Aufgabenstellung

\*Ziel: \*Interaktives Tool, in dem verschiedene Szenarien an räumlichen Vorhersagemodellen getestet und verglichen werden können. Es soll somit zum verbesserten Verständnis von Herausforderungen räumlicher Vorhersagemodellen und deren Güteabschätzung beitragen.  
  
\*Daten: \*Das Ganze läuft auf simulierten virtuellen Daten: Prädiktoren werden simuliert, anhand von nutzerdefinierten Kriterien (z.B Anzahl Prädiktoren, räumliche Autokorrelation etc.).  
Auch die Response muss simuliert werden damit wir sie flächendeckend (zum Vergleich) zur Verfügung haben. Simulieren könnte man die Response aus einem Subset der Prädiktoren mit Hilfe einer Funktion (Komplexität und Rauschen ist nutzerdefiniert), d.h. wir stellen sicher, dass wir aus den Prädiktoren auf die Response schließen können und Aufgabe des Algorithmus (z.B. RF) ist es das dann später auch zu tun.  
  
Damit wir den RF trainieren können simulieren wir Trainingsdaten (Parameter: N, Verteilung (geclustert, zufällig, regulär),...).  
  
\*Modellierung: \*Mit den Prädiktoren und den Trainingsdaten wird ein Algorithmus (von mir aus erstmal nur Random Forest) trainiert. Tuning & Variablenselektion findet mit & ohne Variablenselektion statt und verschiedene Kreuzvalidierungsmethoden (zufällig und räumlich wobei es für die räumliche optional mehrere Varianten geben könnte) werden nicht nur zum Tuning sondern auch zur Validierung verwendet. Im Anschluss wird die AOA berechnet.  
  
\*Visueller/Quantitativer Output: \*Zufälliger CV Fehler, Räumlicher CV Fehler, realer Fehler, das Ganze mit & ohne Variablenselektion, Karte mit Differenz Real-Vorhersage, Karte mit DI/AOA  
  
\*Wichtig: \*Das Tool soll weiterentwickelt werden können. Deshalb unbedingt von Anfang an darauf achten, dass alles ordentlich dokumentiert wird!\*  
\*  
  
\*Einstiegsliteratur: \*Eine Schwierigkeit ist sicherlich die sinnvolle Simulation der virtuellen Prädiktoren und der Response. Carles hat in seiner Masterarbeit (Anhang) eine virtuelle Landschaft mit Prädiktoren und Response simuliert. Der Code ist auch auf Github. Vielleicht könnte man das sogar genauso übernehmen. Auch im Virtualspecies Paket (<https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.01388>) wird sowas gemacht (wobei hier die Prädiktroren real sind). Im Anhang außerdem ein Artikel von mir in dem mal verschiedene Szenarien an Kreuzvalidierung und Variablenselektion getestet wurden.

Fragen:

* Mein Eindruck und was mir eventuell fehlt:
  + Viel Reproduzieren und korrekt anwenden aber nicht selbst forschen oder eine eigene Fragestellung beantworten
    - Was könnte man für eine eigene Fragestellung noch ergänzen?
    - Ist die Beantwortung einer Frage nicht notwendig? Was wäre dann die Fragestellung?
    - Ich würde gerne Sachen herausfinden und eventuell neue Zusammenhänge erklären.
  + Soll ich die BA auf Englisch verfassen? Auf jeden Fall.
  + Fokus auf interaktives Tool und später nutzen, um eventuell Hinweise zu generieren
* Simulation von Landschaften:
  + Anzahl Prädiktoren (umsetzbar)
    - Was passiert mit den 8 Variablen nicht genutzt werden`? (S.16)
  + Räumliche Autokorrelation
    - Womit? Mit Gauss?
    - “Covariates are 2-dimensional stationary and isotropic Gaussian random fields subject to spatial autocorrelation, expressed with spherical semivariograms”
    - Was genau heißt das? Was hat Carles angenommen? Wieso eignet sich das besonders gut? Wieso nicht eins der anderen?
    - In welcher Range? 5, 10, 20, und 40% Autokorrelation?
    - Wie modelliere ich Komplexität und Rauschen? Random und Spatial Noise?
  + Wie groß soll die Landschaft sein?
  + Was noch?
* Trainingsdaten:
  + N:
    - 50, 100, 150 oder noch mehr?
    - Er hat nur „Punkte“ keine Polygone benutzt ABER geclusterte Punkte repräsentativ für Polygone, oder?
  + Verteilung:
    - (random, regular, clust1, clust2 und nonunif)?
* Model:
  + Welche Modelle? Unterscheidung von räumlichen und Standard-Modellen oder generell andere Algorithmen?
* Validierung:
  + Was ist die normale Surface Validation?
  + Was ist der genaue Unterschied zwischen sbLOOrange und sbLOOndm CV?
  + Welche CV Methode haben wir bereits im Kurs genutzt?
  + Hättest du sie gerne im Tool?
  + Welche Selektionsalgorithmen möchtest du einbetten?
    - RFE, FFS jeweils mit räumlicher und zufälliger Kreuzvalidierung?
* Was ist der reale Fehler?

**Aufbau der App**

* Auswahl der Prädiktoren:
  + Parameter: Anzahl, Landschaftsmodell (NLM), räumliche Autokorrelation?
* Simulieren der Response
  + Simulieren könnte man die Response aus einem Subset der Prädiktoren mit Hilfe einer Funktion (Komplexität und Rauschen ist nutzerdefiniert), d.h. wir stellen sicher, dass wir aus den Prädiktoren auf die Response schließen können und Aufgabe des Algorithmus (z.B. RF) ist es das dann später auch zu tun
* Simulieren der Trainingsdaten
  + Damit wir den RF trainieren können simulieren wir Trainingsdaten
  + Parameter: N, Verteilung (random, regular, weak clustering, strong clustering, uniform)
* Modellierung
  + Mit den Prädiktoren und den Trainingsdaten wird ein Algorithmus (Random Forest) trainiert
  + Tuning & Variablenselektion finden mit & ohne Variablenselektion statt und verschiedene Kreuzvalidierungsmethoden (zufällig und räumlich, wobei es für die räumliche optional mehrere Varianten geben könnte) werden nicht nur zum Tuning, sondern auch zur Validierung verwendet
  + Berechnung der AOA
* Visueller/Quantitativer Output
  + \*Zufälliger CV Fehler, Räumlicher CV Fehler, realer Fehler, das Ganze mit & ohne Variablenselektion, Karte mit Differenz Real-Vorhersage, Karte mit DI/AOA

Wichtig: Das Tool soll weiterentwickelt werden können. Deshalb unbedingt von Anfang an darauf achten, dass alles ordentlich dokumentiert wird!

Größte Unklarheiten bei Programmierung:

* Wie lasse ich die NLM auswählen?
* Wie modelliere ich konkret die Response und wie füge ich Rauschen bzw. Noise hinzu?
* Implementieren verschiedener Auswahlmöglichkeiten für Anzahl und Verteilung
* Räumliche CV/ Zufällige CV
* Wie mach ich räumliche Kreuzvalidierung, wenn ich keine Polygone habe?

**2. Fragestunde (09.03.22)**

Zeigen, was ich bisher gemacht habe:

* Erste Erfahrungen mit RShiny gesammelt
* Prototyp der App entwickelt
  + Erstes Layout der App
  + Bereits einige Funktionalitäten implementiert

NLMs:

* Anzahl lässt sich über ausgewählte NLMs bestimmen
* Nur 10 verschiedene NLMs in App integriert aufgrund von Problemen (lässt sich beheben)
* Viele Parameter können bei der Erzeugung der einzelnen NLMs übergeben werden aber aktuell sind sie nicht einstellbar (siehe Funktion generate\_predictors()) (Experten-Modus)
* Soll das eventuell geändert werden?
* Sollen vom gleichen NLM-Typen mehrere NLMs generiert werden können? Hätte das einen Nutzen? (Experten-Modus)
* Keine Legenden
* Maß für Korrelation der simulierten Ausgabe?
* Virtual species R-Package
* Nicht alle Prädiktoren mit in die Berechnung der Ausgabe einbringen
* 🡪 Auswahl ermöglichen
* Anzahl Variablen und dann random generieren oder specify

Sample Points:

* Sind die Variablen so in Ordnung?
* Sinn für N-Cluster? Cluster Dichte?
* Vielleicht den Output Plot etwas verändern und nur sichtbar machen, sobald man an den Parametern herumspielt

Simulated outcome:

* Funktion, die aus allen Prädiktoren einen Ausdruck generiert, sodass die verschiedenen Prädiktoren-Layer anschließend miteinander verrechnet werden können
* Wird jetzt zufällig generiert, aber soll es immer die gleiche sein? Welche Operanden soll ich mit einbringen?
* Noch kein Rauschen/Noise eingefügt
* Kann erst berechnet werden, wenn vorher Prädiktoren generiert wurden
* Skalieren des Outcomes
* Noise auswählbar (random und spatially correlated noise)
  + Wie viel Noise? Einfach addieren ist nicht so gut, weil es zu viel Noise wird (Kurzfristige Lösung: Mit 0.2 multipliziert)
* Knopf für Set seed -> ABER immer der gleiche seed (Eine bestimmte Auswahl lässt sich bisher nicht speichern)
* SEED einstellbar über Input Field
* Save seed überall

Prediction:

* Trainingsdaten werden auf Basis der sampling points extrahiert
* Random Forest Modell
* Noch ohne Kreuzvalidierung
  + Was für eine räumliche Kreuzvalidierungsmethode soll ich implementieren?
  + Nearest Distance Matching
  + Global\_validation/ Sb Blockwise Caret?
  + BlockCV
  + Blockwise? Oder jeden Punkt einmal auslassen?
  + Nochmal erläutern wie blockwise funktioniert
* Literatur für Kreuzvalidierungsmethoden
* Was geben die Fehlerwerte nochmal an? RMSE, R2 und MAE?
* MAE selbst berechnet, stimmt das so?
* Mit oder ohne Variablenselektion?

Difference:

* Wird berechnet
* Leider nicht dieselbe Legende wie outcome und prediction

Weitere Ausgaben:

* AOA, DI und Tabelle der verwendeten Trainingsdaten

**Weitere Fragen:**

* Worauf wollen wir hinarbeiten?
* Welche Zusammenhänge sollen hier dem Nutzer klar werden?
* Wie kann ich diese besser unterstreichen?
* Was soll alles als Ergebnis angezeigt werden?
* Wie soll ich die Ergebnisse anordnen?
  + Verschiedene Seiten? Aber dann kann man schöne Zusammenhänge nicht so gut sehen
* Konkrete Aufgabenstellung?

TO-DO:

CAST::plot\_geodist

Fragen:

* Wie stelle ich spatial blocks her für die spatial cross validation?

Notizen:

* Welche Prediction benutzen wir?

Was muss noch erledigt werden?

* Implementierung der NNDM LOO CV Methode
* Variablenselektion RFE einbetten
  + Was soll mit der Variablenselektion gezeigt werden?
  + Welche Prädiktoren ausgewählt bzw. rausgeschmissen wurden?
* Visualisierung:
  + Gleiche Legende für Outcome und Prediction -> große Schwierigkeiten: Ideen dazu?
  + Welche Farben verwenden?
  + Prediktoren im selben Stil
  + Zeitlicher Delay bei Berechnung (Panels erst ein bisschen später anzeigen lassen)
  + Plots nicht dynamisch verändern X und Y sollten immer gleiche Ausdehnung haben
* Code aufräumen/kommentiern!
* Code vereinfachen
* Eingabe einer eigenen Formel
* Plot of distances

Wo brauche ich Hilfe?

* RFE: Da läuft irgendwas noch nicht richtig
* Variogramm – bei mir stellt sich keine wirkliche Range ein, was tun?
  + Daten zu stark korreliert -> stetig steigendes Variogram
* Inwiefern kann man da die räumliche Autokorrelation wirklich bestimmen? Und inwiefern könnte uns das etwas bringen?

Aufgetretene Fehler:

* Wenn als CV Methode nur NNDM LOO ausgewählt wir, öffnet sich auch LOO CV (die includes()-Abfrage ist nämlich wahr. Problem: Wenn ich NNDM LOO CV + eine andere Methode wähle, tritt dieser Fehler nicht auf, warum???
* Legende korrigieren (auch wenn mehrere Predictions gezeigt werden)

Folgendes muss noch implementiert werden:

* Eine Eingabebox zur eigenen Aufstellung einer Formel für die Berechnung der Simulation
* Plot der „distances“
* Variablenwichtigkeit plotten
* Variablenselektion RFE
* Bestimmung der Autokorrelation, aber wie?
* Vorhersage und AOA für alle Modelle berechnen und plotten (AOA optional)
* Mtry konstant auf 2 setzen –> nicht Bestandteil unserer Simulation (für RF schon erledigt, aber auch für SVM implementieren)
* Shiny App veröffentlichen (shinyserver.io?)
* Legende berichtigen (Zlim: bei Achsenlegende setzen) 🡪 Lesezeichen gesetzt

**Hypotheses**

* It is necessary to account for spatial autocorrelation during the validation of map accuracy
* Depending on the distribution of the sampling points (i.e., training data), some cross-validation strategies are more suitable than others
* Random CV is suitable for random und regular sampling points
* Random CV is overoptimistic for clustered sampling points
* Spatial CV is more suitable for clustered sampling points than random cv
* NNDM CV fits best for all types of sampling points
* For clustered samples:
  + AOA becomes smaller for random cv
* Variable selection:
  + Example lat/lon: Explain why models tend to see it as important (because it is easy to distinguish things about it). This means that the model is good at discriminating data based on these predictors, but it has nothing to do with the actual context

-> When using RFE, these never fall out, no matter whether with random CV or spatial CV

-> For FFS, large differences should occur in the selected predictors