Analýza prostorově závislých funkcionálních dat

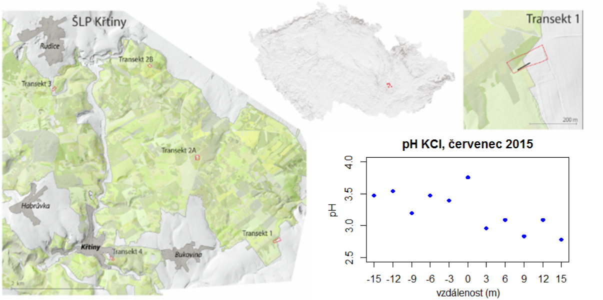
V. Římalová, A. Menafoglio, A. Pini, E. Fišerová

Robust 2018

25. ledna 2018

Motivace – Data a náhled lokace

- Měsíční měření (březen-říjen 2015 a 2016) 5 chemických ukazatelů
- 5 lokací (transektů) v okolí obce Křtiny na Brněnsku
- V transektu se nachází 11 odběrných míst umístěných v přímce za sebou, vzdálenost sousedících OM je 3 metry.
- Uprostřed ležící odběrné místo, ekoton, rozděluje lokaci na polní a lesní část.



Cíl

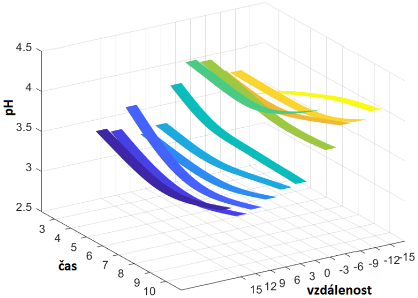
- Ověřit, zda se naměřené hodnoty chemických ukazatelů liší v závislosti na typu půdy a vzdálenosti od ekotonu.
- V příspěvku prezentujeme výsledky pro pH chlornanu draselného naměřeného v transektu 1 v hloubce 5 cm.

Funkcionální geostatistika

- Nechť $(\Omega, \mathcal{F}, \mathbb{P})$ je pravděpodobnostní prostor a $\mathcal{H}$ je Hilbertův prostor, na němž je definovaný skalární součin $\langle . \, , . \rangle$ indukující normu $||.||$.
- Funkcionální náhodnou proměnnou nazveme měřitelnou funkci $\mathcal{X} : \Omega \to H$
- Množina $\{\mathcal{X}_s, s \in D \subset \mathbb{R}^d\}$ je funkcionální náhodné pole.
- Funkcionální dataset $\mathcal{X}_{s_1}, \dots, \mathcal{X}_{s_n}$ je množinou n pozorování náhodného pole vzhledem k lokacím $s_1, \dots, s_n \in D$

Funkcionální geostatistika

- Nechť $(\Omega, \mathcal{F}, \mathbb{P})$ je pravděpodobnostní prostor a $\mathcal{H}$ je Hilbertův prostor, na němž je definovaný skalární součin $\langle . \, , . \rangle$ indukující normu $||.||$.
- Funkcionální náhodnou proměnnou nazveme měřitelnou funkci $\mathcal{X} : \Omega \to H$
- Množina $\{\mathcal{X}_s, s \in D \subset \mathbb{R}^d\}$ je funkcionální náhodné pole.
- Funkcionální dataset $\mathcal{X}_{s_1}, \dots, \mathcal{X}_{s_n}$ je množinou n pozorování náhodného pole vzhledem k lokacím $s_1, \dots, s_n \in D$



Funkcionální geostatistika

- Množina $\{\mathcal{X}_s, s \in D \subset \mathbb{R}^d\}$ je nestacionární náhodné pole, jehož prvky lze vyjádřit jako $\mathcal{X}_s = m_s + \delta_s$.
- Drift lze modelovat jako $m_s(t) = \Sigma_{l=0}^{L} \beta_l(t) f_l(s), s \in D, t \in T$, $\beta_l(t)$ jsou funkcionální regresní parametry a $f_l(s)$ regresory.
- Nechť $\delta_{s_1}, \dots, \delta_{s_n}$ jsou realizace reziduálního procesu $\{\delta_s, s \in D\}$, který má nulový průměr, stacionaritu druhého řádu a je isotropní [Menafoglio, Secchi 2016]. Empirický semivariogram tohoto reziduálního procesu má tvar

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \Sigma_{(i,j) \in N(h)} ||\delta_{s_i} - \delta_{s_j}||^2,$$

  kde $N(h) = \{(i,j) : ||s_i - s_j|| = h\}$ a $|N(h)|$ je mohutnost této množiny.
- Empirický variogram je definován jako $2\hat{\gamma}(h)$

Nechť $\mathcal{X}_{s_i}^{(1)}$ a $\mathcal{X}_{s_i}^{(2)}$ jsou dva náhodné výběry funkcí, $i = 1, \ldots, n_g, g = 1, 2$ je populační index. Testujeme hypotézu

$$H_0 : \mathcal{X}_s^{(1)} \sim \mathcal{X}_s^{(2)}$$

proti alternativě $H_1 : \mathcal{X}_s^{(1)} \nsim \mathcal{X}_s^{(2)}$.

- Testujeme pomocí intervalových permutačních testů pro funkcionální data.
- Výhody: Umožní identifikovat, v které části definičního oboru dochází k porušení H_0.
- Neklade žádné předpoklady na rozdělení funkcionálních dat.

Permutační testy – testování významnosti regresních parametrů

- Uvažujme lineární model

$$\mathcal{X}_s = \beta_0(t) + \Sigma_{l=1}^{L}\beta_l(t)f_l(s) + \delta_s, t \in \langle a, b \rangle, s \in D$$

- Chceme testovat hypotézu

$$H_0 : \beta_1(t) = \ldots = \beta_L(t) = 0,\ \forall t \in \langle a, b \rangle$$

proti alternativě $H_1 : \beta_l(t) \neq 0$ pro nějaké $l \in 1, \ldots, L$.

- Předpoklad: zaměnitelnost pozorování $\mathcal{X}_s$ [Pini & Vantini, 2017].

Model pro data se stejným rozptylem

Za předpokladu $\mathcal{X}_s^{(1)} \sim \mathcal{X}_s^{(2)}$, uvažujme model

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t)d + \delta_s(t).$$

Prostorové vztahy jsou dány indikátorovou funkcí

$$d = \begin{cases} 0 & \text{pro } s \in \{-15, -12, -9, -6, -3\}, \\ 1 & \text{pro } s \in \{3, 6, 9, 12, 15\}. \end{cases}$$

Ověření shody rozptylů pomocí permutačních testů

$H_0 : \mathcal{X}^{(1)} \sim \mathcal{X}^{(2)}$ testujeme na reziduích $\delta_s^{(g)}, g = 1, 2$, z modelu

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t)d + \delta_s(t)$$

pomocí statistiky

$$T^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{|\mathcal{I}|} [\widehat{\mathrm{var}}[\delta^{(1)}(t)] - \widehat{\mathrm{var}}[\delta^{(2)}(t)]]^2 dt,$$

kde $\mathcal{I}$ je libovolný podinterval $\langle a, b \rangle$, $t \in \langle a, b \rangle$ a $\widehat{\mathrm{var}}[\delta^{(g)}(t)], g = 1, 2$, značí rozptyl reziduí z 1. nebo 2. populace.

Ověření shody rozptylů pomocí permutačních testů

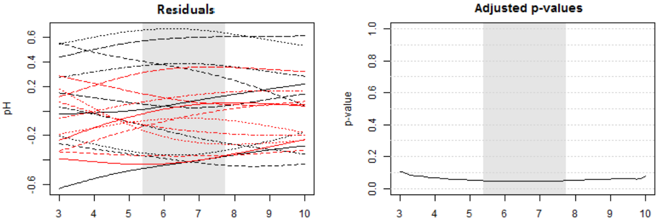
$H_0 : \mathcal{X}^{(1)} \sim \mathcal{X}^{(2)}$ testujeme na reziduích $\delta_s^{(g)}, g = 1, 2$, z modelu

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t)d + \delta_s(t)$$

pomocí statistiky

$$T^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{|\mathcal{I}|} [\hat{\mathrm{var}}[\delta^{(1)}(t)] - \hat{\mathrm{var}}[\delta^{(2)}(t)]]^2 dt,$$

kde $\mathcal{I}$ je libovolný podinterval $\langle a, b \rangle$, $t \in \langle a, b \rangle$ a $\hat{\mathrm{var}}[\delta^{(g)}(t)], g = 1, 2$, značí rozptyl reziduí z 1. nebo 2. populace.



Model pro data s různým rozptylem

Není-li splněn předpoklad $\mathcal{X}_s^{(1)} \sim \mathcal{X}_s^{(2)}$ a funkcionální pozorování mají různý rozptyl v závislosti na typu půdy, lze použít model

$$\mathcal{X}_s^{(g)}(t) = \beta_0(t) + \beta_1(t)d + \delta_s^{(g)}(t), g = 1, 2.$$

$\delta_s^{(g)}(t) = \sigma^{(g)}\delta_s(t)$

Prostorové vztahy modelujeme indikátorovou funkcí

$$d = \begin{cases} 0 & \text{pro } s \in \{-15, -12, -9, -6, -3\}, \\ 1 & \text{pro } s \in \{3, 6, 9, 12, 15\} \end{cases}$$

Rozptyly $\sigma^{(g)}$ odhadneme z parciálních modelů

$$\mathcal{X}_s^{(g)}(t) = \beta_0^{(g)}(t) + \delta_s^{(g)}(t), g = 1, 2.$$

V modelu $\mathcal{X}_s^{(g)}(t) = \beta_0(t) + \beta_1(t)d + \delta_s^{(g)}(t), g = 1, 2$, testujeme hypotézu $H_0 : \beta_1(t) = 0 \, \forall t \in \langle a, b \rangle$ proti alternativě $H_1 : \beta_1 \neq 0$ pomocí testové statistiky

$$T_0 = \int \left[[C\hat{\beta}(t)]'(CZ'WZC')^{-1}[C\hat{\beta}(t)] \right] dt,$$

kde W je matice vah, Z designová matice a vektor $C = (0, 1)$.

Permutační schéma:

1. Redukovaný model $\hat{\mathcal{X}}_s^{(g)}(t) = \hat{\beta}_0(t)$
2. Vydělit $\hat{\delta}_s^{(g)}(t)$ příslušným rozptylem a získaná $\hat{\delta}_s(t)$ permutovat $\rightarrow \mathcal{X}_s^*(t) = \hat{\beta}_0(t) + \hat{\sigma}^{(g)}\hat{\delta}_s^*(t)$.
3. Z $\mathcal{X}_s^*(t)$ odhadnout parametry původního modelu a vypočítat testovou statistiku.
4. p-hodnota permutačního testu je vypočtena jako poměr permutací vedoucích k vyšší hodnotě testové statistiky, než byla ta z původních dat, k celkovému počtu permutací.

Závěr

- Návrh modelů pro nekorelovaná funkcionální data se stejnými a různými rozptyly.
- Test pro 2 populace na reziduích prostorového lineárního modelu s funkcionální závisle proměnnou.
- Testování významnosti parametrů funkcionální regrese s využitím permutačních testů.

A co dál?

- Intervalový test významnosti regresních parametrů.
- Model pro v prostoru korelovaná data.

Literatura

- J.O. Ramsay, B.W. Silverman (2005): Functional Data Analysis. Springer, New York.
- A. Menafoglio, P. Secchi (2016): Statistical analysis of complex and spatially dependent data: a review of Object Oriented Spatial Statistics, European Journal of Operational Research, 258(2), pages 401–410.
- A. Pini & S. Vantini (2017): Interval-wise testing for functional data, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2017.1306627
- Abramowicz, K.; Häger, C.; Pini, A.; Schelin, L.; Sjöstedt de Luna, S.; Vantini, S.: Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament, MOX technical report 30/2016, Politecnico di Milano