Authors: @OTI2020, @Timo12345689

Spatial Cross Validation (räumliche Kreuzvalidierung)

Zusammenfassung

Spatial Cross Validation ist eine Kreuzvalidierungsmethode um mit räumlicher Partition die Effekte von Toblers First Law of Geography zu verringern. Meist wird der K Means Algorithmus genutzt um die Daten zu unterteilen. Der Effekt, der umgangen werden soll, wird auch als "Sneak Preview" bezeichnet. Meistens wird eine Verschachtelung von Cross Validation mit dem K Means Algorithmus genutzt.

1. Was ist räumliche Kreuzvalidierung (Spatial Cross Validation)?

- Idee: Datensatz wird wiederholt in einen Trainings- und einen Testsatz aufgeteilt
- Trainingsdaten werden zur Anpassung an ein Modell verwendet, welches dann auf den Testsatz angewendet wird
- Vergleich der vorhergesagten Werte mit den bekannten Antwortwerten (aus dem Testdatensatz) -> Bewertung möglich, ob Modell passt (Ziel ist es, die Fähigkeit des Modells Werte (aus unabhängigen Daten) vorherzusagen, zu erfassen)
- Unterschied zu herkömmlicher Kreuzvalidierung: räumliche Partitionierung als Möglichkeit Toblers First Law of Geography zu umgehen

2. Warum benutzen wir räumliche Kreuzvalidierung?

- Toblers First Law of Geography besagt, dass Punkte, die nahe beieinander liegen, im Allgemeinen ähnlicher sind als Punkte, die weiter entfernt sind

  > Punkte sind statistisch gesehen nicht unabhängig, da Trainings- und Testpunkte in konventioneller Kreuzvalidierung (Cross Validation) oft zu nahe beieinander liegen

- Trainingsbeobachtungen, die sich in der Nähe der Testbeobachtungen befinden können eine Art "Sneak Preview" entstehen lassen

  > Sneak Preview: Trainingsdatensatz erhält Informationen, die ihm eigentlich nicht zur Verfügung stehen sollten

- Umgehung dieses Problems durch "räumliche Partitionierung" -> Beobachtungen werden in räumlich unzusammenhängende Teilmengen aufgeteilt

- "räumliche Partition" ist (praktisch) einziger Unterschied von räumlicher Kreuzvalidierung zu herkömmlicher Kreuzvalidierung

- räumliche Kreuzvalidierung führt zu einer verzerrungsreduzierten Bewertung der Vorhersageleistung eines Modells -> Vermeidung von Overfitting (Überanpassung)

  > Beispiel:
  >
  > **Macht es Sinn, für die Validierung eines Modells Pixel/Orte anzuschauen, die direkt benachbart zu denen sind, auf denen das Modell traininiert wurde? (vor allem, wenn der**

> **zu analysierende Prozess starke räumliche Autokorrelation aufweist)?**
>
> Nein, da durch Toblers First Law of Geography die Gefahr der "Sneak Preview" entsteht. Der Trainingsdatensatz enthält Informationen, die er eigentlich nicht erhalten sollte und erschafft dadurch ein verfälschtes Ergebnis, was die Validierung des Modells erschwert. Hier würde räumliche Partitionierung die Validierung des Modells vereinfachen, da die starke räumliche Autokorrelation des zu analysierenden Prozesses entsprechend umgangen wird.

3. Wann kann diese Methode benutzt werden?

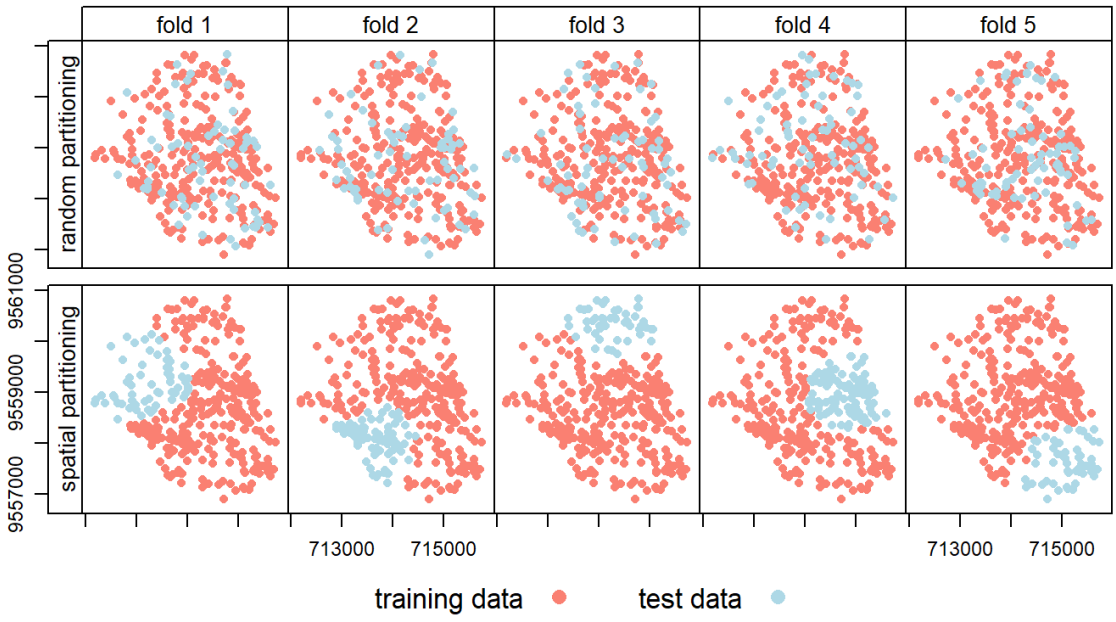
- Wenn die gegebenen Daten eine hohe Autokorrelation haben, um Overfitting/Überanpassung dieser zu verhindern
- hohe Autokorrelation = Korrelation (Beziehung zwischen zwei oder mehreren Merkmalen) eines Punktes mit sich selbst zu einem früheren Zeitpunkt
- Überanpassung/Overfitting: Möglichkeit dass das Modell die Trainingsdaten zu gut modelliert.

4. Wie wird diese Validierungmethode angewendet?

- Verschachtelung von herkömmlicher Kreuzvalidierung
- Beispiel: 100x 5-fache Kreuzvalidierung mit einer räumlichen Partition durch k-means Clustering mit k = 5
- k-means Clustering = Aus einer Menge von ähnlichen Elementen wird eine vorher bekannte Anzahl von k Gruppen gebildet
- Am häufigsten verwendete Technik zur Gruppierung, da schnelles Erkennen von Clusterzentren, Algorithmus bevorzugt Gruppen mit geringer Varianz und ähnlicher Größe

---

Spatial Partioning im Vergleich zu Random Partitioning



- Sichtbar ist, dass bei Random Partitoning die Testdaten keinen räumlichen Zusammenhang haben

- Ändert sich auch bei weiteren Iterationen nicht
- Spatial Partitioning sorgt für einen räumlichen Zusammenhang der Testdaten und sorgt außerdem dafür, dass alle Daten je zu Test- und Trainingsdaten im Laufe der Gesamtiteration werden.

Quellen:

- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6352393
- https://geocompr.robinlovelace.net/spatial-cv.html#intro-cv