmengenkarte: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=k_vmengen2019@senstadt |

## Wie oft wird welche Datenquelle importiert?

Regelmäßige Datenimporte werden über sogenannte “Jobs” mit der Software Kubernetes gesteuert.
Die folgende Abbildung gibt einen Überblick über die Frequenz des Imports für die einzelnen Datenquellen. Im Anschluss werden Details zu den einzelnen Datenquellen erläutert.

Details:

- Luftschadstoffmessungen: Obwohl die Daten stündlich geupdated und importiert werden, werden hier stets die vergangenen zwei Monate importiert. Dadurch wird sichergestellt, dass nachträglich korrigierte bzw. gelöschte Werte die älteren Werte in der Datenbank überschreiben.
- CAMS-EU: Die Prognose wird einmal täglich im Laufe des Vormittags zur Verfügung gestellt. Der genaue Zeitpunkt variiert. Um die neueste Prognose möglichst schnell einzubinden, wird der Datenimport stündlich durchgeführt; bereits vorhandene Daten werden dabei einfach überschrieben.




- CAMS-global: Da hier nur historische Daten verwendet werden, wurden diese einmalig importiert.
- Verkehrsstärken: Diese Daten werden manuell etwa einmal monatlich importiert, sobald sie zur Verfügung stehen (Ziel: 15. des Monats). Eine spätere Automatisierung ist geplant und mit wenig Aufwand möglich.
- Ferien und Feiertage: Diese Daten werden manuell importiert und stehen für ca. 2,5 Jahre in die Zukunft zur Verfügung. Daher wird ein Update im jährlichen Rhythmus ausgeführt.
- Stadtstrukturdaten: Da diese Daten keine zeitliche Auflösung besitzen, erfolgt ein einmaliger Import. Ein manuelles Update kann hier nach ein paar Jahren sinnvoll sein.

# Datenverarbeitung

Im Anschluss an den Datenimport erfolgt eine erste Datenverarbeitung. Dabei werden u.a. folgende Schritte ausgeführt:

- Mapping des Koordinatensystems von Latitude/Longitude (EPSG:4326) auf x/y (EPSG:25833), welches später in der Modellierung und Bereitstellung der Prognose verwendet wird.
- Aggregation der Stadtstrukturdaten auf ein 50x50m$^2$-Gitter über Berlin
- Zusammenführen der CAMS-Daten aus den beiden unterschiedlichen Quellen, dazu Reskalierung der CAMS-global-Daten, sodass sie in Mittelwert und Standardabweichung den CAMS-Europe-Daten entsprechen
- Befüllen von Lücken ("Gap Filling") bei zeitlich aufgelösten Daten - dies wird im Folgenden erläutert.

## Gap Filling

Im Gap Filling werden Lücken in den Daten mit Hilfe eines Zeitreihenmodells (ARIMA, s.u.) aufgefüllt. Dies gilt sowohl für Lücken in der Historie als auch für Lücken am Ende der Daten, die z.B. durch fehlende Verfügbarkeit entstehen.
Kurze Lücken (maximal zwei Messzeitpunkte, d.h. zwei Stunden) werden durch den vorangegangenen Wert aufgefüllt. Für längere Lücken wird ein Zeitreihenmodell (sog. ARIMA, Implementierung: R package msarima) verwendet, welches einen Tages- und ggf. einen Wochenzyklus berücksichtigt - je nach Datenquelle. Außerdem unterscheidet sich der Endzeitpunkt, bis zu dem aufgefüllt wird, zwischen den Datenquellen. Die folgende Tabelle gibt einen Überblick darüber, ob und wie das Gap Filling für die einzelnen Datenquellen umgesetzt wurde:

| **Datenquelle** | **Gap Filling umgesetzt?** | **Wochenzyklus im Modell?** | **Tageszyklus im Modell?** | **Endzeitpunkt (bis wann wird aufgefüllt?)** |
|---|---|---|---|---|
| Luftschadstoff-messungen | ja | ja | ja | aktueller Zeitpunkt (d.h. ggf. Extrapolation in die Zukunft bei Ausfällen) |
| Wetterdaten (Beobachtungen) | ja | nein | ja | maximaler Zeitpunkt in den Daten |

| | | | | |
|---|---|---|---|---|
| Wetterdaten (Prognosen) | nein, weil die Prognosen lückenlos sind | | | |
| CAMS-Daten (Europe) | nein, weil die Prognosen lückenlos sind | | | |
| CAMS-Daten (global) | nein; da NO2-Werte aber nur alle 3 Stunden geliefert werden, werden die Lücken linear interpoliert | | | |
| Verkehrsstärken | nein, weil es sehr viele Lücken gibt und daher im späteren Verlauf der Modellierung ein Ansatz gewählt wurde, der mit lückenhaften Daten arbeiten kann | | | |

Sollten zukünftig doch Lücken in den Wetterprognosen auftreten, z.B. für eine Koordinate keine Werte geliefert werden oder die Schnittstelle für wenige Stunden nicht verfügbar sein, wird automatisch die neueste verfügbare Prognose verwendet.

## Code und Code-Dokumentation

Sämtlicher Code ist von Anfang an auf der Plattform Github hinterlegt, welche auch für die Versionskontrolle des Codes und Code-Reviews verwendet wird. Alle Funktionen sind dokumentiert, um anderen Entwickler*innen die Wartung und Weiterentwicklung zu erleichtern. Außerdem sind automatisierte Tests auf korrekte Funktionsweise sowie automatisierte Tests des Code-Styles (z.B. Formatierung, Einrückungen etc.) eingerichtet, um eine hohe Qualität sicherzustellen.

Der Code ist stark modularisiert, indem beispielsweise für jede Datenquelle ein separates Github-Repository angelegt ist. Im Laufe des Gesamtprojekts werden die Repositories mit Readmes ausgestattet, um Entwickler*innen den Einstieg zu erleichtern. Außerdem ist geplant, alle Repositories open source zur Verfügung zu stellen.

Im Folgenden werden alle Repositories, die mit Datenimport und -verarbeitung zusammenhängen, aufgelistet und kurz beschrieben.

| **Repository-Name** | **Quelle** | **Ziel-Datenbankschema** | **Funktion** |
|---|---|---|---|
| fairq-data-cams | CAMS-APIs | fairq-raw | CAMS-Daten abrufen und in Datenbank hinterlegen |
| fairq-data-dwd | Brightsky-API | fairq-raw | DWD-Daten abrufen und in Datenbank hinterlegen |
| fairq-data-kba | KBA-Webseite | fairq-raw | KBA-Daten abrufen und in Datenbank hinterlegen |
| fairq-data-messstationen | Webseite des Berliner Luftgütemessnetzes | fairq-raw | Luftschadstoff-Messdaten abrufen und in Datenbank hinterlegen |
| fairq-data-stadtstruktur | FIS-Broker | fairq-raw | Stadtstrukturdaten abrufen und in Datenbank hinterlegen |
| fairq-data-traffic-detectors | DPS | fairq-raw | Verkehrsdaten abrufen, Filterung auf relevante Messquerschnitte |
| fairq-dbtools | - | - | Hilfsfunktionen zur Kommunikation mit der Datenbank (für fast alle Arbeitspakete relevant) |
| fairq-features-datetime | schulferien.org | fairq-features | Berechnung von Ferien, Feiertagen etc. |
| fairq-features-stadtstruktur | fairq-raw und fairq-features | fairq-features | Aufbereitung der Stadtstrukturdaten, v.a. Aggregation auf 50x50m$^2$-Grid im Ziel-Koordinatensystem EPSG:25833 |
| fairq-gap-filling | fairq-raw | fairq-features | Berechnung des Gap filling (s. vorheriger Abschnitt) |

# Abweichungen von der ursprünglichen Planung

In Absprache mit der Auftraggeberin weicht die Umsetzung in einzelnen Punkten von der ursprünglichen Planung ab. Diese werden im Folgenden erläutert und begründet.

- *Import-Häufigkeit der Verkehrsdaten.* Die Verkehrsdaten werden nicht stündlich über eine API importiert, sondern monatlich. Grund dafür ist, dass eine stündliche Bereitstellung trotz höherem Aufwand keinen nennenswerten Mehrwert in der Modellierung bringen würde.
- *Modellierte stündliche Verkehrsstärken und -zustände inklusive der modellierten NO2-, NOx- und PM10-Daten an den Hauptverkehrsstraßen*. Diese Daten werden nicht importiert. Eine Bereitstellung der Daten durch SenUMVK hat sich verzögert, wird aber ab 2023 gewährleistet. Zum aktuellen Zeitpunkt liegen die Daten über eine API nicht vor; eine manuelle Bereitstellung durch SenUMVK würde einen unverhältnismäßigen Aufwand bedeuten und würde das Projekt evtl. verzögern.
- *Stadtstrukturdaten.* Hier wurden zusätzliche Daten importiert: Details zu den Schadstoff-Messcontainern, Emissionen, Straßennetz und Verkehrsmenge.
- *CAMS-Daten.* Hier wurden zwei Quellen statt einer angebunden, weil die CAMS-Prognose für Europa nur für die drei vergangenen Jahre zur Verfügung steht. Für ältere Zeitpunkte wird die zeitlich und räumlich etwas niedriger aufgelöste Quelle CAMS-Global verwendet.
- *Wetterdaten.* Da über die verwendete Brightsky-API keine historischen -Vorhersagen zur Verfügung stehen, werden hier zunächst historisch die beobachteten Daten abgerufen, während zusätzlich stündlich die Prognosen abgerufen werden, um schrittweise eine Historie aufzubauen, die später im Modelltraining verwendet werden kann. Als alternative Quelle für historische Vorhersagen wurde Pamore (https://www.dwd.de/DE/leistungen/pamore/pamore.html) geprüft. Hier sind jedoch keine historischen Vorhersagen nach der MOSMIX-Methode verfügbar, sodass die Verwendung von Pamore keinen Mehrwert liefert.
- *Flottenzusammensetzung.* Statt der Flottenzusammensetzung in Berlin werden in Abstimmung mit der SenUMVK die oben aufgeführten KBA-Daten verwendet, die vergleichbare Informationen enthalten, aus denen das Modell langfristige Veränderungen der Zusammensetzung der KFZ erlernen kann.

Es ist nicht ausgeschlossen, dass Datenquellen, die im Rahmen von AP3 nicht importiert wurden, zu einem späteren Zeitpunkt in Absprache mit der Auftraggeberin noch angebunden werden.

# Risikomanagement

## Risikokategorie 1: Fehlende Werte

Diese kleineren, zeitlich begrenzten Lücken können entweder in historischen Daten vorkommen oder dadurch entstehen, dass der Dienst eines Datenlieferanten für kurze Zeit nicht zur Verfügung steht.
Diese Lücken werden automatisch durch das oben beschriebene Gap Filling aufgefüllt. Dadurch können Modelle, die auf lückenlose Zeitreihen angewiesen sind, stets unterbrechungsfrei weiter verwendet werden. Dies betrifft sowohl die Modellschätzung als auch die Prognoseerstellung.

## Risikokategorie 2: Größere Lücken

Ursache hierfür sind größere Lücken in den historischen Daten (z.B. längerer Ausfall eines Messcontainers) oder der Ausfall eines Dienstes eines Datenlieferanten über mehrere Tage.
Auch diese Lücken werden durch das Gap Filling automatisch befüllt, um lückenlose Zeitreihen zu gewährleisten. Beim ersten Datenimport wurde überprüft, dass Lücken in der Vergangenheit immer noch klein genug sind, um auf diese Weise ergänzt zu werden. Zusätzlich ist ein automatisiertes Monitoring eingerichtet, welches an das gesamte INWT-Luftschadstoffprognose-Team Alarme sendet, falls es längere Datenausfälle gibt. Diese Alarme decken sowohl komplette Ausfälle ab als auch - je nach Quelle - Teil-Ausfälle, z.B. fehlende Schadstoffdaten für eine bestimmte Messstation,fehlende Wetterprognosen für ein Koordinatenpaar oder eine verspätete Bereitstellung der CAMS-EU-Prognosen.
INWT prüft dann im Einzelfall die Implikationen für die Modellschätzung. Ab einem Ausfall von drei aufeinanderfolgenden Tagen hält INWT Rücksprache mit der SenUMVK und ggf. mit den Datenlieferanten.

## Risikokategorie 3: Ausfall zentraler Daten über mehrere Wochen oder Einstellung der Datenlieferung durch einen Datenlieferanten

Längere Ausfälle werden durch das Monitoring ebenfalls erkannt. In solchen Fällen wird Rücksprache mit der SenUMVK gehalten.
Welche Datenquellen in der Modellschätzung verwendet werden, wird im Rahmen von AP2 zentral in den jeweiligen Repositories hinterlegt (z.B. in Form einer R-Modellformel oder als Parameterliste in einer JSON-Datei). Sollte eine Datenquelle eingestellt werden, kann diese aus dem Modell entfernt und das Modell mit den verbliebenen Eingangsvariablen neu kalibriert werden. Eine bloße Entfernung von Eingangsvariablen aus dem Modell mit anschließender Neukalibrierung ist innerhalb von Stunden bis wenigen Tagen möglich. Da hierbei im Grunde ein ganz neues Modell aufgesetzt wird, wird zusätzlich ein detailliertes Modell-Review empfohlen, inkl. Anpassung der sog. Modell-Hyperparameter und genauer Analyse der resultierenden Prognosen. Dabei prüfen INWT und SenUMVK u.a. gemeinsam die Plausibilität der gefundenen Zusammenhänge. Dadurch lassen sich die Auswirkungen der weggefallenen Datenquelle analysieren, während gleichzeitig mit den verbliebenen Variablen eine möglichst gute Prognose erstellt wird.
Als Beispiel sei angenommen, dass CAMS nur noch Prognosen für die Hintergrundbelastung durch PM10, aber nicht mehr für PM2.5 liefert. Die CAMS-Prognose für PM2.5 ist aber wahrscheinlich eine relevante Eingangsvariable bei der Prognose der PM2.5-Werte an den Messcontainern. Das PM2.5-Modell könnte derart angepasst werden, dass es als Ersatz die PM10-CAMS-Prognose verwendet. Das neue Modell muss dann dahingehend geprüft werden, ob sich die Modellgüte durch die Veränderung relevant verschlechtert und ob die Zusammenhänge der Zielvariablen (PM2.5) mit allen erklärenden Variablen weiterhin plausibel sind. Falls nicht, können weitere Anpassungen am Modell vorgenommen werden, um die gewünschte Qualität der Prognosen wieder zu erreichen.

## Bereitstellung der Prognose

Die Prognosen werden über eine API zur Verfügung gestellt (AP7).

In der Modellierung und Prognose-Erstellung sind die folgenden drei Schritte voneinander entkoppelt:

1. Modellentwicklung inkl. Variablenselektion und Optimierung von Hyperparametern (gelegentlich, z.B. im Abstand einiger Wochen bis Monate)
2. Schätzung (Kalibrierung, Neu-Eichung) des Modells mit den bis zu diesem Zeitpunkt verfügbaren Daten (regelmäßig, z.B. im Abstand von wenigen Wochen)
3. Erstellung der Prognosen (häufig, mindestens alle 24 Stunden)

Die exakte Festlegung der Häufigkeiten erfolgt im Rahmen von AP7, wenn die Prognose automatisiert wird.

Das in Schritt 1 entwickelte und in Schritt 2 geschätzte Modell wird in einer Datenbank hinterlegt. Von dort kann es für Schritt 3 geladen werden, um Prognosen zu erstellen. Sollte es bei der geplanten Modell-Neuschätzung Probleme geben, kann solange das bisherige Modell verwendet werden, bis die Probleme gelöst sind. Dadurch können weiterhin ohne Unterbrechung Prognosen erstellt werden.

Außerdem wird mit Containerisierung, Code-Versionierung sowie einer Trennung von Entwicklungs- und Produktivumgebung gearbeitet. Dadurch ist u.a. sichergestellt, dass Änderungen am Modell zuerst in der Entwicklungsumgebung umfangreich getestet werden können und den Produktivbetrieb nicht stören.
Ein Caching der API wird im Rahmen von AP7 umgesetzt, um eine Überlastung des Systems (der von Hetzner gehostete Server, auf dem die API läuft) zu verhindern und kurzzeitige Systemausfälle automatisch zu überbrücken.