

# Spatial Cross Validation (räumliche Kreuzvalidierung)

---

## 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)

## 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)

## 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:

Beispiel: Erhöhter Bedarf von E Ladesäulen an Autobahnkreuzen (?)(durch Kreuzung zweier Autobahnen wird ein erhöhter Bedarf festgestellt und berechnet)

## 4. Wie wird diese Validierungsmethode 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
- 

## Spation Partioning im Vergleich zu Random Partitioning



---

### Quellen:

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