An inferential framework for the analysis of spatio-temporal geochemical data

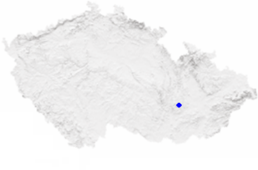
V. Římalová, A. Menafoglio, A. Pini, E. Fišerová

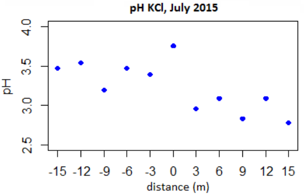
Palacký University in Olomouc, Czech Republic

Workshop on Functional Data Analysis
Prague, July 12, 2018

Motivation – Data description

- Monthly measurements (March-October 2015) of the potassium chloride pH
- Site located near Brno, Czech Republic
- Mean altitude $526,8\,m$, mean slope $2,7^\circ$, surface oriented to southwest
- The transect contains 11 sampling points (on a straight line), 3 meters from each other
- Central sampling point, ecotone, divides the site into field and forest part





Functional geostatistics

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let H be a separable Hilbert space (e.g. L^2 space) endowed with inner product $\langle , \rangle$ and induced norm $||.|| = \sqrt{\langle , \rangle}$ defined on H.
- We call functional random variable a measurable function $\mathcal{X} : \Omega \to H$, its realisation x is a functional datum.
- Let $\{\mathcal{X}_s, s \in D \subset \mathbb{R}^d\}$ be a functional random field.
- Functional dataset $\mathcal{X}_{s_1}, \ldots, \mathcal{X}_{s_n}$ is a collection of n observations of the random field related to locations $s_1, \ldots, s_n \in D$

Modelling spatial observations - drift

- Functional observations $\mathcal{X}_s$ of non-stationary random field $\{\mathcal{X}_s, s \in D \subset \mathbb{R}^d\}$ can be expressed as

$$\mathcal{X}_s = m_s + \delta_s.$$

- Drift m_s can be expressed through a linear model

$$m_s(t) = \sum_{l=0}^{L} \beta_l(t) f_l(s), s \in D, t \in T,$$

- $\beta_l(t), l = 0, \ldots, L$, are unknown functional coefficients independent on the spatial location
- $f_l(s), l = 0, \ldots, L$, are known functions of spatial variable $s \in D$, constant with respect to $t \in T$.

Modelling spatial observations - residuals and variogram

- Let $\delta_{s_1}, \ldots, \delta_{s_n}$ be a realization of zero-mean, second-order stationary and isotropic residual process $\{\delta_s, s \in D\}$ [Menafoglio, Secchi 2016]
- Spatial correlation among residuals can be measured via the semivariogram:

$$\gamma(h) = \frac{1}{2}E[||\delta_{s_i} - \delta_{s_j}||^2], s_i, s_j \in D, h = ||s_i - s_j||.$$

- The empirical semivariogram of process is

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

- The empirical variogram is defined as $2\hat{\gamma}(h)$.

Permutation tests for comparison of two functional populations

- Let $\epsilon_{s_i(1)}, i = 1, \ldots, n_1$, and $\epsilon_{s_i(2)}, i = 1, \ldots, n_2$, be two random independent samples of functions in L^2.
- Test of hypothesis

$$H_0 : \mathrm{E}(\epsilon_{s(1)}) = \mathrm{E}(\epsilon_{s(2)}) \text{ and } \mathrm{Var}(\epsilon_{s(1)}) = \mathrm{Var}(\epsilon_{s(2)}), \text{against}$$

$$H_1 : \mathrm{E}(\epsilon_{s(1)}) \neq \mathrm{E}(\epsilon_{s(2)}) \text{ or } \mathrm{Var}(\epsilon_{s(1)}) \neq \mathrm{Var}(\epsilon_{s(2)}).$$

- using test statistics measuring L^2 distance between two sample means and variances:

$$T_m^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{|\mathcal{I}|} [\bar{\epsilon}_{s(1)}(t) - \bar{\epsilon}_{s(2)}(t)]^2 dt,$$

$$T_v^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{|\mathcal{I}|} [\hat{\mathrm{Var}}[\epsilon_{s(1)}(t)] - \hat{\mathrm{Var}}[\epsilon_{s(2)}(t)]]^2 dt.$$

Permutation tests for comparison of two functional populations

- The procedure adapted from [Pini, Vantini 2017] is interval-wise; aims at identifying parts of functional domain where the two groups of data significantly differ.
- Let $\mathcal{I} \subseteq T$ be an arbitrary interval of form (t_1, t_2) or its complement $T \backslash (t_1, t_2)$, where $(t_1, t_2) \subseteq T$. Let $p^{\mathcal{I}}$ be the p-value of functional test

$$H_0^{\mathcal{I}} : \mathrm{E}(\epsilon_{s(1)})^{\mathcal{I}} = \mathrm{E}(\epsilon_{s(2)})^{\mathcal{I}} \text{ and } \mathrm{Var}(\epsilon_{(1)})^{\mathcal{I}} = \mathrm{Var}(\epsilon_{(2)})^{\mathcal{I}}, \text{ against}$$

$$H_1^{\mathcal{I}} : \mathrm{E}(\epsilon_{s(1)})^{\mathcal{I}} \neq \mathrm{E}(\epsilon_{s(2)})^{\mathcal{I}} \text{ or } \mathrm{Var}(\epsilon_{(1)})^{\mathcal{I}} \neq \mathrm{Var}(\epsilon_{(2)})^{\mathcal{I}}.$$

- The adjusted p-value of the test is, for each $t \in T$, defined as

$$p(t) = \sup_{\mathcal{I} \ni t} p^{\mathcal{I}}, \forall\, t \in T.$$

Testing for significance in spatial regression model with functional response

Functional-on-scalar linear model for the drift:

$$\mathcal{X}_s(t) = \sum_{l=0}^{L} \beta_l(t) f_l(s) + \delta_s(t), s \in D, t \in T,$$

Residuals $\delta_s(t), t \in T$ zero-mean, independent and identically distributed random functions with finite total variance.
We aim at testing the hypothesis:

$$H_0 : \beta_1(t) = \ldots = \beta_L(t) = 0 \ , \ \forall l \in \{1, \ldots, L\}, \ \forall t \in T, \ \text{against}$$

$$H_1 : \beta_l(t) \neq 0 \text{ for some } l \in \{1, \ldots, L\} \text{ and some } t \in T,$$

using test statistic

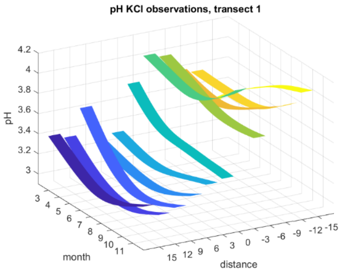
$$T_0 = \int [(C\hat{\beta}(t))' [C(F'\Sigma^{-1}F)C']^{-1} (C\hat{\beta}(t))] dt.$$

Freedman and Lane permutation scheme

1. Estimate residuals of the reduced model (model under H_0).
2. Permute residuals of the reduced model.
3. Compute permuted responses through the reduced model and permuted residuals.
4. Estimate parameters of the full model from permuted responses.
5. Calculate the test statistic T_0.

The global p-value of the test is obtained as the proportion of permutations leading to higher value of test statistic than the one of observed data.

Functional observations



- Data preprocessed using B-spline basis (cubic splines, knots placed at data points, 10 basis functions)
- Observations were smoothed using PENSSE (penalized residual sum of squares) criterion
- Penalisation parameter selected via generalized cross-validation ($\lambda = 10$)

The data are treated as functions of time distributed in space.

Exploring spatial dependence among observations

Drift modelled as:

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t) \cdot soil(s) + \delta_s(t),$$

where $soil(s)$ is the indicator function such that:

$$soil(s) = \begin{cases} 0 & \text{for } s \in \{-15, -12, -9, -6, -3\}, \\ 1 & \text{for } s \in \{3, 6, 9, 12, 15\} \end{cases}$$

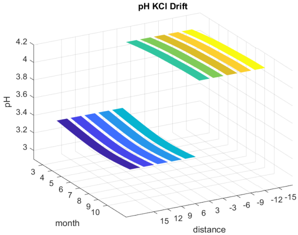
Exploring spatial dependence among observations

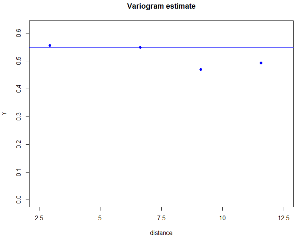
Drift modelled as:

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t) \cdot soil(s) + \delta_s(t),$$

where $soil(s)$ is the indicator function such that:

$$soil(s) = \begin{cases} 0 & \text{for } s \in \{-15, -12, -9, -6, -3\}, \\ 1 & \text{for } s \in \{3, 6, 9, 12, 15\} \end{cases}$$





Testing for differences in field and forest residuals

Let $\delta_{s_i(1)}, i = 1, \ldots, 5$, and $\delta_{s_i(2)}, i = 1, \ldots, 5$, denote the residuals from field and forest soil, respectively. The aim is to test the hypothesis

$$H_0 : \mathrm{E}(\delta_{s(1)}) = \mathrm{E}(\delta_{s(2)}) \text{ and } \mathrm{Var}(\delta_{s(1)}) = \mathrm{Var}(\delta_{s(2)}), \text{against}$$

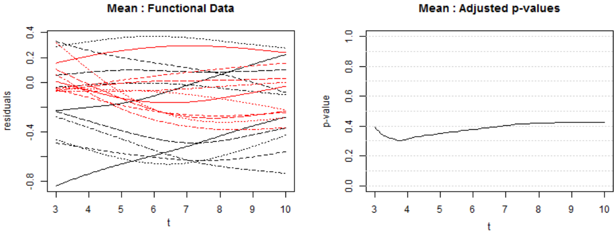
$$H_1 : \mathrm{E}(\delta_{s(1)}) \neq \mathrm{E}(\delta_{s(2)}) \text{ or } \mathrm{Var}(\delta_{s(1)}) \neq \mathrm{Var}(\delta_{s(2)}).$$

Testing for differences in field and forest residuals

Let $\delta_{s_i(1)}, i = 1, \ldots, 5$, and $\delta_{s_i(2)}, i = 1, \ldots, 5$, denote the residuals from field and forest soil, respectively. The aim is to test the hypothesis

$$H_0 : \mathrm{E}(\delta_{s(1)}) = \mathrm{E}(\delta_{s(2)}) \text{ and } \mathrm{Var}(\delta_{s(1)}) = \mathrm{Var}(\delta_{s(2)}), \text{against}$$

$$H_1 : \mathrm{E}(\delta_{s(1)}) \neq \mathrm{E}(\delta_{s(2)}) \text{ or } \mathrm{Var}(\delta_{s(1)}) \neq \mathrm{Var}(\delta_{s(2)}).$$

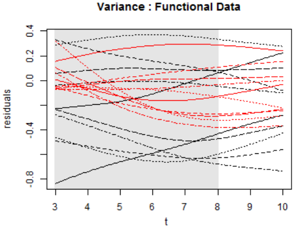


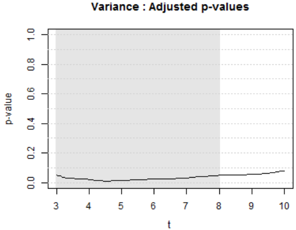
Testing for differences in field and forest residuals

Let $\delta_{s_i(1)}, i = 1, \ldots, 5$, and $\delta_{s_i(2)}, i = 1, \ldots, 5$, denote the residuals from field and forest soil, respectively. The aim is to test the hypothesis

$$H_0 : \mathrm{E}(\delta_{s(1)}) = \mathrm{E}(\delta_{s(2)}) \text{ and } \mathrm{Var}(\delta_{s(1)}) = \mathrm{Var}(\delta_{s(2)}), \text{against}$$

$$H_1 : \mathrm{E}(\delta_{s(1)}) \neq \mathrm{E}(\delta_{s(2)}) \text{ or } \mathrm{Var}(\delta_{s(1)}) \neq \mathrm{Var}(\delta_{s(2)}).$$





Model for data with different variances

- Although the residuals were spatially independent, the test for two population showed that was still some influence of the soil type with respect to variance.
- Instead, a new model is proposed:

$$\mathcal{X}_{s(j)}(t) = \beta_0(t) + \beta_1(t) \cdot soil(s) + \delta_{s(j)}(t), j = 1, 2,$$

$$soil(s) = \begin{cases} 0 & \text{for } s \in \{-15, -12, -9, -6, -3\}, \\ 1 & \text{for } s \in \{3, 6, 9, 12, 15\} \end{cases}$$

- where $\delta_{s(j)}(t) = \sigma_{(j)}\epsilon_s(t), j = 1, 2,$
- $\sigma_{(j)}$ is a standard deviation of residuals changing according the type of soil,
- $\epsilon_s(t)$ are spatially independent identically distributed (and thus permutable) residuals.

Model for data with different variances

- The drift is estimated via weighted least squares with diagonal weight matrix:

$$W = diag\Bigg\{\underbrace{\frac{1}{\hat{\sigma}_{(1)}},\ldots,\frac{1}{\hat{\sigma}_{(1)}}}_{5},\underbrace{\frac{1}{\hat{\sigma}_{(2)}},\ldots,\frac{1}{\hat{\sigma}_{(2)}}}_{5}\Bigg\}.$$

- The variances $\hat{\sigma}^2_{(j)}, j = 1, 2$, estimated from variograms of partial models

$$\mathcal{X}_{s(j)}(t) = \beta_{0(j)}(t) + \delta_{s(j)}, j = 1, 2,$$

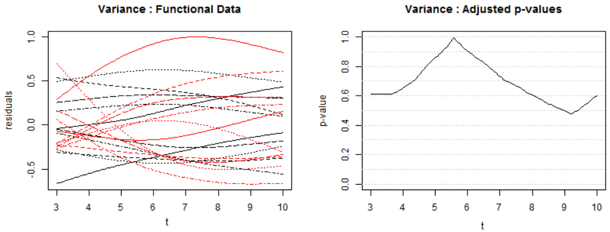
  for field and forest part separately, as a sill of each variogram.
- The estimates are $\hat{\sigma}^2_{(1)} = 0,9452$ and $\hat{\sigma}^2_{(2)} = 0,1684$.
- The variance of field soil residuals is more than 5 times higher than of forest soil residuals.

Model for data with different variances

Let $\epsilon_{s_i(1)}, i = 1, \ldots, n_1$, and $\epsilon_{s_i(2)}, i = 1, \ldots, n_2$, denote the residuals from field and forest soil, respectively. We test the hypothesis

$$H_0 : \mathrm{E}(\epsilon_{s(1)}) = \mathrm{E}(\epsilon_{s(2)}) \text{ and } \mathrm{Var}(\epsilon_{s(1)}) = \mathrm{Var}(\epsilon_{s(2)}), \text{against}$$

$$H_1 : \mathrm{E}(\epsilon_{s(1)}) \neq \mathrm{E}(\epsilon_{s(2)}) \text{ or } \mathrm{Var}(\epsilon_{s(1)}) \neq \mathrm{Var}(\epsilon_{s(2)}).$$



Testing for significance of regression parameters

In model

$$\mathcal{X}_{s(j)}(t) = \beta_0(t) + \beta_1(t) \cdot soil(s) + \sigma_{(j)}\epsilon_s(t), j = 1, 2,$$

we test the null hypothesis:

$$H_0 : \beta_1 = 0, \text{ against } H_1 : \beta_1 \neq 0,$$

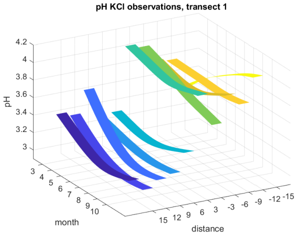
Modified permutation scheme

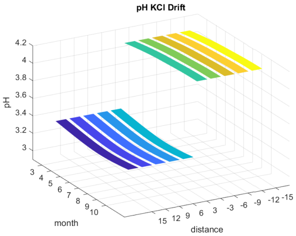
Initial step: estimate $\hat{\sigma}^2_{(j)}, j = 1, 2$, from the partial models.

1. Estimate residuals $\hat{\delta}_{s(j)}(t)$ of the reduced model $\mathcal{X}_{s(j)}(t) = \beta_0(t) + \delta_{s(j)}(t), j = 1, 2$.
2. Divide $\hat{\delta}_{s(j)}(t)$ by corresponding standard deviation $\hat{\sigma}_{(j)}, j = 1, 2 \rightarrow$ exchangeable residuals $\hat{\epsilon}_s(t)$.
3. Permute $\hat{\epsilon}_s(t)$.
4. Compute permuted responses $\mathcal{X}^*_{s(j)}(t)$ through reduced model and permuted residuals $\hat{\delta}^*_{s(j)}(t) = \hat{\sigma}_{(j)}\hat{\epsilon}^*_s(t), j = 1, 2$.
5. Estimate parameters of the full model from permuted responses $\mathcal{X}^*_{s(j)}(t), j = 1, 2$.
6. Calculate the test statistic T_0.

Conclusion

- A total number of 1000 permutation was performed and the resulting global *p*-value = 0 was computed.
- The null hypothesis is rejected on the significance level $\alpha = 0,05$. The type of soil significantly affects the potassium chloride pH.





Future steps

- Extend the methodology to more complex spatial structures.
- Develop a functional test for spatial dependence.

References

- 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