Räumliche Interpolation von Feinstaub-Sensordaten mit Hilfe von Kriging

Aaron Braatz

Hamburg University of Applied Science
Faculty of Computer Science and Engineering
Department of Computer Science
Berliner Tor 7, 20099 Hamburg
aaron.braatz@haw-hamburg.de

Zusammenfassung. In dieser Arbeit wird ein Ansatz vorgestellt, wie aus Feinstaubdaten eines dynamischen Sensornetzes ein reguläres räumliches Gitter an Daten geschätzt werden kann. Dadurch wird es ermöglicht Methoden für zeitliche Analysen und Prognosen zu nutzen, die einen klar definierten Eingangsvektor brauchen.

Schlüsselwörter: Geostatistik · Regression-Kriging · Kriging · Variogramm · raumzeitliche Datenanalyse · Feinstaubanalyse

1 Einleitung

Räumliche Interpolation ist eine Teildisziplin der raumzeitlichen Datenanalyse (*engl. Spatio-temporal Data Analysis*; STDA) und wird genutzt, um aus einer Menge von Messungen Karten und Gitter zu schätzen. Auf diese Weise soll, aus den Messungen eines dynamischen Sensornetzes für Feinstaub, ein gleichmäßiges, räumliches Gitter bestimmen. *Kriging* ist eine Methode der räumlichen Interpolation, die ihren Ursprung in der Bergbauindustrie der 1950er-Jahre hat, um den Verlauf von Gesteinsvorkommen zu schätzen. Seitdem wurde das Modell auf viele weitere Bereiche ausgeweitet und wird für alle Daten genutzt, die eine räumliche Autokorrelationsstruktur haben. In dieser Arbeit soll *Kriging* ein regelmäßiges Gitter an Feinstaubdaten generieren.
Die Daten stammen aus dem Citizen Science Projekt *Sensor.Community*. Sie bieten jeder Person die Möglichkeit eine eigene Sensorstation aufzustellen und Daten zu Feinstaub, Temperatur, Luftfeuchtigkeit und weiteren Umwelteinflüssen zu sammeln. Diese Flexibilität bietet allerdings eine Herausforderung für Analysemodelle, die auf eine klare Definition der Eingangsdaten angewiesen sind.
Diese Arbeit entsteht als Teil eines Projekts für STDA von Feinstaubdaten (siehe Abb. 1). In diesem Teil des Projekts erfolgt die Vorarbeit, damit weitere Modelle für die zeitliche Analyse von Feinstaub genutzt werden können, und um die weitere Entwicklung zu prognostizieren. Ziel ist es, Erkenntnis über folgende Fragestellungen zu erhalten: Inwiefern ist der weltweite Rückgang des Transports während Corona erkennbar? Inwiefern verbessern Werksferien von Industriebetrieben die Feinstaubbelastung? Wie wirken sich eine Rushhour oder Ferien auf

Feinstaub aus? Wie bewegt sich eine Ballung erhöhter Feinstaubbelastung und wie gut lässt sich die Bewegung vorherzusagen?
In den Kapiteln 2 und 3 werden STDA und räumliche Interpolation genauer erklärt, wobei diesbezüglich relevante Modelle in Kapitel 4 vorgestellt werden. Kapitel 5 beschreibt den experimentellen Aufbau. Dabei wird genauer auf den Datensatz und dessen Besonderheiten eingegangen. Außerdem werden der Ablauf und die genutzten Technologien erklärt. Kapitel 6 stellt Metriken vor mit denen das *Kriging*-Modell evaluiert werden kann und in Kapitel 7 werden die bisherigen Ergebnisse vorgestellt und ein Ausblick für den weiteren Verlauf des Projekts gegeben.

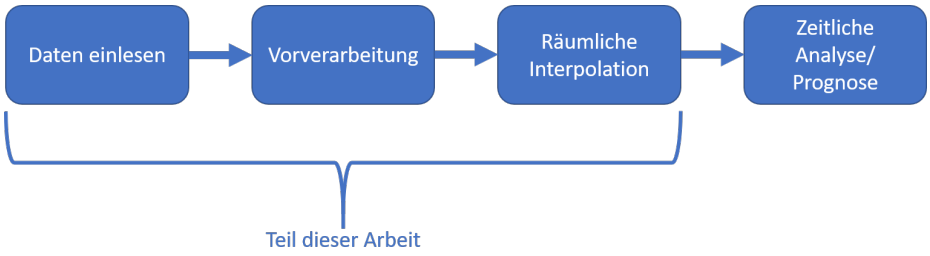

Abb. 1. Schematischer Ablauf des Gesamtprojekts

2 Raumzeitliche Datenanalyse

Raumzeitliche Datenanalyse (*engl. Spatio-temporal Data Analysis*; STDA) bietet verschiedene Methoden zur Analyse von Daten, die geographische Koordinaten und gegebenenfalls zeitliche Informationen enthalten:

- Bildverarbeitung für Fernerkundungsdaten
- Punktmusteranalyse für Punkt- und Linienobjekten
- Geostatistik für kontinuierliche räumliche Merkmale (Felder)
- Geomorphometrie für Topographien

Diese Methodensind auf zwei Wissenschafsbereiche zurückzuführen: Geoinformationswissenschaft (GIS) und raumzeitliche Statistik. Nach [Ripley, 2005, S. 1] dient die räumliche Statistik zur Datenzusammenfassung von räumlichen Daten und Erklärung von räumlichen Mustern durch theoretische Modelle. Die Geostatistik hingegen entstammt der Bergbauindustrie und hat sich daher vorrangig mit der Geologie beschäftigt. Im Gegensatz zur ursprünglichen Anwendung, bekommt heutzutage die zeitliche Komponente immer mehr Relevanz. So finden sich Techniken der Geostatistik mittlerweile in diversen Bereichen wieder, zum Beispiel der Bodenkartierung, Meteorologie, Ökologie, Ozeanographie, Geochemie, Epidemiologie, Humangeographie, Geomorphometrie und weiteren. [Hengl, 2007, S. 1]

3 Räumliche Interpolation

Bei räumlicher Interpolation oder auch räumlicher Vorhersage werden Werte einer Zielvariable für unbeobachtete Orte innerhalb eines definierten Bereichs geschätzt. Diese Vorhersagen werden in Karten oder Bildern dargestellt. In der Geostatistik liegt bei der räumlichen Interpolation der zu schätzende Ort umgeben von den Stichprobenorten und ist innerhalb des räumlichen Autokorrelationsbereich. Innerhalb Autokorrelationsbereich ändert sich der Wert einer Variablen an einem Ort ähnlich zu den Nachbarn. Die räumliche Vorhersage beinhaltet zusätzlich die räumliche Extrapolation. Dabei werden auch Werte außerhalb des bekannten Stichprobenbereichs geschätzt. Grundsätzlich gilt für die räumliche Extrapolation, dass der zu schätzende Ort außerhalb des praktischen Bereichs liegt, also der Vorhersagefehler größer ist, als die globale Varianz. Dementsprechend können bei der Extrapolation keine signifikanten Vorhersagen gemacht werden.[Hengl, 2007, S. 3][Anselin, 2002, S. 256]

4 Modelle zur räumlichen Interpolation

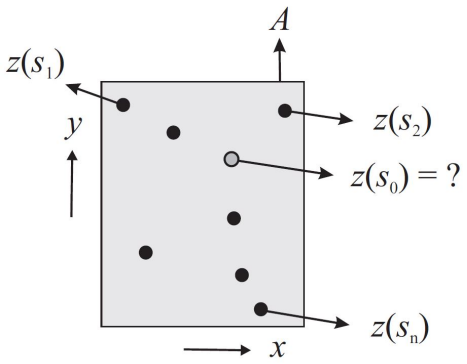

Abb. 2. Räumliche Interpolation in einem Bereich mit bekannten Beobachtungen[Hengl, 2007, Abb. 1.7a]

In einer idealen Umgebung ließe sich die Variabilität einer Umgebungsvariable durch bekannte physikalische Gesetze und eine endliche Menge an Parametern definieren. In einer solchen Umgebung könnten auch die Werte einer Zielvariablen exakt bestimmt werden. In der Wirklichkeit sind diese Umstände meist nicht gegeben und die Prozesse lassen sich nicht präzise modellieren [Heuvelink and Webster, 2001, S. 294]. Daher wird versucht, anhand von realen Messungen ein passendes Modell zu schätzen. In mathematischer Notation wird eine Menge von Beobachtungen einer Zielvariablen Z als $z(s_1), z(s_2), \ldots, z(s_n)$ bezeichnet. Dabei ist $s_i = (x_i, y_i)$ ein Ort mit x_i und y_i als geografische Koordinaten. n ist

die Anzahl der Beobachtungen und A bezeichnet den betrachteten Bereich (siehe Abb. 2). Für das Modell wird angenommen, dass die Stichproben repräsentativ, nicht präferentiell und konsistent sind, damit ein Vorhersagemodell für einen unbekannten Ort s_0 die Werte der Zielvariablen bestimmen kann. Daraus ergibt sich:

$$\hat{z}(s_0) = E\{Z|z(s_i), q_k(s_0), \gamma(h), s \in A\} \tag{1}$$

wobei $z(s_i)$ der Eingabepunktdatensatz, $\gamma(h)$ das Kovarianzmodell, welches die räumliche Autokorrelationsstruktur definiert und $q_k(s_0)$ die Liste der deterministischen Prädiktoren ist. Die Prädiktoren werden auch als Kovariaten oder erklärende Variablen bezeichnet und sind zusätzliche Informationen, die für jeden Ort innerhalb A bekannt sein müssen. [Hengl, 2007, S. 9]
[Jin Li, Andrew D. Heap, 2008] gibt eine Übersicht über mehr als 40 räumliche Interpolationsmodelle, darunter zwölf nicht-geostatistische, 22 geostatistische und acht kombinierte Modelle. Diese Modelle wurden von unterschiedlichen Domänen für ihre jeweiligen Bedürfnisse entwickelt. Zu einem großen Teil dieser Modelle bilden Kriging oder Regression Kriging die Basis. Daher betrachten wir in dieser Arbeit diese grundlegenden Modelle.
Die Gleichung 1 beschreibt ein allgemeines Modell, allerdings beschränken sich einige Modelle auf eine Submenge der Eingaben. Den Fokus auf die Prädiktoren hat*Multi Linear Regression* (4.2) und verzichtet auf das Kovarianzmodell, wohingegen *Ordinary Kriging* (4.3) nur das Kovarianzmodell und keine Prädiktoren nutzt. *Regression Kriging* 4.4 kombiniert beide Ansätze. Wiederum nutzt *Inverse Distance Interpolation* 4.1 nur die Entfernung zwischen den Orten. Außerdem gibt es *Konditionalen Autoregressiven Modelle*, die in dieser Arbeit nicht betrachtet werden, da es schwierig ist mit dieser Art von Modellen gute Schätzungen zu erhalten [Florax and Nijkamp, 2003, S. 13].

4.1 Inverse Distance Interpolation

Inverse Distance Interpolation (IDI) [Shepard, 1968] ist eines der ältesten räumlichen Interpolationsmodelle [Hengl, 2007, S. 12]. IDI nutzt für die Interpolation an einem neuen Ort einen gewichteten Mittelwert der umliegenden Messungen:

$$\hat{z}(s_0) = \sum_{i=1}^{n} \lambda_i(s_0) \cdot z(s_i) \tag{2}$$

wobei λ_i das Gewicht einer benachbarten Messung i ist. Die Summe der Gewichte muss eins betragen um keinen Bias einzuführen. Die Matrixform ist:

$$\hat{z}(s_0) = \lambda_i(s_0)^{\mathrm{T}} \cdot \mathrm{z} \tag{3}$$

Die Gewichte werden durch die Inverse der Distanz bestimmt:

$$\lambda_i(s_0) = \frac{\frac{1}{d^{\beta}(s_0,s_i)}}{\sum_{j=0}^{n} \frac{1}{d^{\beta}(s_0,s_j)}}; \qquad \beta > 1 \tag{4}$$

wobei $d(s_0, s_i)$ die Distanz zwischen dem zu schätzenden und einem bekannten Ort ist. β ist der Koeffizient um die Gewichtung anzupassen und muss manuell gewählt werden. Ziel ist es mit dem Koeffizienten möglichst gut die Stärke der Autokorrelation abzubilden. IDI setzt mit dem Modell Tobler's erstes Gesetz der Geographie um:

"Everything is related to everything else but nearby things are more related than distant things" [Tobler, 1970, S. 236]

4.2 Multiple Linear Regression

Multiple Linear Regression (MLR) ist ein häufig genutztes Regressionsmodell für räumliche Interpolation aus dem Bereich *umweltbedingte Korrelation.*[Draper and Smith, 1998][Kutner, 2005] und ist wiefolgt definiert:

$$\begin{aligned} \hat{z}_{MLR}(s_0) &= \hat{b}_0 + \hat{b}_1 \cdot q_1(s_0) + ... + \hat{b}_p \cdot q_p(s_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0); \qquad q_0(s_0) \equiv 1 \end{aligned} \tag{5}$$

bzw. in Matrix-Schreibweise:

$$\hat{z}_{MLR}(s_0) = \hat{\beta}^T \cdot \mathrm{q} \tag{6}$$

$q_k(s_o)$ sind dabei die Werte der Prädiktoren an der Zielposition, p ist die Anzahl der Prädiktoren und $\hat{\beta}_k$ sind die Regressionskoeffizienten, welche mit *Ordinary Least Squares* (OLS) [Kutner, 2005, S. 257] ermittelt werden:

$$\hat{\beta} = \left(\mathrm{q}^T \cdot \mathrm{q}\right)^{-1} \cdot \mathrm{q}^T \cdot \mathrm{z} \tag{7}$$

q beschreibt hierbei die Matrix der Prädiktoren $(n \times (p+1))$ und z ist der Vektor mit den Werten der Messungen. Der Vorhersagefehler einer MLR ist:

$$\hat{\sigma}^2_{MLR}(s_0) = MSE \cdot \left[1 + \mathrm{q}_0^T \cdot \left(\mathrm{q}^T \cdot \mathrm{q}\right)^{-1} \cdot \mathrm{q}_0\right] \tag{8}$$

dabei ist MSE die mittlere quadratische Abweichung zur Regressionsgeraden:

$$MSE = \frac{\sum_{i=1}^{n} \left[z(s_i) - \hat{z}(s_i)\right]^2}{n-2} \tag{9}$$

und q_0 ist der Vektor mit den Prädiktoren für die zu schätzenden Orte. Für eine univariate lineare Regression kann die Varianz des Schätzfehlers ermittelt werden mit:

$$\hat{\sigma}^2(s_0) = MSE \cdot \left[1 + \frac{1}{n} + \frac{\left[q(s_0) - \bar{q}\right]^2}{\sum_{i=1}^{n} \left[q(s_i) - \bar{q}\right]^2}\right] = MSE \cdot \left[1 + v(s_0)\right] \tag{10}$$

wobei v die Krümmung des Konfidenzbands um die Regressionsgerade ist. Ein Kritikpunkt an Regressionsmodellen wie MLR ist, dass die räumliche Verteilung der Messpunkte nicht berücksichtigt wird. Um dem entgegenzuwirken, kann *Geographically Weigthed Regression* (GWR)[Brunsdon et al., 1996, S. 5] anstelle von OLS bei der Bestimmung der Regressionskoeffizienten genutzt werden:

$$\hat{\beta}_{WLS} = \left(\mathrm{q}^T \cdot \mathrm{W} \cdot \mathrm{q}\right)^{-1} \cdot \mathrm{q}^T \cdot \mathrm{W} \cdot \mathrm{z} \tag{11}$$

wobei W die Matrix der Gewichte ist. Sie kann mit unterschiedlichen Distanzgewichtenden Kernels berechnet werden, wie zum Beispiel dem Bisquare [Wheeler and Páez, 2010, S. 464]:

$$w(s_i, s_j) = \left[1 - \left(\frac{d(s_i, s_j)}{\gamma}\right)^2\right]^2 \tag{12}$$

wobei $w(s_i, s_j)$ das Gewicht und $d(s_i, s_j)$ die Distanz zwischen den beiden Orten s_i und s_j angibt. γ gibt die Bandbreite an und ermöglicht es den Grad der Lokalität einzustellen. Die Bandbreite bietet die Möglichkeit zu ergründen, wie sich die Lokalität an den Messorten verhält, allerdings muss sie manuell angepasst werden. [Bidanset and Lombard, 2014]. Im gegensatz zur Brandbeite, die alle Orte berücksichtigt, gibt es auch die Option eines *n-nearest-neighbors*-Ansatzes. Hierbei werden nicht alle umliegenden Messorte betrachtet und gewichtet, sondern nur die n nächsten zur Zielposition.

4.3 Ordinary Kriging

Ordinary Kriging (OK) gehört zu den statistischen räumlichen Vorhersagemodellen und ist die Standardvariante von *Kriging*. *Kriging* wurde lange Zeit synonym zu geostatistischer Interpolation genutzt. Erst einige Jahre später wurden das Model von [Matheron, 1962] mathematischen beschrieben.
Das Interpolationsmodell wird definiert durch:

$$Z(s) = \mu + \varepsilon'(s) \tag{13}$$

wobei μ eine konstante stationäre Funktion (globaler Mittelwert) ist und $\varepsilon'(s)$ die Varianz stochastisch über die räumliche Korrelation beschreibt. Die Interpolation wird ähnlich wie in Gleichung 2 durchgeführt:

$$\hat{z}_{OK}(s_0) = \sum_{i=1}^{n} w_i(s_0) \cdot z(s_i) = \lambda_0^T \cdot \mathrm{z} \tag{14}$$

wobei λ_0 der Vektor mit den Kriging-Gewichten w_i und z der Vektor mit den Messungen an den n bekannten Orten ist. Im Gegensatz zu IDI müssen bei der Bestimmung der Gewichte keine Parameter manuell gesetzt werden, sondern durch die *Semivarianz*-Unterschiede zwischen benachbarten Messungen bestimmt:

$$\gamma(h) = \frac{1}{2} E\left[\left(z(s_i) - z(s_i + h)\right)^2\right] \tag{15}$$

dabei ist $z(s_i)$ der Wert an einer gemessenen Position und $z(s_i + h)$ der Wert an einer benachbarten Position mit einem Abstand von $s_i + h$. Daraus resultiert ein empirisches Variogramm, für welches anschließend auf ein theoretisches Variogramm angepasst wird. Dieser Prozess wurde bereits in [Braatz, 2019, S. 13ff.] beschrieben. Das theoretische Variogramm wird beschrieben durch:

- *Sill*, der maximalen Semivarianz im Variogramm
- *Range*, Entfernung als *Lag*, ab dem *Sill* erreicht wird
- *Nuggeteffekt*, die Abweichung, der Semivarianz am Ursprung

Dabei beschreibt *Lag* ein Entfernungsintervall.

Die Parameter eines Variogramms können iterativ mittels *Least Squares Estimator* bestimmt werden auf Basis der Anzahl der Punktpaare oder der Abstände. [Hengl, 2007, S. 16]
Mit dem theoretischen Variogramm können die Gewichte wiefolgt erhalten werden:

$$\lambda_0 = \mathrm{C}^{-1} \cdot c_o; \qquad C(|h| = 0) = C_0 + C_1 \tag{16}$$

dabei ist C die Kovarianz-Matrix der $n \times n$ Messungen und c_o der Kovarianz-Vektor an einer neuen Position. Allerdings hat C die Größe $(n+1) \times (n+1)$ um sicherzugehen, sodass die Summe der Gewichte gleich eins ist.

$$\begin{bmatrix} C(s_1,s_1) & \dots & C(s_1,s_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(s_n,s_1) & \dots & C(s_n,s_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C(s_0,s_1) \\ \vdots \\ C(s_0,s_n) \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(s_0) \\ \vdots \\ w_n(s_0) \\ \varphi \end{bmatrix} \tag{17}$$

dabei ist φ der sogenannte *Lagrange-Multiplikator*.
Die Schätzvarianz für OK (auch OK-Varianz) beschreibt die Unsicherheit des Modells für eine Schätzung an einem bestimmten Ort. Die OK-Varianz wird beschrieben durch[Webster and Oliver, 2007, S. 158]:

$$\hat{\sigma}^2_{OK} = (C_0 + C_1) - c_o^T \cdot \lambda_0 = C_0 + C_1 - \sum_{i=1}^{n} w_i(s_0) \cdot C(s_0, s_i) + \varphi \tag{18}$$

Sofern die OK-Varianz größer oder gleich der globalen Varianz ist, ist das Modell maximal unpräzise, ist die OK-Varianz hingegen null, dann ist das Modell absolut präzise. Allerdings beschreibt dies nur wie gut das Modell an die Werte angepasst ist und nicht wie gut das Modell unbekannte Orte interpoliert. Dafür wird ein Testdatensatz benötigt, wie in Kapitel 6 beschrieben.
Bei OK wird davon ausgegangen, dass die Trendfunktion μ und das Variogramm konstant den gesamten betrachteten Bereich ist und die Zielvariable annähernd normal verteilt ist. Diese Annahme in der Realität nicht immer erfüllt.[Hengl, 2007, S. 16]

4.4 Regression-Kriging

In [Matheron, 1969] wird ein Modell vorgestellt, mit dem der Wert einer Zielvariablen durch die Summe aus einem stochastischen und deterministischen Teil dargestellt wird:

$$Z(s) = m(s) + \varepsilon'(s) + \varepsilon'' \tag{19}$$

In den bisherigen Modellen haben wir mit OK die statistische Variation von räumlichen Daten und mit MLR die deterministische Variation durch umweltbedingte Korrelation modelliert. *Regression Kriging* (RK) verbindet diese beiden Ansätze:

$$\hat{z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=0}^{n} \lambda_i \cdot e(s_i) \tag{20}$$

wobei $\hat{m}(s_0)$ der modellierte deterministisch Teil ist. $\hat{e}(s_0)$ beschreibt das interpolierte Residuum, $\hat{\beta}_k$ steht für die Model Koeffizienten und λ_i sind die *Kriging*-Gewichte bezüglich der Autokorellationsstruktur der Residuen $e(s_i)$. $\hat{\beta}_k$ kann durch OLS oder durch *Generalized Least Squares* (GLS) [Cressie, 2001] bestimmt werden:

$$\hat{\beta}_{GLS} = \left(\mathrm{q}^T \cdot \mathrm{C}^{-1} \cdot \mathrm{q}\right)^{-1} \cdot \mathrm{q}^T \cdot \mathrm{C}^{-1} \cdot \mathrm{z} \tag{21}$$

dabei ist $\hat{\beta}_{GLS}$ der Vektor mit den geschätzten Regressionskoeffizienten, C ist die Kovarianzmatrix, q ist die Matrix der Prädiktoren und z sind die Messungen. Die Matrixform von RK ist:

$$\hat{z}_{RK}(s_0) = \mathrm{q}_0^T \cdot \hat{\beta}_{GLS} + \lambda_0^T \cdot \left(\mathrm{z} - \mathrm{q} \cdot \hat{\beta}_{GLS}\right) \tag{22}$$

wobei $\hat{z}(s_0)$ der interpolierte Wert am Ort s_0 ist. q_0 ist der Vektor mit $p+1$ Prädiktoren und λ_0 ist der Vektor mit n Kriging-Gewichten. Dieses Modell wird als *Best Linear Predictor of spatial data* bezeichnet [Christensen, 1991]. Die dazugehörige Schätzvarianz ist definiert durch:

$$\begin{aligned} \hat{\sigma}^2_{RK} = (C_0 + C_1) - \mathrm{c}_0^T \cdot \mathrm{C}^{-1} \cdot \mathrm{c}_0 \\ + \left(\mathrm{q}_0 - \mathrm{q}^T \cdot \mathrm{C}^{-1} \cdot \mathrm{c}_0\right)^T \cdot \left(\mathrm{q}^T \cdot \mathrm{C}^{-1} \cdot \mathrm{q}\right)^{-1} \cdot \left(\mathrm{q}_0 - \mathrm{q}^T \cdot \mathrm{C}^{-1} \cdot \mathrm{c}_0\right) \end{aligned} \tag{23}$$

wobei $C_0 + C_1$ die Sill-Variation und $c0$ der Vektor der Kovarianzen zu den Residuen an den zu schätzenden Orten ist.
Neben RK gibt es noch *Universal Kriging* (UK) und *Kriging with external Drift* (KED). Diese sind aber grundsätzlich sehr ähnlich zu RK. Auch gibt es noch weiter Verfeinerung von RK, auf die in dieser Arbeit nicht eingegangen wird, aber detailliert in [Hengl, 2007, S. 36ff.] aufgeführt sind.

5 experimenteller Aufbau

Der experimentelle Aufbau orientiert sich an dem *Knowledge Discovery in Databases* (KKD)-Prozess von [Fayyad et al., 1996]. Dieser hat das Ziel, Wissen aus Daten zu extrahieren. Dabei werden fünf Schritte iterative durchlaufen (siehe Abb. 3).

- Bei der *Datenselektion* werden relevante Datensätze identifiziert, oder falls nötig erstellt
- Die *Vorverarbeitung* dient zur Behandlung von fehlenden und falschen Daten
- Die *Datentransformation* dient dazu die Daten in eine Form zu überführen, sodass sie für das gewählte Modell verarbeitbar werden
- Das *Data Mining* beschreibt den eigentlichen Prozess der Wissensextraktion durch ein Modell
- Während der *Interpretation / Evaluation* werden die Ergebnisse diskutiert. Je nach Bedarf kann hiernach an einem vorherigen Schritt neu angesetzt und Verbesserungen vorgenommen werden

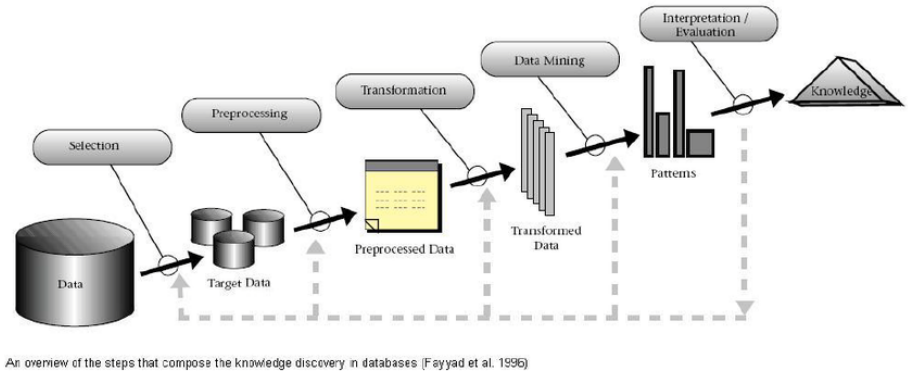

Abb. 3. Schritte des KDD-Prozesses[Fayyad et al., 1996, Abb. 1]

5.1 Datensatz

Der in dieser Arbeit betrachtete Datensatz wird bereitgestellt von dem Projekt Sensor.Community[1] (ehem. Luftdaten.info), welches ins Leben gerufen wurde durch das OK Lab Stuttgart. Laut ihrer Website ist das OK Lab Stuttgart

[1] https://sensor.community/de/

,,[...] Teil des Programms Code for Germany[2] der Open Knowledge Foundation Germany[3]. Ziel des Programms ist es, Entwicklungen im Bereich Transparenz, Open Data und Citizen Science zu fördern." luftdaten.info – Feinstaub selber messen [05.01.2021]. In diesem Zusammenhang wird jeder Person die Möglichkeit gegeben, selbst Sensoren aufzustellen und die durch sie gewonnenen Daten zu teilen. Hierbei können unterschiedliche Umwelteinflüsse gemessen werden, wie Feinstaub, Temperatur, Luftfeuchtigkeit, Luftdruck und Lärm. Über Sensor.Community werden die Daten aller Sensoren zusammengefasst und in Karten visualisiert.
Diese Arbeit beschränkt sich auf die Feinstaubdaten, welche als PM10 (Particulate matter ≤ 10 μm) und PM2,5 (Particulate matter $\leq 2{,}5$ μm) zur Verfügung gestellt werden. Die ersten Feinstaubmessungen stehen seit Oktober 2015 zur Verfügung, damals noch mit dem PPD42NS-Sensor. Seit Juli 2016 wird vorwiegend der SDS011-Sensor genutzt.
Der SDS011 wurde bereits in [Braatz, 2019] genauer evaluiert. Dabei wurde darauf hingewiesen, dass die Messungen des Sensors durch starke Veränderungen in Temperatur und Luftfeuchtigkeit beeinflusst werden können. Das wird in dieser Arbeit vernachlässigt und eventuell in einer zukünftigen Arbeit betrachtet.
Die Daten sind öffentlich zugänglich und werden täglich für alle Sensoren gebündelt im CSV-Format[4] zur Verfügung gestellt. Außerdem gibt es monatliche Zusammenfassungen je Sensortyps im CSV- und Parquet-Format[5]. Allerdings sind die Parquet-Dateien nur bis Ende 2020 verfügbar. Aufgrund des hohen Speicherbedarfs wurde entschieden, die Daten nur noch als CSV zur Verfügung zu stellen. Dadurch erhöht sich der Aufwand in der Daten-Vorverarbeitung, da die Daten nicht mehr mit den Datentyp-Informationen gespeichert werden.
Der Datensatz enthält 12 Spalten:

1. sensor_id - numerische Identifikation für den Sensor
2. sensor_type - Typ des Sensors (hier nur 'SDS011')
3. location - numerische Identifikation des Sensorknotens
4. lat - Breitengrad des Sensorknotens mit drei Nachkommastellen
5. lon - Längengrad des Sensorknotens mit drei Nachkommastellen
6. timestamp - Zeitpunkt der Messung; größtenteils in ISO8601, aber auch in Unixzeit in Millisekunden
7. P1 - PM10 in ppm (Parts per million), teilweise aber auch als Text (z.B. 'PM10', '['PM10']', '%vla2%')
8. durP1 - Überrest des PDD42NS; wird nicht betrachtet
9. ratioP1 - Überrest des PDD42NS; wird nicht betrachtet
10. P2 - PM2,5 in ppm (Parts per million), teilweise aber auch als Text (z.B. 'PM25', '['PM25']', '%vla2%')
11. durP2 - Überrest des PDD42NS; wird nicht betrachtet
12. ratioP2 - Überrest des PDD42NS; wird nicht betrachtet

[2] https://codefor.de/
[3] https://okfn.de/
[4] Comma-seperated values
[5] https://parquet.apache.org/documentation/latest/

Die Daten werden von den Sensorstationen über eine API kommuniziert. Bei der Anbindung, der Station, können die Betreibenden der Sensorstation frei wählen, mit welcher Frequenz die Messungen übertragen werden. Dadurch gibt es sowohl Frequenzen von einer Messung pro Sekunde, aber auch von weniger als einer Messung innerhalb von fünf Tagen.
Eine weitere Schwierigkeit ist, dass nicht genauer bekannt ist, wo die Sensoren aufgestellt sind. Für die Analyse besonders relevant ist der Unterschied zwischen Indoor und Outdoor, da die Indoor-Sensoren nicht zur Autokorrelationsstruktur beitragen.
Die Messungen können auch extreme Werte annehmen. Selbst bei der Betrachtung des monatlichen Mittelwerts können diese extremen Werte nachgewiesen werden. So ist der maximale Mittelwert in dem beobachteten Zeitraum bei 11.404.888 ppm, was auf ein falschen Wert schließen lässt. Die Verteilung des monatlichen Mittelwerts wird in dem kumulativen Verteilungsdiagramm (Abb. 5) dargestellt. Dort ist zu erkennen, dass bereits 90 % der Mittelwerte bis 100 ppm abgebildet werden.
Des Weiteren ist die Verteilung der Sensoren nicht homogen. Tendenziell sind mehr Sensoren im Westen Deutschlands und grundsätzlich gibt es Ballungen bei größeren Städten (siehe Abb. 6).

5.2 Pipeline

In dem folgenden Kapitel wird der Ablauf der Datenverarbeitung und Modellerstellung detailliert beschrieben.

Datenselektion Der erste Schritt der Pipeline ist das Einlesen der Daten. Das kann direkt aus dem Datenarchiv der Sensor.Community erfolgen. Dabei ist zu beachten, dass die Daten im CSV-Format in einem anderen Verzeichnis abgelegt sind als die im Parquet-Format. Außerdem sind die Dateinamen der Parquet-Dateien nicht mit einem menschenlesbaren Namen versehen und daher nur über den Ort im Verzeichnis zu identifizieren.

Vorverarbeitung In der Vorverarbeitung werden grundsätzlich Fehler und fehlende Daten behandelt. In diesem Anwendungsfall müssen die fehlenden Messungen nicht durch synthetische Daten ergänzt werden, da das Ziel ist, aus den VORHANDENEN Daten ein gleichmäßiges Gitter an Daten zu schätzen. Die fehlenden Daten werden in dem Prozess durch die umliegenden Messungen geschätzt. Dabei wird bewusst akzeptiert, dass für das Modell weniger Stützpunkte zur Verfügung stehen.
Sofern Daten in den Spalten ,,timestamp", ,,lat" oder ,,lon" fehlen, werden die dazugehörigen Datenpunkte entfernt, da es keine Möglichkeit gibt, den raumzeitlichen Zusammenhang aufzuarbeiten.
Für die Vergleichbarkeit der Daten müssen alle Stationen auf eine gemeinsame Datenfrequenz gebracht werden. In diesem Fall wurde eine Frequenz von

zehn Minuten gewählt[6]. Hierfür wird ein *zeitlich gewichteter Mittelwert* genutzt. Das heißt, wenn auf die erste Messung zwei Minuten später eine zweite Messung folgt, dann geht die erste Messung mit einem zwei Minuten-Gewicht in den 10 Minuten-Mittelwert ein. Sind für einen Takt keine Daten verfügbar, wird der Datenpunkt entfernt. Diese Entscheidung wurde getroffen, da aufgrund der Flüchtigkeit von Feinstaub nicht gewährleistet werden kann, dass die letzte Messung noch valide ist.
Ausreißer werden über *spatial outlier detection* identifiziert und entfernt, aus den bereits genannten Begründungen. [Chen et al., 2008] beschreibt zwei Algorithmen: Einmal auf Basis des *Medians* und einmal mit Hilfe von *k nearest neighbor*. Welcher der beiden Algorithmen genutzt wird, wird in einer zukünftigen Arbeit erörtert.

Transforamtion Bei der Transformation wird der Datensatz in die einzelnen Zeitstempel aufgeteilt, damit im nächsten Schritt durch Kriging pro Zeitschritt das regelmäßige Gitter bestimmt werden kann. Außerdem werden PM10 und PM2,5 separat betrachtet, daher wird der Datensatz auch diesbezüglich aufgeteilt. Ein Datenpunkt besteht dann aus einem Messwert zusammen mit den räumlichen Koordinaten in Form von Längen- und Breitengrad.

Data Mining Als Modell soll RK genutzt. Dies ermöglicht in Zukunft auch noch weitere Prädiktoren einzuführen, wie zum Beispiel Winddaten. [Hengl, 2007] stellt einen Entscheidungsbaum (Abb. 7) zur Verfügung, welcher zuordnet, welches der vorgestellten Modelle gewählt werden soll. OK wurde bereits in [Braatz, 2019] für den Datensatz angewendet.
Bei einer ersten Analyse der Daten konnte festgestellt werden, dass deutschlandweit kein einheitliches Variogramm gefunden werden konnte, sondern nur für einzelne Bereiche, wie zum Beispiel für den Raum Wuppertal Abb. 8. Daher müssen noch Untersuchungen vorgenommen werden, ob RK als *moving window Kriging* genutzt werden kann. Dadurch könnte die jeweilige lokale Autokorrelationsstruktur betrachtet werden. Laut [Hengl, 2007, S. 41] gibt es dafür aber noch kaum Implementierungen.

5.3 Toolchain

In diesem Kapitel wird beschrieben, welche Technologien für die Umsetzung der Pipeline genutzt werden. Alle hier genutzten Technologien sind mindestens Open-Source und frei zugänglich.

Python Für die Durchführung des Projekts wird die Programmiersprache *Python*[van Rossum and Drake, 2009] genutzt. Python ist weitverbreitet, besonders durch ihre reiche Auswahl an wissenschaftlichen Bibliotheken. Diese ermöglichen die Verarbeitung, Analyse und Visualisierung von Daten.

[6] Diese Frequenz wurde gewählt, da in Zukunft eventuell Winddaten ergänzt werden sollen. Diese sind in einem 10 Minuten Takt verfügbar

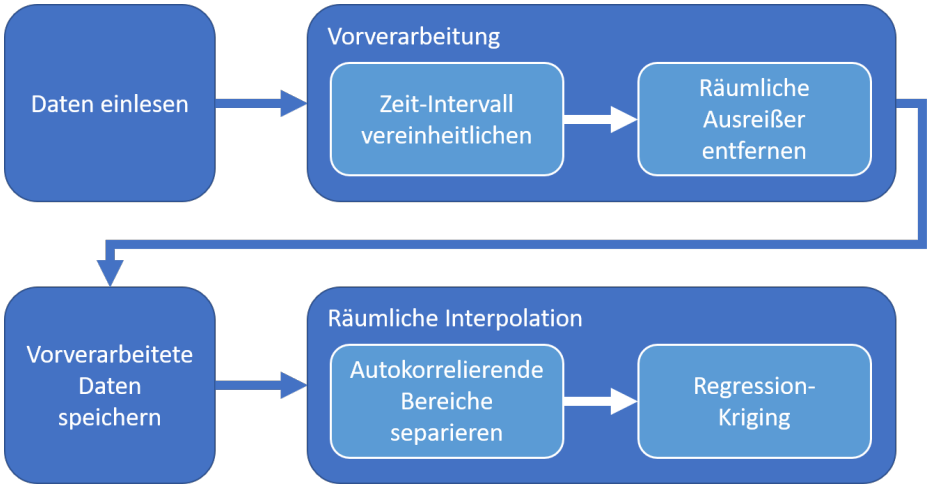

Abb. 4. Kriging-Ablauf

Requests Das *Requests*-Modul ist eine HTTP-Bibliothek und ermöglicht es HTTP-Anfragen zu stellen [Reitz, 03.08.2021]. Diese Modul wird genutzt, um aus dem Archiv von Sensor.Community den Namen der Parquet-Dateien abzufragen.

pandas ist ein Python-Modul zur Verarbeitung von Daten [McKinney, 2010] [Jeff Reback et al., 2021]. Es bietet die Möglichkeit, Daten in einer Tabellenstruktur performant im Hauptspeicher zu manipulieren. Dabei können Daten aus diversen Formaten und direkt aus dem Internet geöffnet und in ebenso vielen Formaten gespeichert werden. Es gibt ähnliche Funktionalitäten wie zum Beispiel in SQL und zusätzlich noch dedizierte Funktionen zur Behandlung von unter anderem fehlenden Daten und Zeitreihen. *pandas* wird in dem Projekt für das gesamte Einlesen und Vorverarbeiten der Daten genutzt.

SciKit GStat [Mirko Mälicke et al., 2021] ist ein Python-Modul für Variogramm-Analysen und OK. Es bietet alle Funktionen des *Kriging*-Prozesses. Es ist möglich, eine Minimalmenge an verfügbaren Messpunkten anzugeben, damit eine Schätzung fundiert getroffen wird. Allerdings müssen alle Parameter manuell gesetzt werden.

GSTools [Müller and Schüler, 2021] bietet eine ähnliche Funktionalität wie *SciKit GStat*, kann aber zusätzlich Variogramme automatisiert an die Daten anpassen. Weiter gibt es noch UK und KED als Interpolationsmodelle. Bisher konnte mit der automatisierten Anpassung der Variogramme nicht die gleiche Qualität der Ergebnisse von *SciKit GStat* erreicht werden.

Apache Parquet [par, 02.03.2021] ist ein freies und quelloffenes spaltenorientiertes Datenspeicherformat, welches ursprünglich aus dem Apache-Hadoop-Ökosystem stammt. Es bietet effiziente Datenkomprimierungs- und Kodierungsverfahren auch für großen Mengen an Daten an. *Apache Parquet* wird genutzt, um die Daten nach der Vorverarbeitung zwischenzuspeichern, ohne dabei Informationen über Datentypen zu verlieren.

GeoPandas [Kelsey Jordahl et al., 2020] ist eine Erweiterung von *pandas* für die Verarbeitung von geografischen Daten. *GeoPandas* wird vorwiegend für die Visualisierung von Daten im räumlichen Kontext genutzt.

Matplotlib ist ein weiteres Python-Modul, welches umfassende Möglichkeiten zur Visualisierung von Daten bietet. Außerdem ist es Backend vieler anderer Module zur graphischen Darstellung von Daten. *Matplotlib* wird direkt oder indirekt für alle Visualisierung genutzt.

statsmodel [Seabold and Perktold, 2010] ist ein Python-Modul für statistische Analysen, Tests und Datenexploration. *statsmodel* wird in dem Projekt bei der Datenexploration genutzt, für die Generierung von kumulativen Verteilungsdiagrammen.

6 Evaluation

Auch wenn OK die Kriging-Varianz bietet, beschreibt diese nur, wie gut das Model auf die Daten angepasst wurde. Wie gut das Modell bezüglich unbekannter Daten ist, kann nur über Validierungsdaten getestet werden. Dafür werden geschätzte Werte ($\hat{z}(s_j)$) mit tatsächlichen Beobachtungen an Validierungspunkten ($z^*(s_j)$) verglichen. Dabei werden hauptsächlich zwei Metriken genutzt.
Der mittlere Vorhersagefehler (*engl. mean prediction error*; ME):

$$ME = \frac{1}{l} \cdot \sum_{j=1}^{l} [\hat{z}(s_j) - z^*(s_j)]\,; \qquad E\{ME\} = 0 \tag{24}$$

Und die Wurzel aus dem mittleren, quadrierten Vorhersagefehler (*engl. root mean square prediction error*; RMSE):

$$RMSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} [\hat{z}(s_j) - z^*(s_j)]^2}; \qquad E\{RMSE\} = \sigma(h=0) \tag{25}$$

dabei ist l die Anzahl der Validierungspunkte.
Zusätzlich ist es möglich den Fehler auch auf der Basis der Vorhersagevarianz zu standardisieren, die durch das Modell gegeben wird:

$$RMNSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} \left[\frac{\hat{z}(s_j) - z^*(s_j)}{\hat{\sigma}_j} \right]^2}; \qquad E\{RMNSE\} = 1 \tag{26}$$

Grundsätzlich ist darauf zu achten, dass ME, RMSE und RMNSE nur auf Basis einer Teilmenge der Daten bestimmt werden. Ist die Teilmenge unglücklich gewählt, kann die Aussagekraft der Metriken sinken und die Qualität des Modells ist eventuell weniger gut als angenommen.
Um diese Problematik zu umgehen, bietet sich *Kreuzvalidierung (engl. cross-validation)* an. Dazu gibt es unterschiedliche Optionen [Bivand et al., 2013, S. 221-226]:

- *k-fold cross-validation*: Der Datensatz wird in k Teile geteilt und jeder Teil wird für die Kreuzvalidierung genutzt
- *leave-one-out cross-validation* (LOO): jeder einzelne Datenpunkt wird für die Kreuzvalidierung genutzt
- *Jackknifing*: ähnlich zu LOO, allerdings wird der Bias der Analyse und nicht die Interpolation betrachtet

7 Ergebnisse & Ausblick

In den bisherigen Versuchen wurde ausschließlich Ordinary Kriging genutzt. Die Ergebnisse sind im Anhang in den Abbildungen 9, 10, 11 und 12 dargestellt. Grundsätzlich ließ sich ein regelmäßiges Gitter schätzen und es ist möglich ein beschreibendes Variogramm für die Daten zu finden. Auf der Interpolation, welche mit SciKit GStat erstellt wurde, sind nicht interpolierte Bereiche erkennbar. Diese entstehen dadurch, dass nicht ausreichend Messungen im Umkreis für eine sichere Schätzung waren. Die Abbildung der Kriging-Varianz ist leider gestört. Ein ähnliches Ergebnis gab es leider bei allen Versuchen und ist vermutlich auf das Tool zurückzuführen.
Bei der Interpolation mit GSTools wurde das Variogramm automatisch angepasst. Die Kriging-Varianz ist allerdings deutlich größer, im Vergleich zur Umsetzung mit SciKit GStat und ebenfalls im Verhältnis zu den Messungen.
Die in Kapitel 6 beschriebenen Metriken wurden bisher noch nicht angewendet, werden aber in dem nächsten Schritt des Projekts implementiert.
Bisher wurden Ausreißer mithilfe der 2-Sigma-Regel identifiziert und durch einen *moving median* ersetzt. In Zukunft werden Ausreißer mit dem vorgestellten *spatial outlier detection*-Algorithmen identifiziert und entfernt.
Derzeit konnte noch kein beschreibendes Variogramm für die gesamte Fläche Deutschlands gefunden werden. Daher wird versucht *moving window kriging* umzusetzen.

Außerdem wird im nächsten Schritt des Projekts die zeitliche Dimension betrachtet. Dabei soll ein LSTM-Netz trainiert werden, um Prognosen über die Entwicklung der Feinstaubbelastung zu machen. Außerdem soll mithilfe von Clusteranalysen Anhäufungen von Feinstaub identifiziert werden. Ebenfalls soll die Bewegung dieser Cluster verfolgt und prognostiziert werden.

8 Danksagung

An dieser Stelle möchte ich mich bei Sensor.Community bedanken. Zum einen für das Projekt selbst, die Bereitstellung der Daten, aber insbesondere für die Hinweise und Hilfestellungen die mir gegeben wurden.

Literaturverzeichnis

(02.03.2021) Apache parquet. URL https://parquet.apache.org/

Anselin L (2002) Under the hood issues in the specification and interpretation of spatial regression models. Agricultural Economics 27(3):247–267, https://doi.org/10.1111/j.1574-0862.2002.tb00120.x

Bidanset P, Lombard J (2014) The effect of kernel and bandwidth specification in geographically weighted regression models on the accuracy and uniformity of mass real estate appraisal. Journal of Property Tax Assessment & Administration 10(3), URL https://digitalcommons.odu.edu/publicservice_pubs/27

Bivand R, Pebesma EJ, Gómez-Rubio V (2013) Applied spatial data analysis with R, 2nd edn. Use R!, Springer, New York, NY, URL http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=601853

Braatz A (2019) Raumzeitliches data-mining in dynamischen sensorsystemen: Raumzeitliches data-mining in dynamischen sensorsystemen. Bachelor's thesis, Hochschule für angewandte Wissenschaften Hamburg, URL https://reposit.haw-hamburg.de/handle/20.500.12738/9152?&locale=de

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis 28(4):281–298, https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Chen D, Lu CT, Kou Y, Chen F (2008) On detecting spatial outliers. GeoInformatica 12(4):455–475, https://doi.org/10.1007/s10707-007-0038-8, URL https://link.springer.com/article/10.1007/s10707-007-0038-8

Christensen R (1991) Linear Models for Multivariate, Time Series, and Spatial Data. Springer Texts in Statistics, Springer New York, New York, NY and s.l., https://doi.org/10.1007/978-1-4757-4103-2

Cressie NAC (2001) Statistics for spatial data, rev. ed. edn. Wiley series in probability and mathematical statistics Applied probability and statistics, Wiley, New York, https://doi.org/10.1002/9781119115151, URL http://onlinelibrary.wiley.com/book/10.1002/9781119115151

Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. Wiley Series in Probability and Statistics, Wiley, New York and Chichester

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3):37, https://doi.org/10.1609/aimag.v17i3.1230, URL https://ojs.aaai.org//index.php/aimagazine/article/view/1230

Florax RJ, Nijkamp P (2003) Misspecification in linear spatial regression models. Tinbergen Institute Discussion Papers 2003-081(3), https://doi.org/10.2139/ssrn.459500

Hengl T (2007) A practical guide to geostatistical mapping of environmental variables, EUR. Scientific and technical research series, vol 22904. Publications Office, Luxembourg

Heuvelink G, Webster R (2001) Modelling soil variation: past, present, and future. Geoderma 100(3-4):269–301, https://doi.org/10.1016/S0016-7061(01)00025-8

Jeff Reback, jbrockmendel, Wes McKinney, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, Simon Hawkins, gfyoung, Sinhrks, Matthew Roeschke, Adam Klein, Terji Petersen, Jeff Tratner, Chang She, William Ayd, Patrick Hoefler, Shahar Naveh, Marc Garcia, Jeremy Schendel, Andy Hayden, Daniel Saxton, Richard Shadrach, Marco Edward Gorelli, Vytautas Jancauskas, Fangchen Li, attack68, Ali McMaster, Pietro Battiston, Skipper Seabold, Kaiqi Dong (2021) pandas-dev/pandas: Pandas 1.3.1. https://doi.org/10.5281/zenodo.3509134

Jin Li, Andrew D Heap (2008) A Review of Spatial Interpolation Methods for Environmental Scientists - data.gov.au, vol 2008. Geoscience Australia, URL `https://data.gov.au/data/dataset/a-review-of-spatial-interpolation-methods-for-environmental-scientists`

Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, Geir Arne Hjelle, Alan D Snow, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Nick Eubank, maxalbert, Aleksey Bilogur, Sergio Rey, Christopher Ren, Dani Arribas-Bel, Leah Wasser, Levi John Wolf, Martin Journois, Joshua Wilson, Adam Greenhall, Chris Holdgraf, Filipe, François Leblanc (2020) geopandas/geopandas: v0.8.1. https://doi.org/10.5281/ZENODO.3946761

Kutner MH (2005) Applied linear statistical models, 5th edn. McGraw-Hill/Irwin series Operations and decision sciences, McGraw-Hill Irwin, Boston, Mass., URL `http://www.loc.gov/catdir/bios/mh042/2004052447.html`

luftdateninfo – Feinstaub selber messen (05.01.2021) Home - luftdaten.info – feinstaub selber messen. URL `https://luftdaten.info/`

Matheron G (1962) Traité de géostatistique appliquée. Bureau de recherches géologiques et minières (France)., Paris, URL `http://worldcatlibraries.org/wcpa/oclc/491866302`

Matheron G (1969) Le krigeage universel, vol 1. École nationale supérieure des mines de Paris Paris

McKinney W (2010) Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, SciPy, Proceedings of the Python in Science Conference, pp 56–61, https://doi.org/10.25080/Majora-92bf1922-00a

Mirko Mälicke, Egil Möller, Helge David Schneider, Sebastian Müller (2021) mmaelicke/scikit-gstat: A scipy flavoured geostatistical variogram analysis toolbox. https://doi.org/10.5281/ZENODO.4835779

Müller S, Schüler L (2021) Geostat-framework/gstools: v1.3.0 'pure pink'. https://doi.org/10.5281/ZENODO.1313628

Reitz K (03.08.2021) Requests: Http for humans™ — requests 2.26.0 documentation. URL `https://docs.python-requests.org/en/master/`

Ripley BD (2005) Spatial Statistics, Wiley Series in Probability and Statistics, vol v.575. John Wiley & Sons Inc, Hoboken, URL

http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=129384
Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference
Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: Blue RB, Rosenberg AM (eds) Proceedings of the 1968 23rd ACM national conference on -, ACM Press, New York, New York, USA, pp 517–524, https://doi.org/10.1145/800186.810616
Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Economic geography 46(sup1):234–240
van Rossum G, Drake FL (2009) Python 3 Reference Manual. CreateSpace, Scotts Valley, CA
Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Statistics in practice, Wiley, Chichester and Hoboken, NJ, https://doi.org/10.1002/9780470517277
Wheeler DC, Páez A (2010) Geographically weighted regression. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis, Springer, Berlin, pp 461–486, https://doi.org/10.1007/978-3-642-03647-7_22

9 Anhang

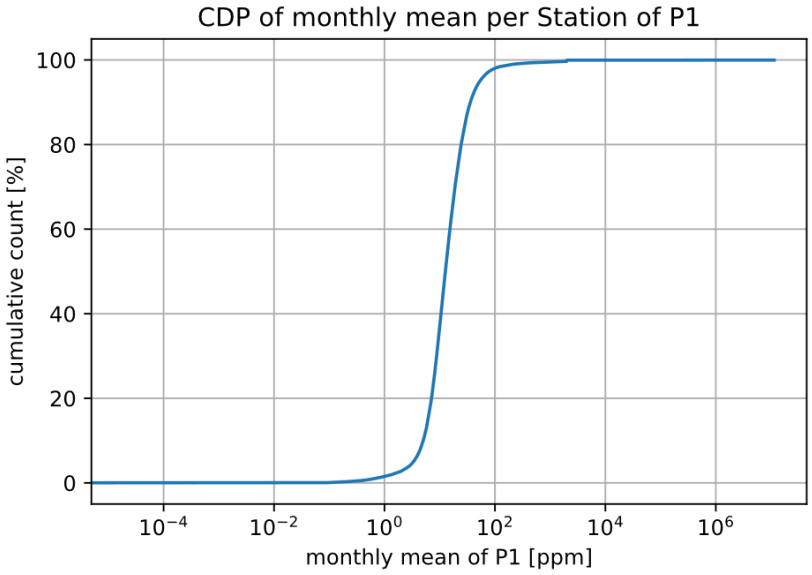

Abb. 5. Sensorstationen in Deutschland

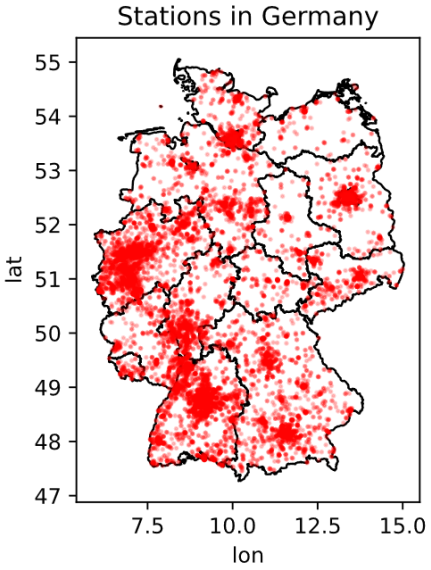

Abb. 6. Sensorstationen in Deutschland

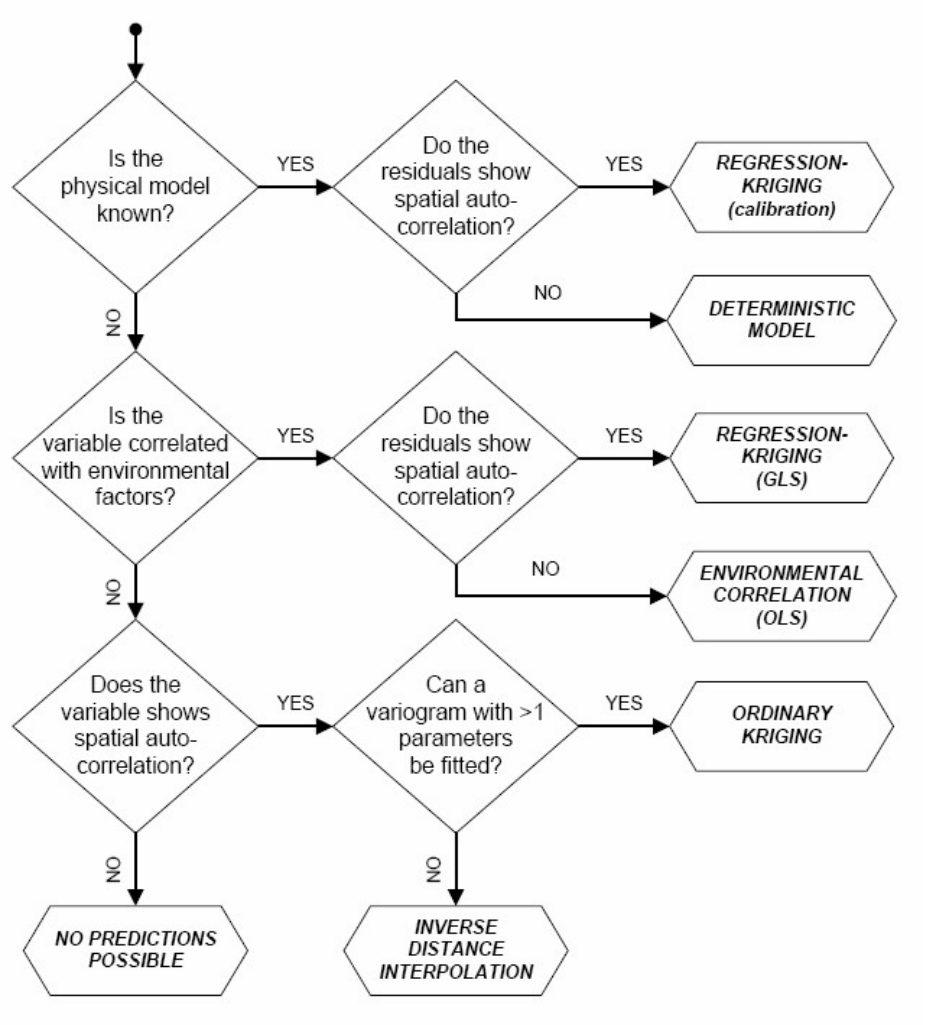

Abb. 7. Grundsätzlicher Entscheidungsbaum zur Auswahl eines geeigneten räumlichen Vorhersagemodells[Hengl, 2007, Abb. 2.3]

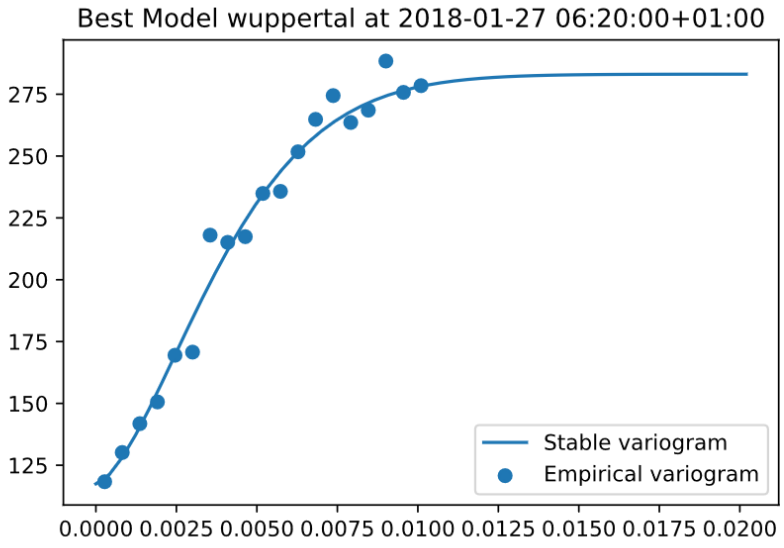

Abb. 8. Ein Variogramm für Wuppertal

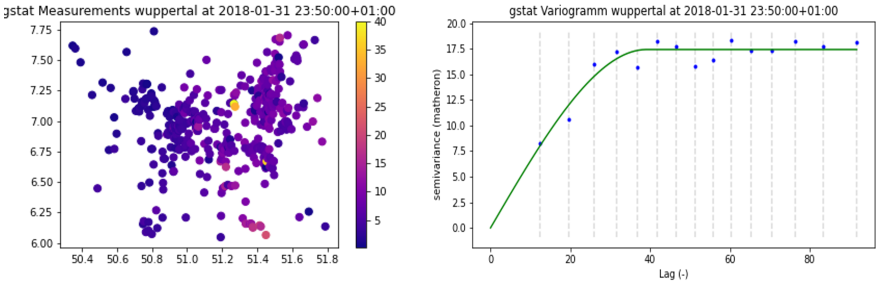

Abb. 9. Variogramm mit SciKit GStat

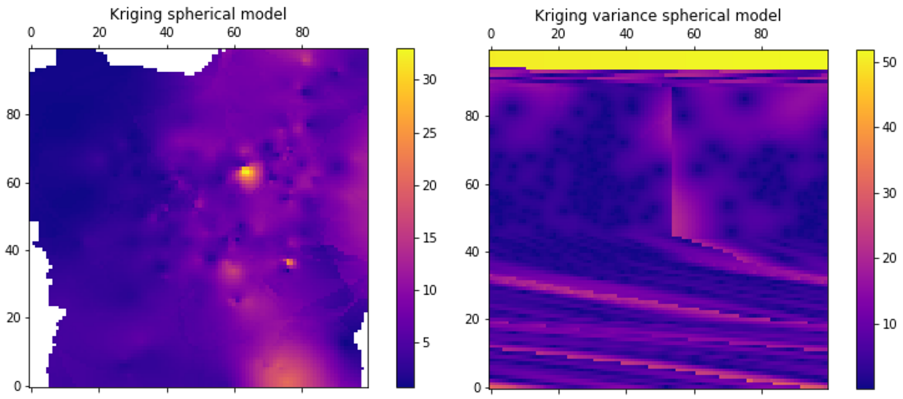

Abb. 10. Ordinary Kriging mit SciKit GStat

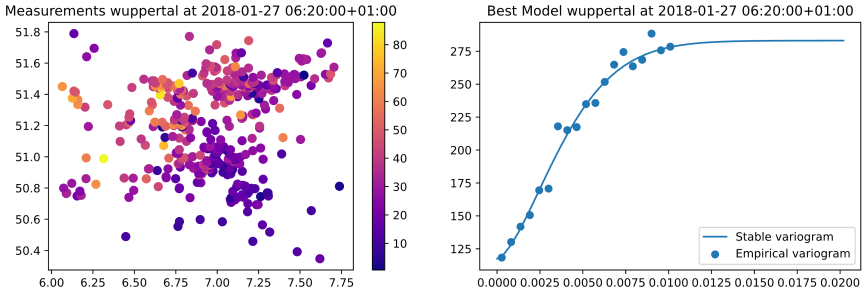

Abb. 11. Variogramm mit GSTools

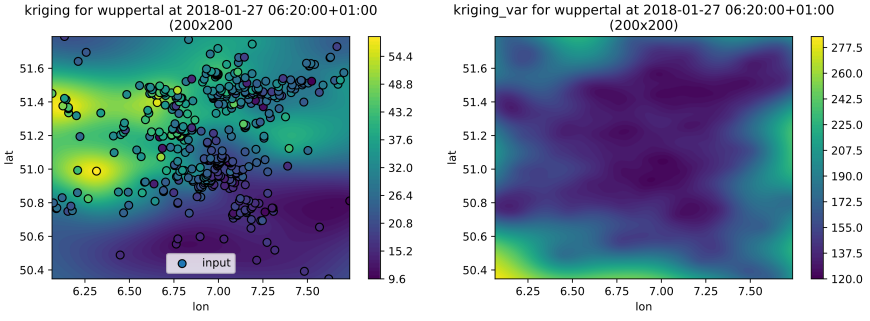

Abb. 12. Ordinary Kriging mit GSTools