Geostatistik und räumliche Vorhersage

am Beispiel von Angebotsmietdaten in Magdeburg

Georg Wiegleb
Amt für Statistik, Wahlen und Digitalisierung Magdeburg
Erfurt, 16.09.2022

Einleitung

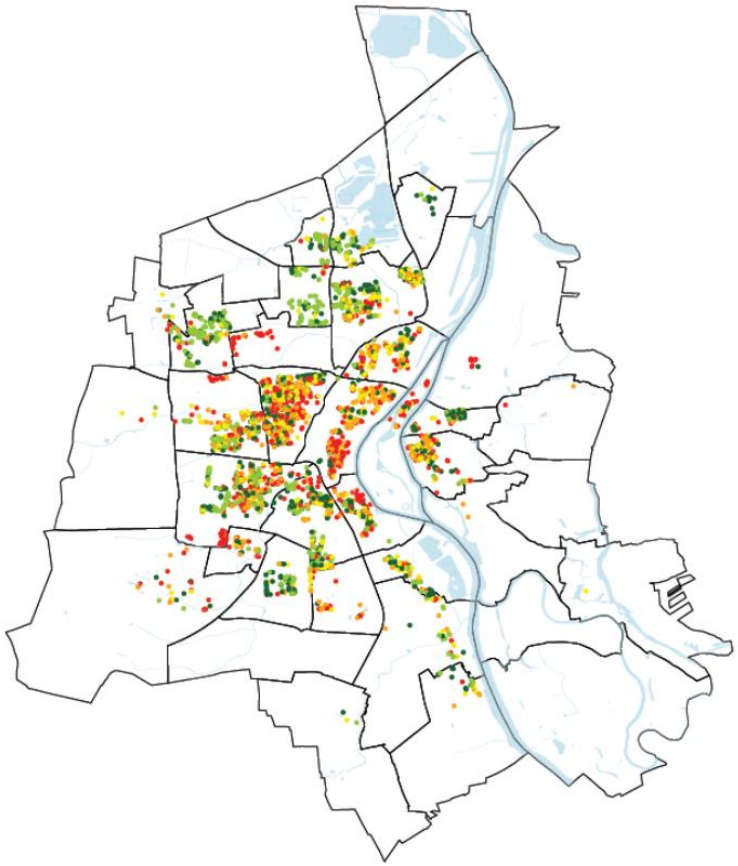
> "... everything is related to everything else, but near things are more related than distant things"
>
> — Tobler's first law of geography

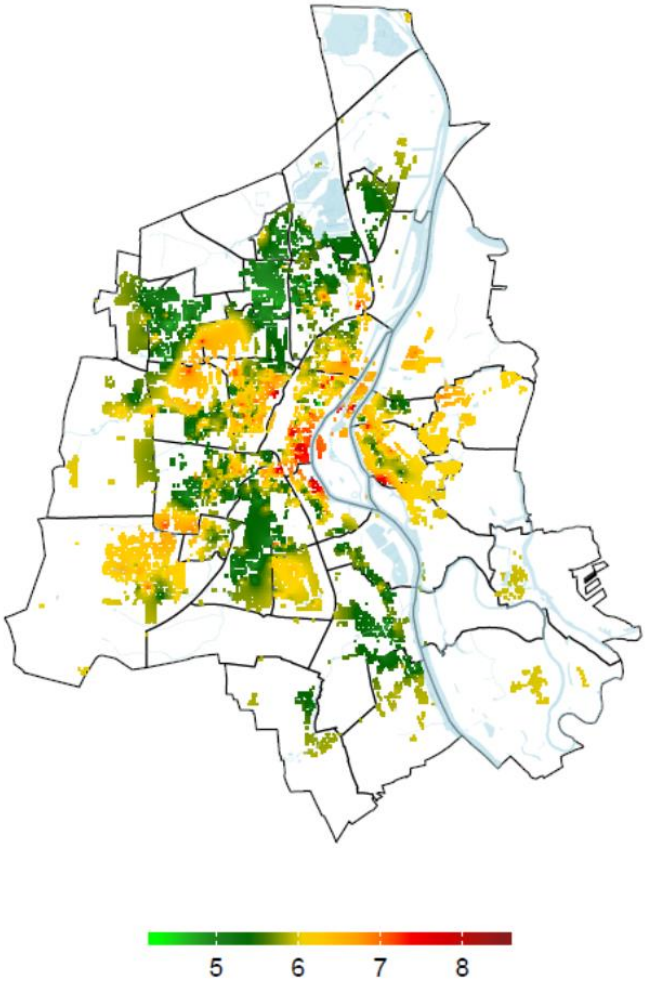
- Wohnungsmietpreise als **prominenter Untersuchungsgegenstand**
- Zentrales Anliegen: **Vorhersage/Prädiktion** von Mietpreisen anhand anderer Merkmale
- Einfluss der **Lage** allgemein anerkannt, aber **problematisch**
- In georeferenzierten Daten liegt großes (häufig ungenutztes) Potenzial Vorhersage/Prädiktion von Mietpreisen zu verbessern

Geostatistik

- Geostatistik Sammlung von stochastischen Methoden zur Modellierung georeferenzierter Daten unter Annahme räumlicher Autokorrelation.
- **Ziel**: Schätzung und Vorhersage/Prädiktion auf Basis der Positionen von beobachteten Untersuchungsobjekten (Kriging)

Angebotsmietpreise in Magdeburg





(Mathematische) Grundlagen

Allg. | Z | Stationarität | Isotropie

- **Traditionelle Statistik**: Merkmalsträger/ Objekte voneinander unabhängig (iid etc.)
- **Geostatistik**: Räumliche Abhängigkeit, Phänomen ist stochastischer Prozess (Zufallsfeld), Abhängigkeit durch Abstand
- $Z(x)$ ist Zufallsvariable in Abhängigkeit von $x \in \mathbb{R}^2$ als Koordinate, z.B. $Z((52.12, 11.65))$
- Zusammenhang über Kovarianzfunktion $C(x, y)$ oder Variogramm $\gamma(x, y)$ mit x und y als 2 Punkten im relevanten Raum $D \subseteq \mathbb{R}^2$
- Annahmen: (intrinsische) Stationarität und Isotropie -> γ und C nur vom (euklidischen) Abstand $|h| = ||y - x||$ abhängig

(Mathematische) Grundlagen

Zufallsfeld

Sei $D \subseteq \mathbb{R}^2$ und $Z(x): \Omega \to \mathbb{R}$ für alle $x \in D$ messbar.

$$Z := \left(Z(x)\right)_{x \in D}$$

heißt Zufallsfeld Z.

Kovarianzfunktion

$$C(x, y) = Cov[Z(x), Z(y)]$$

Variogramm

$$\gamma(x, y) = \frac{1}{2} Var[Z(y) - Z(x)]$$

(Mathematische) Grundlagen

Stationarität

Z heißt **stationär**, wenn $\mathbb{E}[(Z(x)] \equiv \mu$ und wenn C verschiebungsinvariat ist. Für den Abstandsvektor $h = y - x$ gilt: $R(h) := C(x, x + h) = C(x, y)$.

Intrinsische Stationarität

Z heißt **intrinsisch stationär**, wenn für alle $h \in D$ $\left(Z(x + h) - Z(x)\right)_{x \in D}$ Stationarität und Zentriertheit ($\mathbb{E} \equiv 0$) vorliegt. In diesem Fall ist das Variogramm eine Funktion von h: $\gamma(h)$.

- Stationäre Zufallsfelder sind auch intrinsisch stationär (gilt nicht umgekehrt)

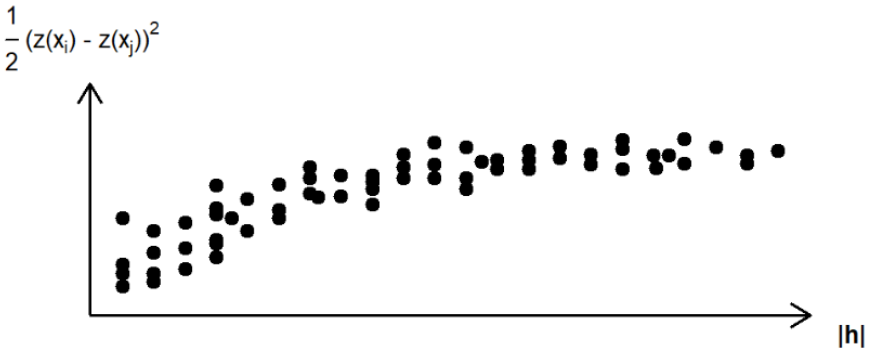
(Mathematische) Grundlagen

Isotropie

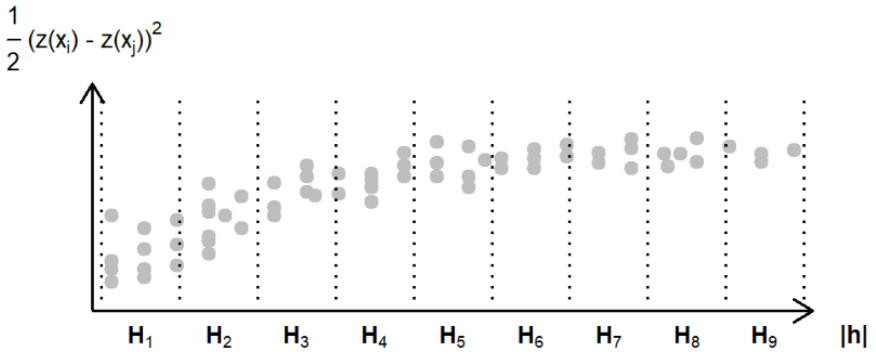
Z heißt **isotrop**, wenn es stationär ist und die Kovarianzfunktion invariat gegenüber der Ausrichtung des Abstandsvektors h ist. In diesem Fall schreibt man $R_0(|h|) = C(x, y)$

- Im isotropen Fall ist nur die Länge des Abstandsvektors bzw. der (euklidische) Abstand zwischen beiden Punkten relevanten
- Analog gilt im isotropen dass sich das Variogramm als eine Funktion des Abstands zwischen zwei Punkten darstellen lässt

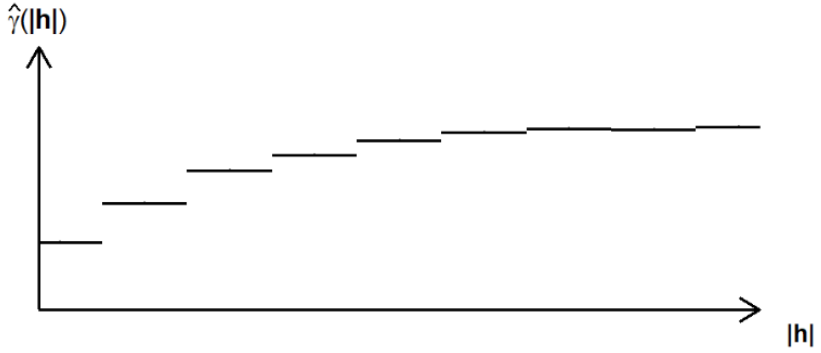
Variographie



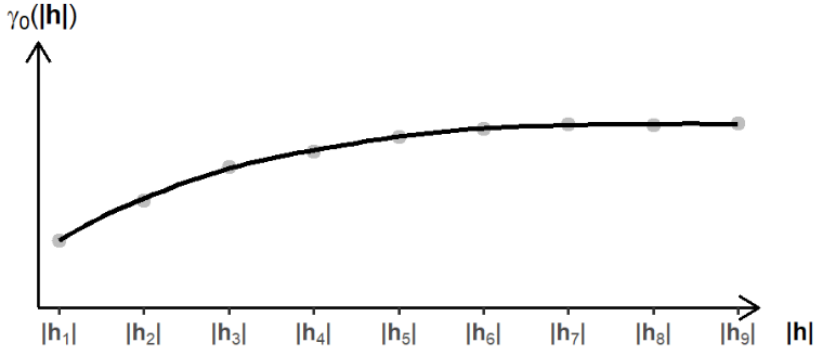
(a) Erstellung der Variogrammwolke



(b) Unterteilung in Distanzgruppen

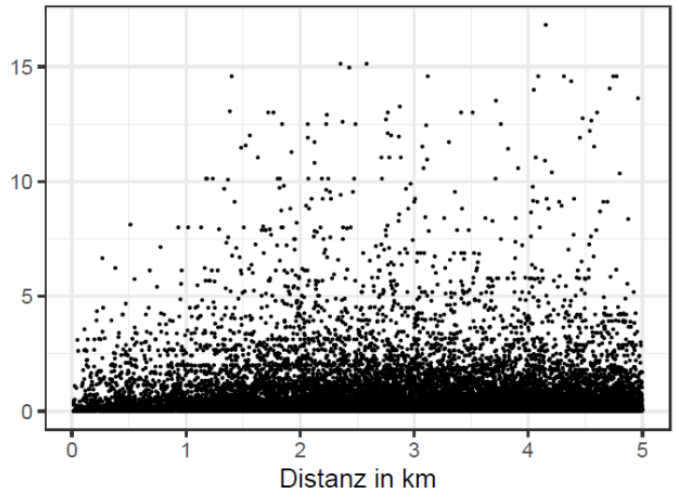


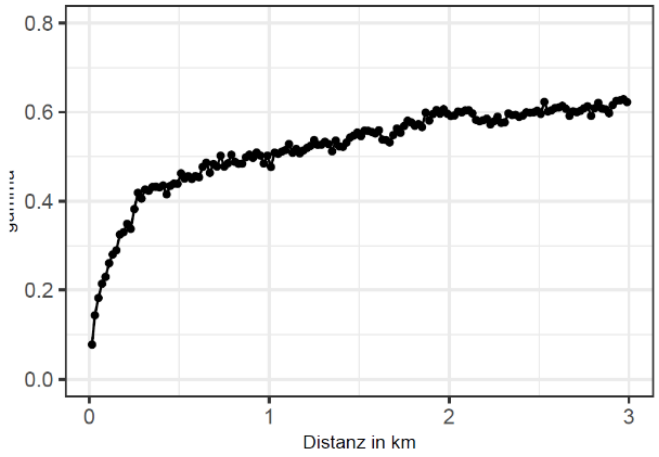
(c) Bestimmung des empirischen Variogramms

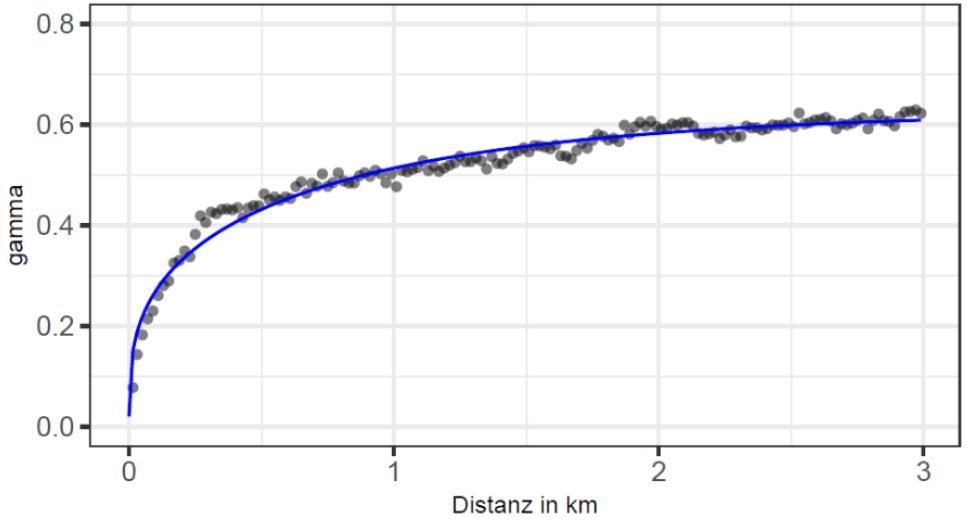


(d) Anpassung der Variogrammfunktion

Variographie



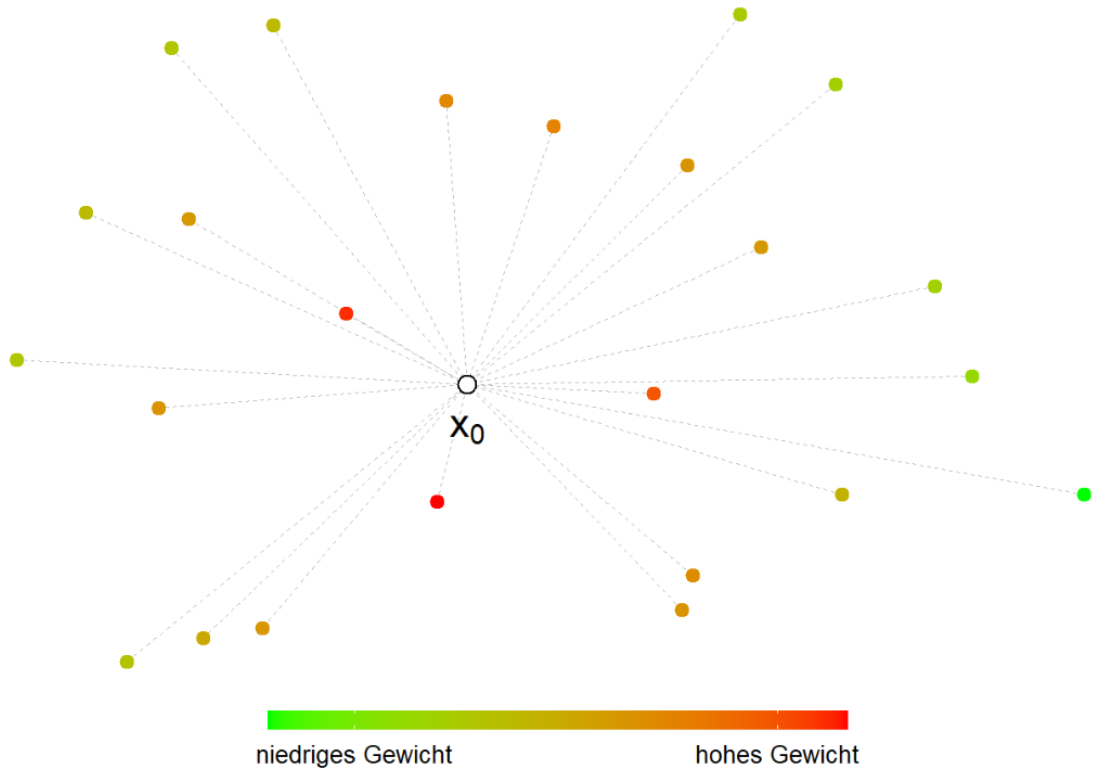




Räumliche Vorhersage

Beobachtungen $z(x_1), \ldots, z(x_n)$ aus Z. Was ist eine **beste Vorhersage/Prädiktion** für $z(x_0)$ für $x_0 \in D$?

Gewichtung und Abstand | Linearer Prädiktor



Räumliche Vorhersage

Beobachtungen $z(x_1), \ldots, z(x_n)$ aus Z. Was ist eine **beste Vorhersage/Prädiktion** für $z(x_0)$ für $x_0 \in D$?

Ein **linearer Prädiktor** $\hat{Z}(x_0)$ für $Z(x_0)$ an der Position x_0 ist definiert als:

$$\hat{Z}(x_0) := \sum_{i=1}^{n} w_i Z(x_i) = \boldsymbol{w}^T Z(\boldsymbol{x})$$

mit $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ und $Z(\boldsymbol{x}) = (Z(x_1), \ldots, Z(x_n))^T$.

$\hat{Z}^*(x_0)$ ist bester, linearer, unverzerrter Prädiktor (BLUP) für $Z(x_0)$, wenn gilt:

$$\textbf{Unverzerrtheit} : Bias[\hat{Z}^*(x_0)] := \mathbb{E}[\hat{Z}^*(x_0) - Z(x_0)] = 0$$

$$\textbf{Minimale Varianz} : Var[Z^*(x_0) - \hat{Z}(x_0)] \leq Var[\hat{Z}(x_0) - Z(x_0)]$$

für alle linear unverzerrten Prädiktoren $\hat{Z}(x_0)$.

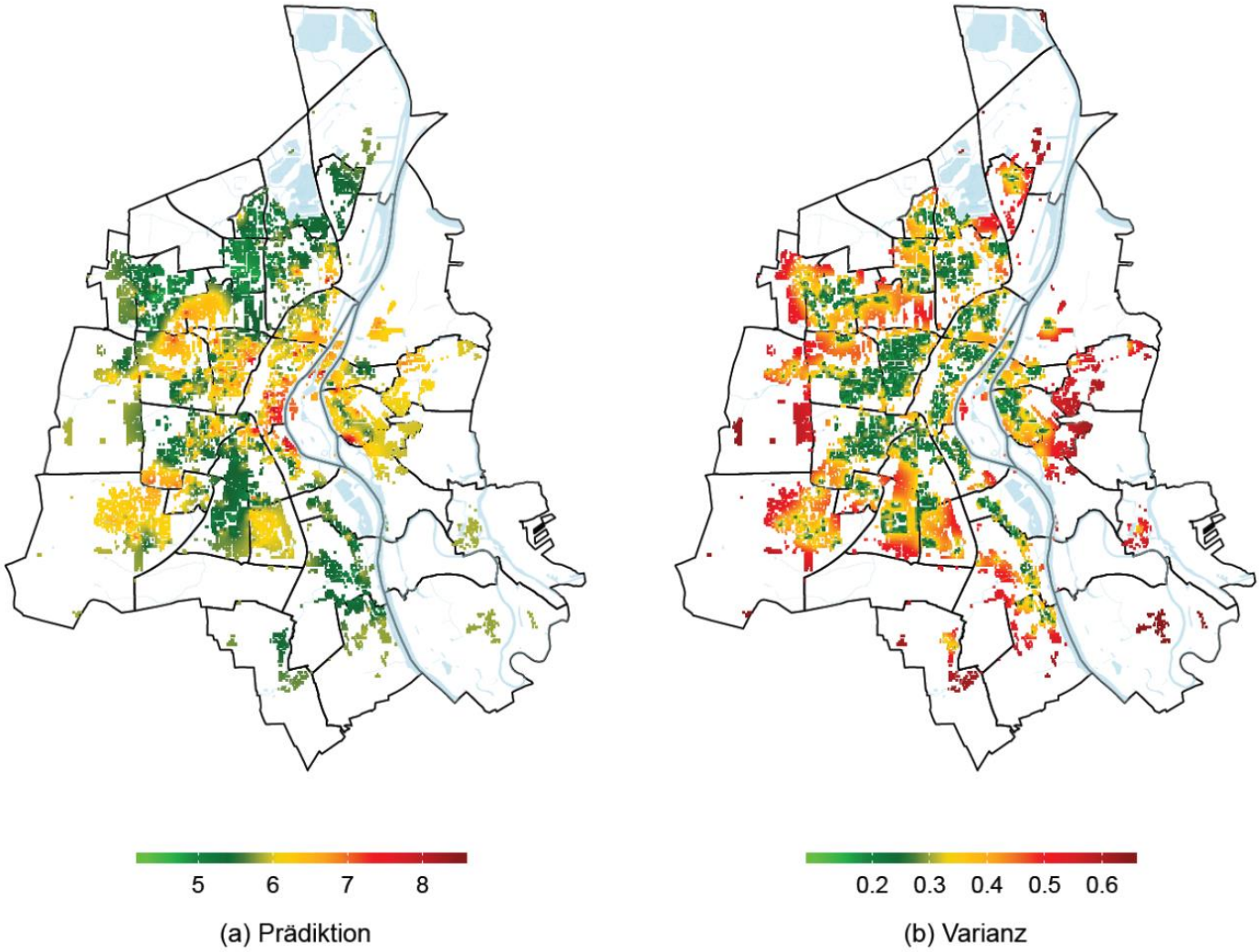
Ordinary Kriging

Annahmen

- Mietpreis wird durch einen durchschnittlichen Wert und ein Zufallsfeld bestimmt
- Der Durchschnittswert ist nicht bekannt
- Das Zufallsfeld hat eine gewisse Stabilität (intrinsische Stationarität + Isotropie)
- Das Variogramm ist bekannt bzw. wurde aus den Daten geschätzt

Dann ist der Ordinary-Kriging-Vorhersage die statistisch beste Prädiktion (BLUP)

Ordinary Kriging



(a) Prädiktion

(b) Varianz

Ordinary Kriging

- Beobachtungen aus intrinisch stationären Zufallsfeld
- konstanter, aber unbekannter Erwartungswert μ

Gegeben Sei ein intrinsisch stationäres Zufallsfeld Z mit $\mathbb{E} \equiv \mu \in \mathbb{R}$ unbekannt und Variogramm γ. Für beobachtete Positionen $x_1, \dots, x_n \in D$ sei $\boldsymbol{V} := [\gamma(x_i - x_j)]_{i,j=1}^n$ invertierbar. Dann ist

$$\hat{Z}(x_0) := \boldsymbol{w}_{OK}^T Z(\boldsymbol{x})$$

BLUP für $Z(x_0)$, wenn $\boldsymbol{w}_{OK} := \boldsymbol{V}^{-1}\left(\boldsymbol{v}_0 - \mathbf{1}_n\left(\frac{\mathbf{1}_n^T \boldsymbol{V}^{-1}\boldsymbol{v}_0 - 1}{\mathbf{1}_n^T \boldsymbol{V}^{-1}\mathbf{1}_n}\right)\right)$ mit

$\boldsymbol{v}_0 := (\gamma(x_1 - x_0), \dots, \gamma(x_n - x_0))^T$.

Ordinary Kriging

$$\begin{pmatrix} 0 & \gamma(x_1 - x_2) & \cdots & \gamma(x_1 - x_n) & 1 \\ \gamma(x_2 - x_1) & 0 & \cdots & \gamma(x_2 - x_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(x_n - x_1) & \gamma(x_n - x_2) & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1^{OK} \\ w_2^{OK} \\ \vdots \\ w_n^{OK} \\ \lambda \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \gamma(x_2 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}$$

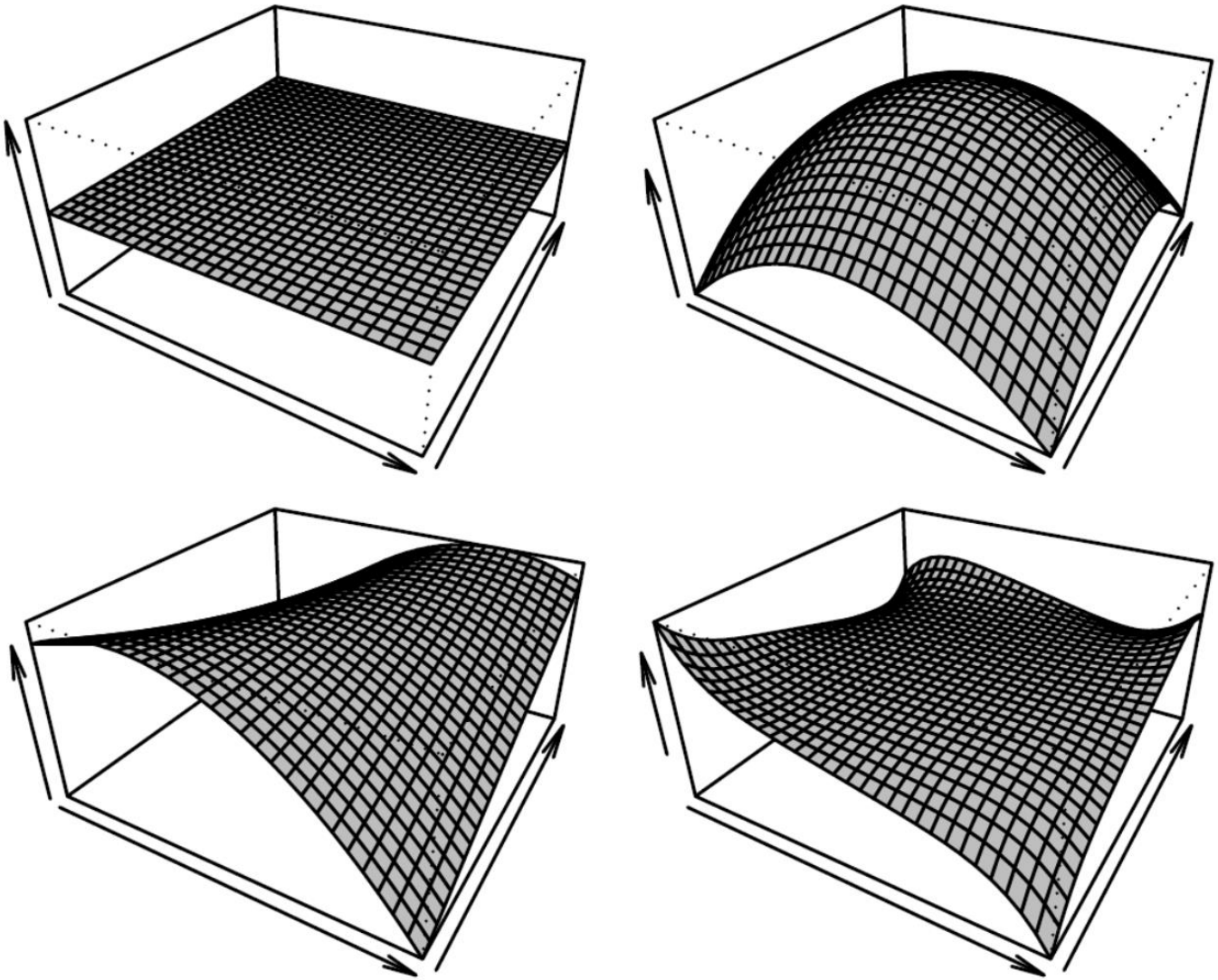
Universal Kriging

Annahmen

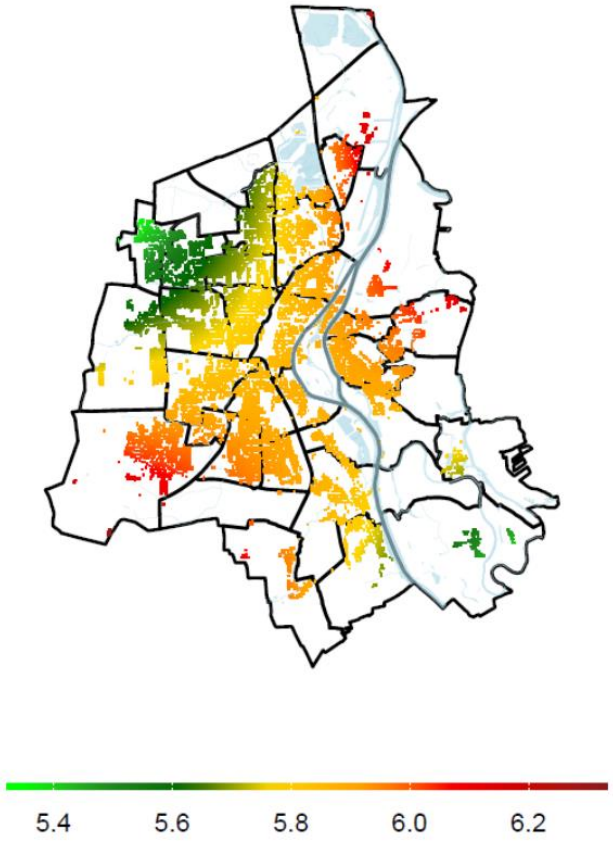
- Mietpreis wird durch eine Trendoberfläche (Durchschnittswert verändert sich) und ein Zufallsfeld bestimmt
- Die Trendoberfläche ist nicht bekannt
- Das Zufallsfeld hat eine gewisse Stabilität (intrinsische Stationarität + Isotropie)
- Das Variogramm ist bekannt bzw. wurde aus den Daten geschätzt

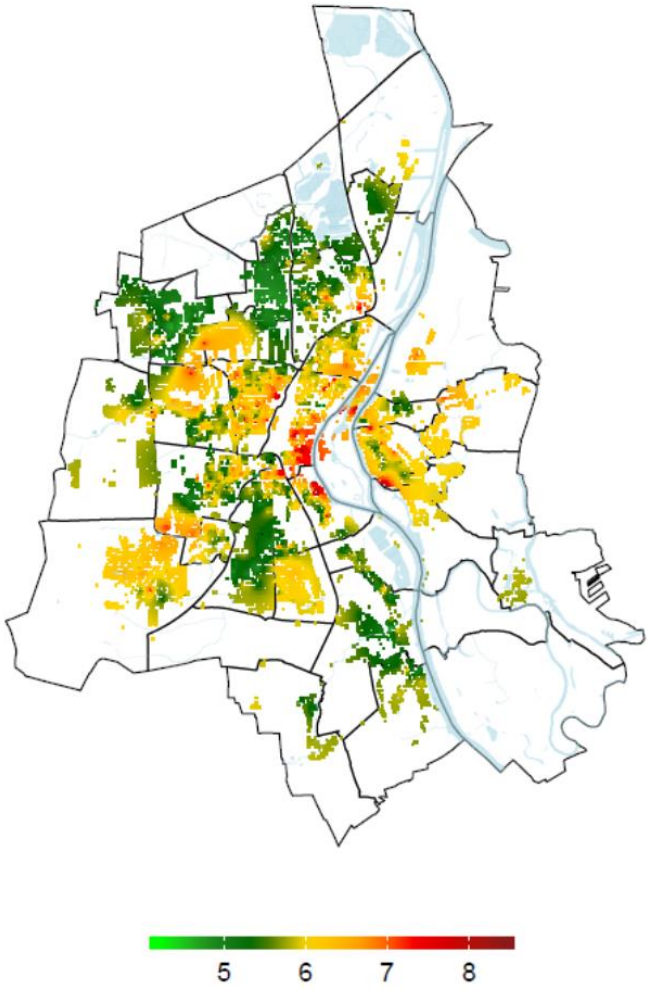
Dann ist der Universal-Kriging-Vorhersage die statistisch beste Prädiktion (BLUP)

Universal Kriging



Universal Kriging





Universal Kriging

- Zufallsfeld $Z(x) = m(x) + \epsilon(x)$ mit $m(x)$ deterministisch und $\epsilon(x)$ zufällig (universelles Modell spatialer Variation)
  - $m(x)$ wird Drift genannt und beschreibt eine (unbekannte) polynomiale Trendoberfläche, die sich durch ein lineares Modell schätzen lässt (Koordinaten als erklärende Variablen)
  - $\epsilon(x)$ ist ein intrinsisch stationäres, zentriertes Zufallsfeld

Sei Z ein Zufallsfeld mit den oben beschriebenen Annahmen. Sei $\boldsymbol{F}$ eine Designmatrix für ein allgemeines lineares Modell und $\boldsymbol{\alpha}$ der zugehörige Parametervektor für die polynomiale Gestalt der Trendoberfläche. Für beobachtete Positionen $x_1, \ldots, x_n \in D$ sei $\boldsymbol{V_\epsilon} := [\gamma_\epsilon(x_i - x_j)]_{i,j=1}^n$ invertierbar. Dann ist der Prädiktor $\hat{Z}(x_0) := \sum_{i=1}^n w_i Z(x_i) = \boldsymbol{w}^T Z(\boldsymbol{x})$ BLUP, wenn

$$\boldsymbol{w} = \boldsymbol{w}_{UK} := \boldsymbol{V_\epsilon}^{-1} \left(\boldsymbol{v}_{\epsilon,0} - \boldsymbol{F} \left(\boldsymbol{F}^T \boldsymbol{V_\epsilon}^{-1} \boldsymbol{F} \right)^{-1} \left(\boldsymbol{F}^T \boldsymbol{V_\epsilon}^{-1} \boldsymbol{v}_{\epsilon,0} - \boldsymbol{f}_0 \right) \right).$$

Universal Kriging

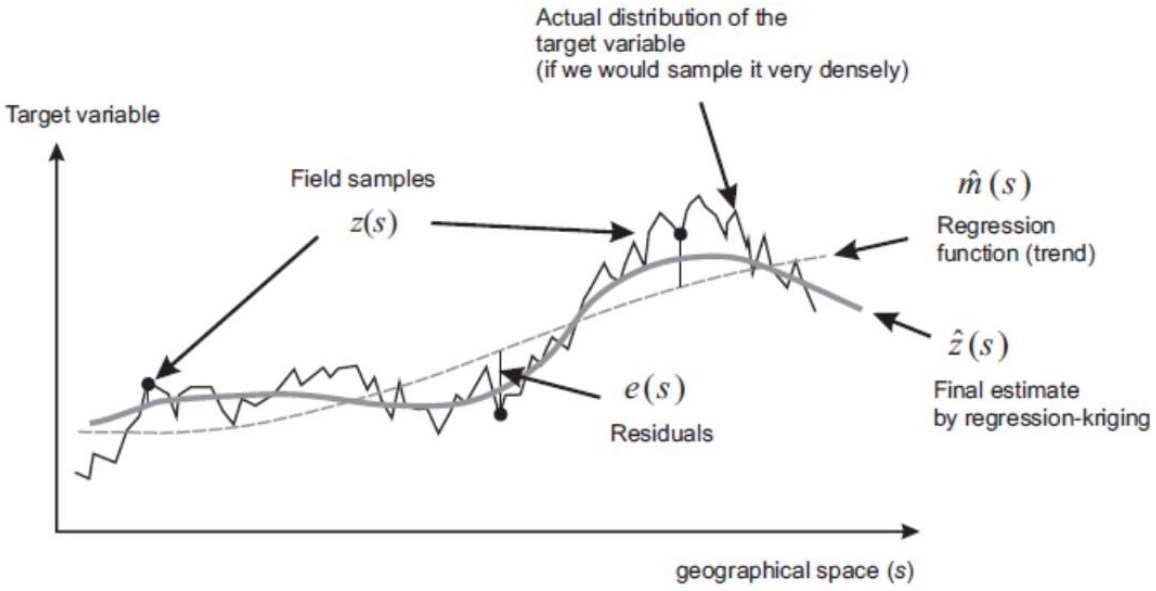
$$\left(\begin{array}{ccc|cccc} \gamma_\epsilon(x_1-x_1) & \cdots & \gamma_\epsilon(x_1-x_n) & 1 & f_1(x_1) & \cdots & f_L(x_1) \\ \gamma_\epsilon(x_2-x_1) & \cdots & \gamma_\epsilon(x_2-x_n) & 1 & f_1(x_2) & \cdots & f_L(x_2) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_\epsilon(x_n-x_1) & \cdots & \gamma_\epsilon(x_n-x_n) & 1 & f_1(x_n) & \cdots & f_L(x_n) \\ \hline 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(x_1) & \cdots & f_1(x_n) & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_L(x_1) & \cdots & f_L(x_n) & 0 & 0 & \cdots & 0 \end{array}\right) \left(\begin{array}{c} w_1^{UK} \\ w_2^{UK} \\ \vdots \\ w_n^{UK} \\ \hline \alpha_0^{UK} \\ \alpha_1^{UK} \\ \vdots \\ \alpha_L^{UK} \end{array}\right) = \left(\begin{array}{c} \gamma_\epsilon(x_1-x_0) \\ \gamma_\epsilon(x_2-x_0) \\ \vdots \\ \gamma_\epsilon(x_n-x_0) \\ \hline f_0(x_0) \\ f_1(x_0) \\ \vdots \\ f_L(x_0) \end{array}\right)$$

$$\begin{pmatrix} \boldsymbol{V}_\epsilon & \boldsymbol{F} \\ \boldsymbol{F}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_{UK} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_{\epsilon,0} \\ \boldsymbol{f}_0 \end{pmatrix}$$

Weitere Spielarten von Kriging

Regression-Kriging | Block-Kriging | Cokriging

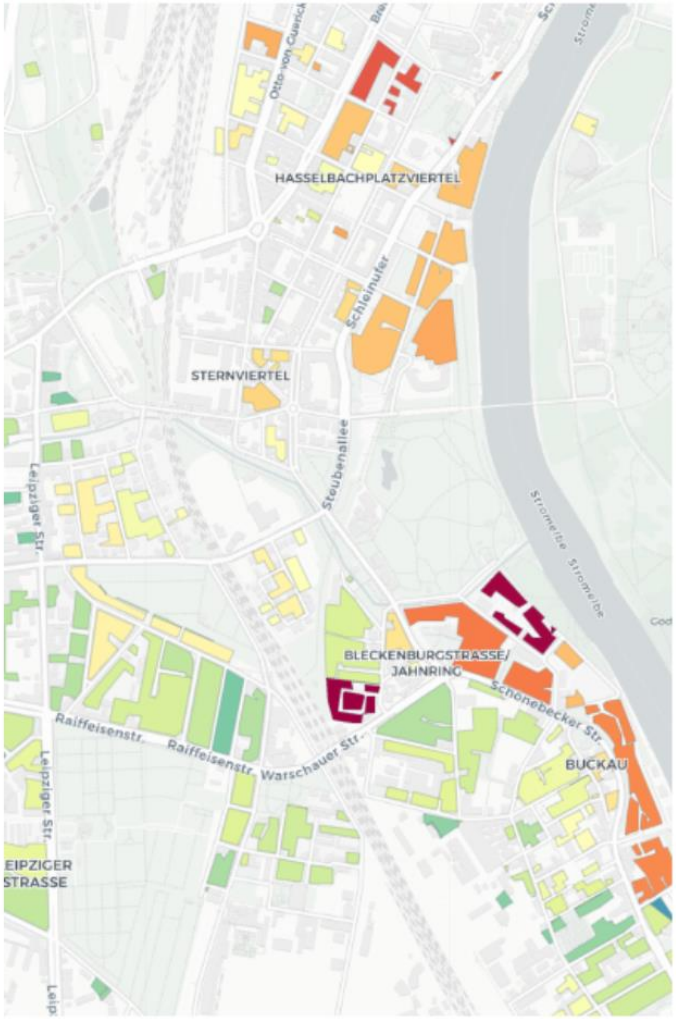
- auch: Gauss-Prozess-Regression
- Methodisch/ Mathematisch äquivalent zu Universal Kriging.
- Strukturelle Eigenschaften statt Geokoordinaten für "Trend" (z.B. Wohnungsgröße, Ausstattung, Baujahr, ...)
- $Z(x, s) = m(s) + \epsilon(x)$
  - $m(s)$ als deterministisch, Resultat der Regressionsfunktion mit Merkmalskombination s
  - $\epsilon(x)$ als Zufallsfeld (intrinisch stationär, zentriert, isotrop)



Weitere Spielarten von Kriging

Regression-Kriging | Block-Kriging | Cokriging

- Statt einer einzelnen Position (Point-Kriging) wird der erwartete Wert über ein gesamtes Polygon $P \subseteq D$ interpoliert
  - Stadtteile, Statistische Bezirke, Blöcke, Flur-Stücke,...
- Definiert als ein Integral des Zufallsfeldes über P, Annährung über Durchschnitte mehrerer Prädiktionen
- Besonderheit:
  - Auch Werte außerhalb der durch die Polygon definierten Bereiche werden berücksichtigt
  - keine Restriktion bzgl. Anzahl der Merkmalsträger
  - Bei Autokorrelation bessere Vorhersage (Standardfehler geringer)



Weitere Spielarten von Kriging

Regression-Kriging | Block-Kriging | **Cokriging**

- Natürliche Erweiterung von Kriging für den multivariaten Fall.
- Viele Anwendungsszenarien:
  - Simulatenes Kriging mehrerer abhängiger Variablen
  - Berücksichtigung mehrerer Stichproben
  - Verbesserung der Prädiktion einer räumlich weniger gut verteilten Variable durch Hilfsvariablen, die regelmäßiger beobachtet wurden (Collocated Cokriging)
  - Berücksichtigung der Korrelation von exogenen Merkmalen
- Denkbare Anwendungsfälle:
  - Verbesserung der Vorhersage von Bestandsmietdaten durch Hinzunahme von Angebotsmietdaten
  - Hinzunahme exogener Variablen, wie Bodenrichtwerte, Abstände zu Grünflächen etc., die an jeder Position ermittelelt werden können

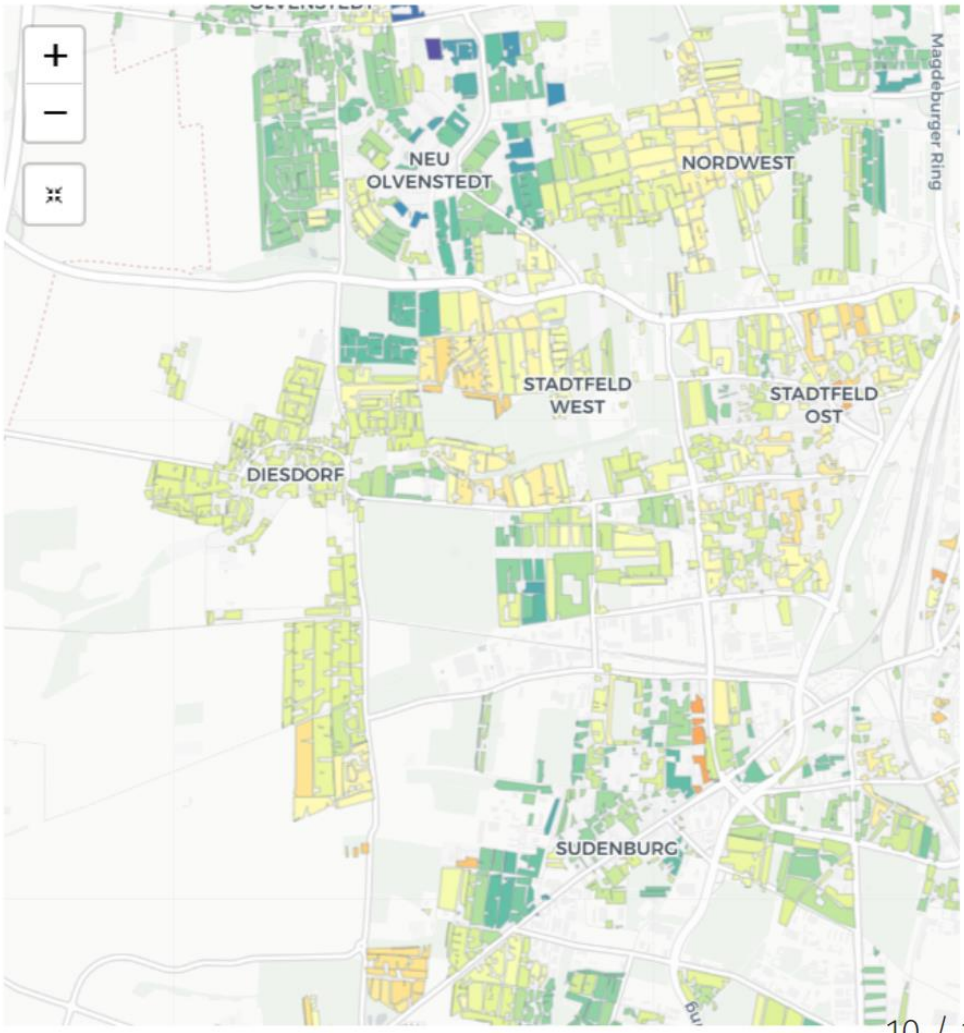
Regression-Block-Kriging Magdeburg

Regressionskoeffizienten

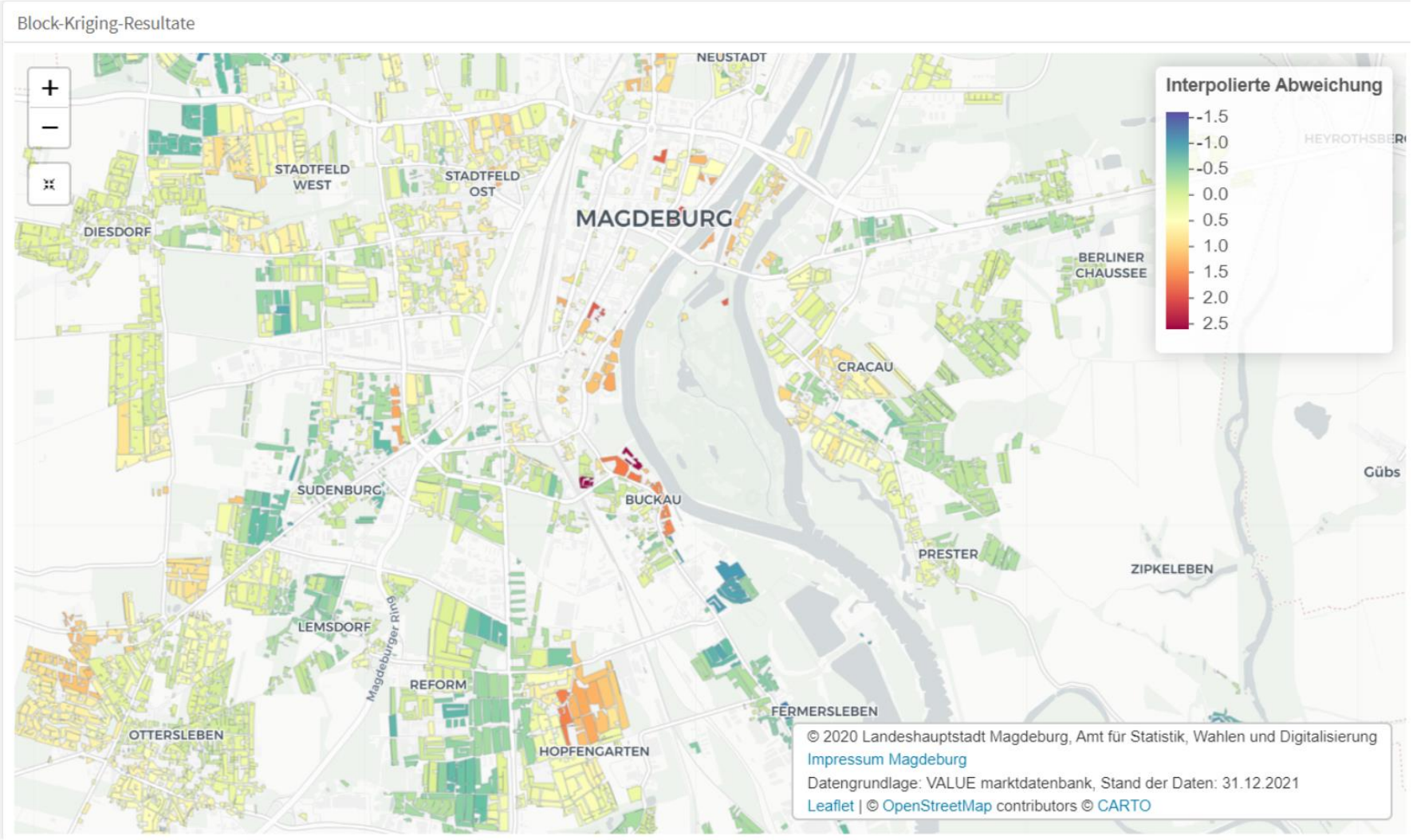
Download ▾ Filter

	Schätzung	P-Wert
Ausgangskaltmiete pro qm	6.473	0
40 bis 59 qm	-0.299	0
60 bis 79 qm	-0.333	0
80 bis 99 qm	-0.15	0.003
110 bis 119 qm	0.17	0.034
120 und mehr qm	0.224	0.012
Lift vorhanden	0.317	0
Einbauküche	0.449	0
Fußbodenheizung	1.057	0
Balkon oder Terasse	0.27	0
Neuwertig	0.17	0.03
Etage (1,2,3,...)	-0.039	0

Block-Kriging-Resultate



Regression-Block-Kriging Magdeburg



Verwendete und weiterführende Literatur

Chica-Olmo, J., Cano-Guervos, R. \& Chica-Rivas, M. (2019). Estimation of Housing Price Variations Using Spatio-Temporal Data. Sustainability, 11(6), 1551.

Chilès, Jean-Paul; Delfiner, Pierre (1999): Geostatistics. Modeling spatial uncertainty. New York, NY: Wiley (A Wiley-Interscience publication).

Cressie, Noel A. C. (1993): Statistics for spatial data. Rev. ed. New York: Wiley (Wiley series in probability and mathematical statistics Applied probability and statistics).

Del Giudice, V. & de Paola, P. (2017). Spatial Analysis of Residential Real Estate Rental Market with Geoadditive Models. In M. D'AMATO \& T. KAUKO (Hrsg.), Advances in Automated Valuation Modeling (Bd. 86, S. 155–162). Studies in Systems, Decision and Control.

Gaetan, Carlo; Guyon, Xavier (2010): Spatial Statistics and Modeling. New York, NY: Springer Science+Business Media LLC (Springer Series in Statistics).

Hengl, Tomislav (2009): A practical guide to geostatistical mapping. 2nd extended ed. Amsterdam: Hengl.

Kuntz, M. & Helbich, M. (2014). Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. International Journal of Geographical Information Science, 28(9), 1904–1921.

Lerbs, Oliver; Sebastian, Steffen (2015): Mietspiegel aus ökonomischer Sicht – Vorschläge für eine Neuregulierung.

Verwendete und weiterführende Literatur

Lichtenstern, Andreas (2013): Kriging methods in spartial statistics. Technische Universität München, München. Department of Mathematics.

Matheron, Georges (1971): The Theory of Regionalized Variables and Its Applications. Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu, Paris.

Olea, R. A. (1999). Geostatistics for Engineers and Earth Scientists.

Schernthanner, Harald (2017): Räumliche Analyse und Visualisierung von Mietpreisdaten. Dissertation. Universität Potsdam.

Stein, Michael L. (1999): Interpolation of Spatial Data. Some Theory for Kriging. New York, NY: Springer (Springer Series in Statistics).

Tobler, W. R. (1970): A Computer Movie Simulating Urban Growth in the Detroit Region. In: Economic Geography 46, S. 234.

Wackernagel, Hans (2003): Multivariate geostatistics. An introduction with applications : 7 tables. 3., completely revised ed. Berlin: Springer.