Aufgabenzettel 07

Gruppe 01

9 6 2020

Laden sie den Workspace yingtan_20_ueb3.Rdata sowie das Paket gstat und überführen sie das Objekt ljz in ein SpatialPointsDataFrame. Reproduzieren sie ihr Variogrammmodell aus Übung 05.

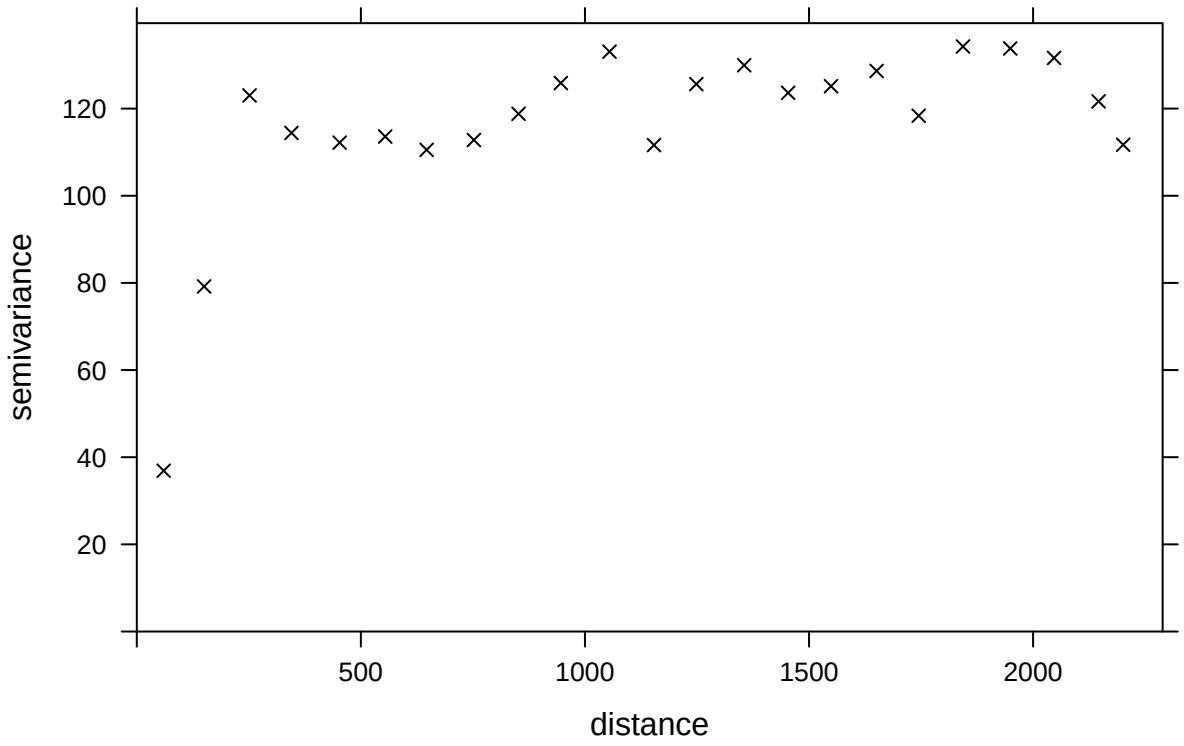
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

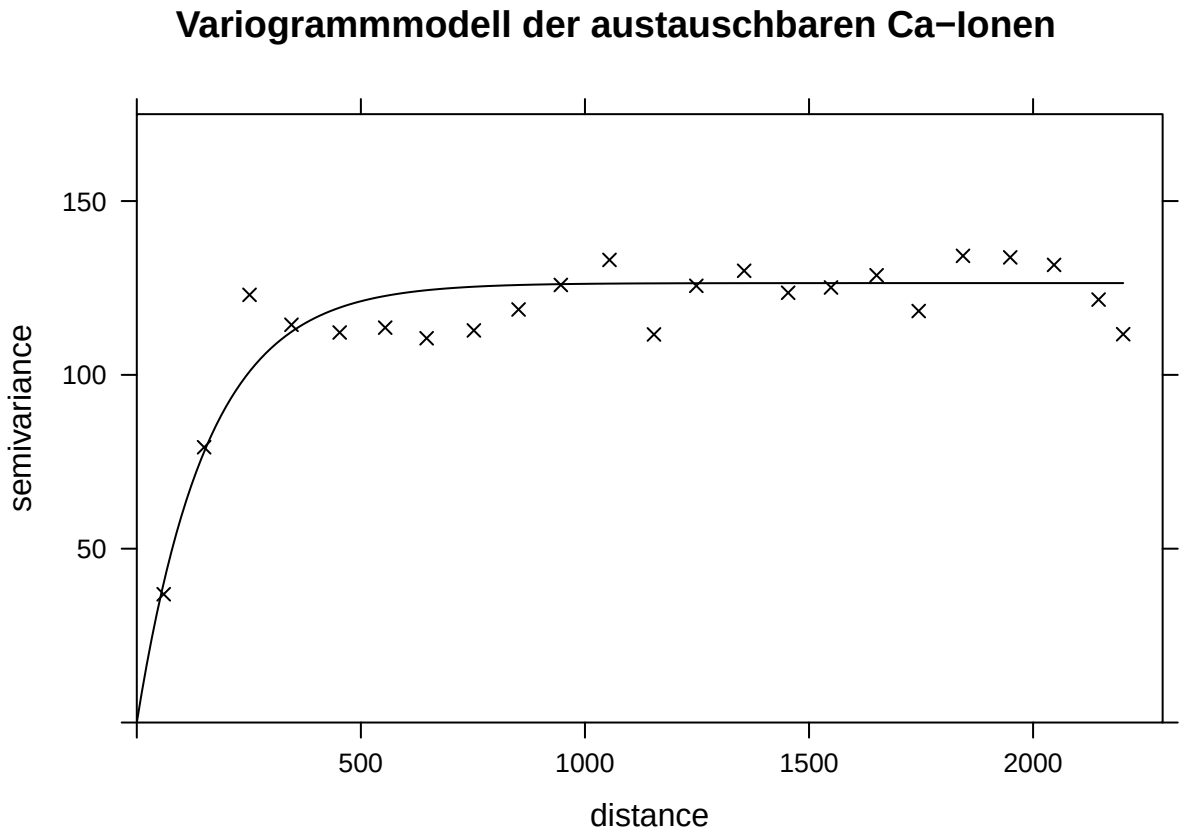
```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
load("data/yingtan_20_ueb3.RData")

##SpatialPointsDataFrame##
library(sp)
SPDFljz <- ljz
coordinates(SPDFljz) <- ~ EAST + NORTH
proj4string(SPDFljz) <- CRS("+proj=utm +zone=50 +ellps=WGS84 +datum=WGS84")

##Reproduktion des Variogrammmodells##
#omnidirektionales empirisches Variogramm
library(gstat)
Ca <- SPDFljz@data$Ca_exch

vario_omni_Ca <- variogram(Ca ~ EAST + NORTH,
                           data = SPDFljz,
                           cutoff = 2202,
                           width = 100)

plot(vario_omni_Ca,
     main = "Omnidirektionales empirisches Variogramm",
     pch = 4,
     col = "black")
```




```
#Modell zum Variogramm
vario_omni_Ca_fit <- fit.variogram(vario_omni_Ca,
                                   vgm(model = "Exp"),
                                   fit.method = 7)

plot(vario_omni_Ca,
     model = vario_omni_Ca_fit,
     cutoff = 2202,
     ylim = c(0, 175),
     pch = 4,
     col = "black",
     main = "Variogrammmodell der austauschbaren Ca-Ionen")
```



Aufgabe 14 Leave-One-Out-Cross-Validation

Die Validierung der Ergebnisse ist ein wichtiger Schritt jeder Modellierung. Um erfolgreich und unabhängig validieren zu können, bedarf es Daten, die nicht in die Kalibrierung des Modells eingeflossen sind. Um den häufig ohnehin schon kleinen Datenpool durch eine Aufteilung in Kalibrierungs- und Validierungsdatensatz nicht noch weiter zu reduzieren wird bei geostatistischen Modellen häufig das LOOCV-Verfahren angewendet. Dabei wird nacheinander ein Probenstandort aus dem Modell entfernt und die Zielgröße für diesen Ort vorhergesagt; so lange, bis alle Beprobungspunkte einmal ausgeschlossen worden sind.

a) Führen Sie mit der Methode krige.cv für die Ca-Ionen eine leave-one-out-cross-validation durch. Verwenden Sie das Variogrammmodell aus Aufg. 13 und notieren Sie ihre R-Syntax im Protokoll. (1 Punkt)

```
#Leave-one-out-Cross-Validation
LOOCV <- gstat::krige.cv(formula = Ca_exch ~ 1,
                  locations = SPDFljz,
                  model = vario_omni_Ca_fit)
```

b) Vergleichen Sie die Struktur des mittels krige.cv generierten Objekts mit dem Ergebnis der krige-Funktion aus Aufg. 13. Welche Daten-Attribute sind hinzugekommen und wofür stehen sie? (1 Punkt)

```
summary(LOOCV)
```

```
## Object of class SpatialPointsDataFrame
## Coordinates:
##           min     max
## EAST   490441  493591
## NORTH 3121290 3125630
```

```
## Is projected: NA
## proj4string : [NA]
## Number of points: 335
## Data attributes:
##    var1.pred         var1.var          observed         residual
##  Min.   : 6.958   Min.   :  1.701   Min.   : 3.772   Min.   :-23.95501
##  1st Qu.:17.146   1st Qu.: 27.358   1st Qu.:13.633   1st Qu.: -5.52411
##  Median :20.593   Median : 81.062   Median :19.491   Median : -1.20433
##  Mean   :21.861   Mean   : 63.792   Mean   :21.926   Mean   :  0.06484
##  3rd Qu.:26.082   3rd Qu.: 91.357   3rd Qu.:28.718   3rd Qu.:  5.11838
##  Max.   :44.167   Max.   :125.531   Max.   :94.311   Max.   : 68.02619
##      zscore               fold
##  Min.   :-3.095274   Min.   :  1.0
##  1st Qu.:-0.689932   1st Qu.: 84.5
##  Median :-0.145780   Median :168.0
##  Mean   : 0.006295   Mean   :168.0
##  3rd Qu.: 0.670534   3rd Qu.:251.5
##  Max.   : 7.136681   Max.   :335.0
```

observed: Tatsächlich gemessenen Werte.

residual: Residuen als Differenz des errechneten Werts zum tatsächlichen Wert, der zur Überprüfung ausgelassen wurde.

zscore: Kriging Standard-Fehler, bei dem die Kriging-Varianz eine Rolle spielt. Mean und Variance sollten dicht an 0 und 1 liegen.

fold: Zeigt zu welcher Teilmenge der jeweilige Datensatz gehört.

(Bivand, Pebesma, and Gomez-Rubio 2008)

c) Wie sähe das Vorhersageergebnis aus, wenn der Probenstandort während der Kreuzvalidierung nicht ausgeschlossen werden würde? Was ergäbe sich konsequenterweise bei der Fehlerberechnung? (1 Punkt)

Die Leave-One-Out-Cross-Validation beruht auf dem theoretischem Ansatz, den vorhandenen Datensatz nur gegen eine einzige Beobachtung zu testen und dieses Verfahren schließlich auf jeden Beobachtungsparameter anzuwenden. Ohne das Ausschließen dieses Wertes, würde das kriging den beobachteten Wert selbst vorhersagen. Das Ergebnis ist dann dasselbe wie beim Ordinary Kriging (s. R Hilfe).

Aufgabe 15 Root-Mean-Squared-Error

Der RMSE gibt Auskunft darüber, wie nah das Modell an die bekannten, tatsächlich gemessenen Daten herankommt. Er ist ein Maß für die 'accuracy' des gewählten Prädiktionsverfahrens.

a) Berechnen Sie den RMSE für die austauschbaren Ca-Ionen. (1 Punkt)

```
#Root-Mean-Square-Error Ca
rmse <- function(x,y) {
  sqrt(mean((x)^2))
}

rmse(x = LOOCV@data$residual)
```

```
## [1] 9.353448
```

b) Was bedeuten die einzelnen Silben des Wortes RMSE und warum wird der Vorhersage-Fehler gerade so beschrieben? (2 Punkte)

RMSE steht für Root-Mean-Square-Error. Zur Fehlerberechnung des Modells wird die Wurzel aus den quadrierten mittleren Residuen des berechneten Werts von den tatsächlichen Werten berechnet. Die Differenzen

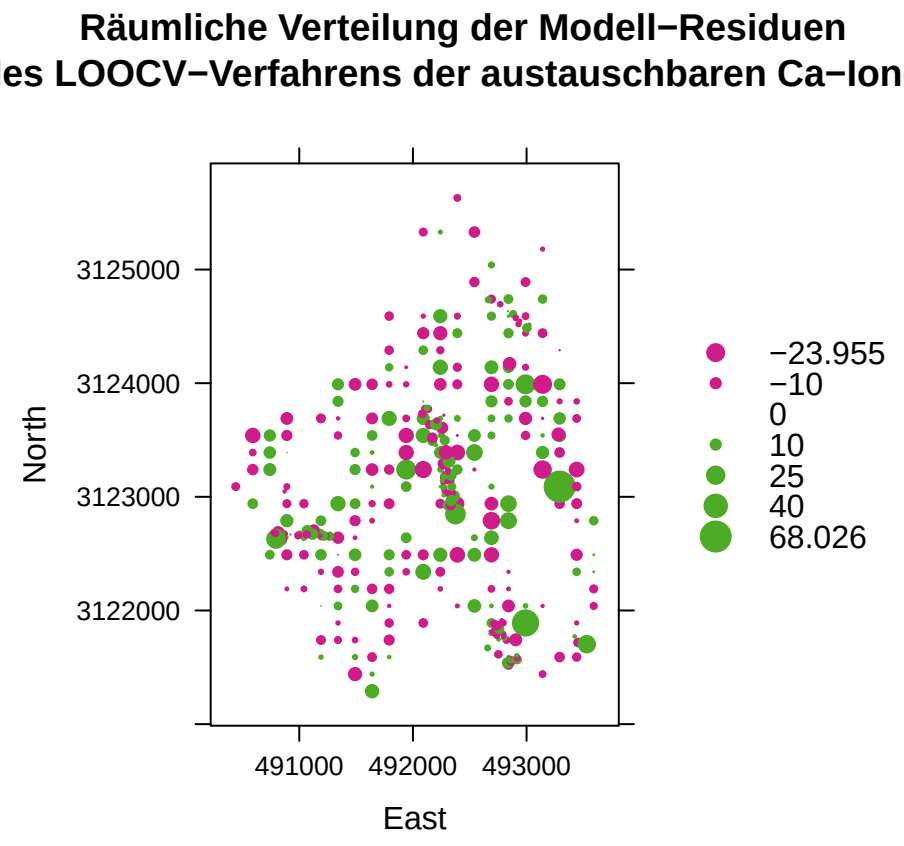
von Schätzwerten und tatsächlichen Messwerten werden zunächst quadriert, da der Betrag der Differenz (= Residuen) nur positiv sein kann. Um die Quadrierung wieder zu negieren, wird nach der Ermittlung des arithmetischen Mittels die Wurzel gezogen. Die Einheit des RMSEs ist dieselbe wie die der Eingangsgröße.

Aufgabe 16 Grafische Validierung

Die Angabe eines Gesamtfehlers reicht nicht aus, um die Güte eines Modells hinreichend zu beschreiben. Eine Darstellung der Verteilung der Fehler im Raum ist ebenso nützlich wie die Betrachtung der Streuung im Werteraum.

a) Erstellen Sie für ihre Modell-Residuen einen ansehnlichen Bubble-Plot und gehen sie der Frage nach, ob räumliche Muster erkennbar sind. (2 Punkte)

```
#Bubbleplot der LOOCV
library(lattice)
bubble(LOOCV, "residual",
       key.space="right",
       key.entries=c(min(LOOCV$residual),
                     -10,0,10,25,40,
                     max(LOOCV$residual)),
       scales=list(tick.number=4, alternating=1),
       maxsize=2,
       xlab="East",
       ylab="North",
       main="Räumliche Verteilung der Modell-Residuen\ndes LOOCV-Verfahrens der austauschbaren Ca-Ionen
```

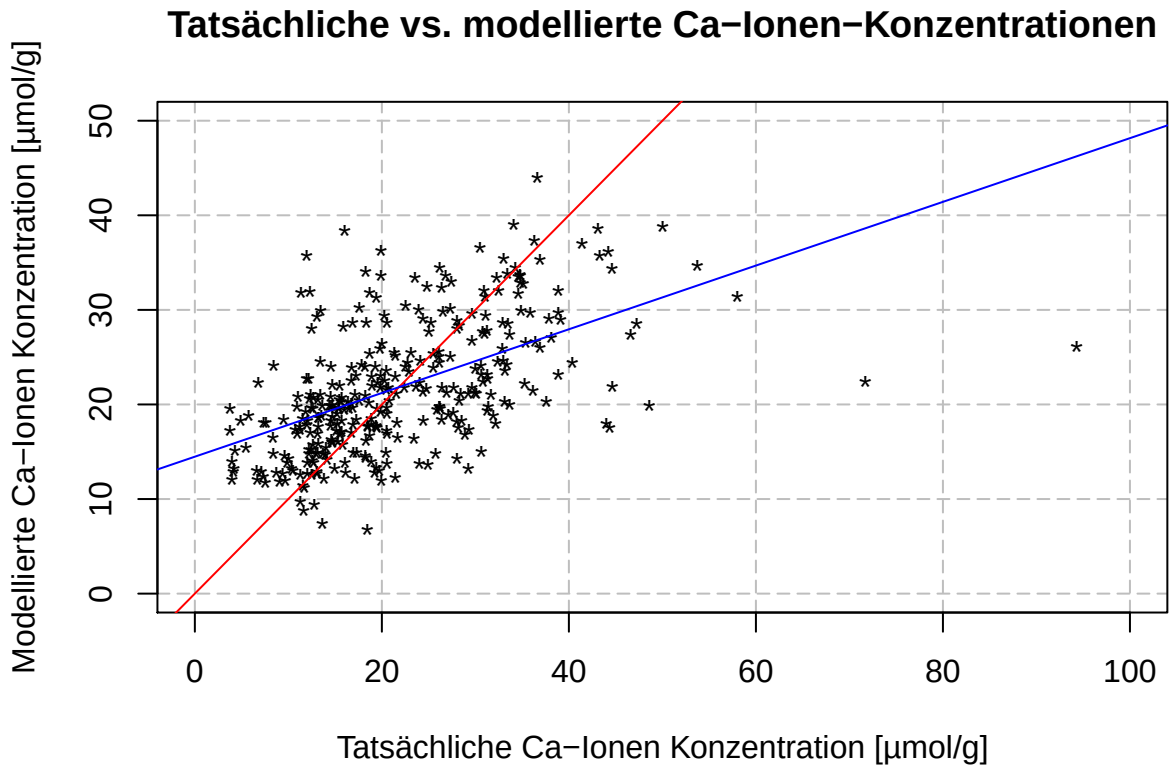


Es ist kein klarer räumlicher Trend für das erstellte Modell zu erkennen. Im westlichen Drittel des Beprobungsgebietes ist die Differenz zwischen den vorhergesagten und den tatsächlich gemessen Werten

etwas höher als im östlichen Teil. Insgesamt ist nicht zu erkennen, ob eine Über- oder Unterschätzung der austauschbaren Ca-Ionen vorliegt. An den Beprobungsstandorten, wo sehr hohe Messwerte gemessen wurden, sind auch die Residuen höher (Ausreißer). Dort, wo viel beprobt wurde, entsteht durch die Häufung der bubbles der Eindruck, dass das Modell stark von den gemessenen Werten abweicht. Das kann aber so nicht bestätigt werden.

b) Plotten Sie die tatsächlichen Ca-Ionen-Konzentrationen gegen die vorhergesagten Werte. Ergänzen Sie eine Ausgleichsgerade mit der Steigung eins und einem Verlauf durch den Ursprung. (1 Punkt)

```
plot(LOOCV$observed,
     LOOCV$var1.pred,
     xlim= c(0,100),
     ylim=c(0,50),
     pch="*",
     cex=1,
     col="black",
     xlab= "Tatsächliche Ca-Ionen Konzentration [µmol/g]",
     ylab= "Modellierte Ca-Ionen Konzentration [µmol/g]",
     main ="Tatsächliche vs. modellierte Ca-Ionen-Konzentrationen",
grid(col= "grey", lty=5));
abline(0,1,col="red")
abline(lm(LOOCV$var1.pred ~ LOOCV$observed),
       col = "blue")
```




```
#Gleichung Regressionsgerade
x <- LOOCV$observed
y <- LOOCV@data$var1.pred
```

```r
reg <- lm(y ~ x, data = LOOCV)

s <- summary.lm(reg)
b <- s$coefficients[1,1]
a <- s$coefficients[2,1]
cat(a,"x +", b, sep=" ", append=TRUE)
```

```
## 0.336812 x + 14.47628
```

c) Bewerten Sie kurz das durchgeführte Interpolationsverfahren. Beziehen Sie sich auf den RMSE und ihre Diagnose-Plots. (2 Punkte)

Da bei dem RMSE von 9,4μmol/g im Vergleich zu einer Spannweite der Werte von knapp über 90μmol/g (etwa 1/10) die Ausreißer eine großen Einfluss auf das Ergebnis haben, ist der Wert des Fehlers eventuell höher als die tatsächliche Abweichung vom Modell (Abweichung von der roten Geraden). Zur genauen Fehlerbestimmung des Modells sollten noch andere Größen, wie z.B. der MAE, in Betracht gezogen werden.

Außerdem zeigt sich auch hier mit einem Korrelationskoeffizient von 0,34, dass kaum ein Zusammenhang zwischen den Werten der austauschbaren Ca-Ionen und ihrer geographischen Lage besteht. Daher sind für ein geeignetes Modell weitere Größen wie Beprobungstiefe oder Reliefgrößen notwendig.

Literatur

Bivand, R. S., E. J. Pebesma, and Gomez-Rubio. 2008. *Applied Spatial Data Analysis with R*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-78171-6.