www.graspa.org

Data Fusion for functional data

Rosalba Ignaccolo [a]
with
Stefano Bande [b], ***Maria Franco-Villoria*** [a] ***and Alex Giunta*** [a]

[a]Università degli Studi di Torino
[b]ARPA Piemonte

Bergamo, July, 24, 2017

https://sites.google.com/site/ephastat/

www.stephiproject.it

Background and Goal

- **Data Fusion in Air Quality**: combining various data sources in order to obtain an accurate air quality assessment.
  In particular
  - observed concentrations gathered from irregularly spaced sites of monitoring networks
  - numerical output of chemical transport models on a regular thick grid

- **Goal:** to propose a spatial data fusion strategy based on a functional kriging model with external drift.

- **Applied Goal:** to implement and compare our proposal to *feasible* alternatives for pollutants in Piemonte, Italy.

Observed Data

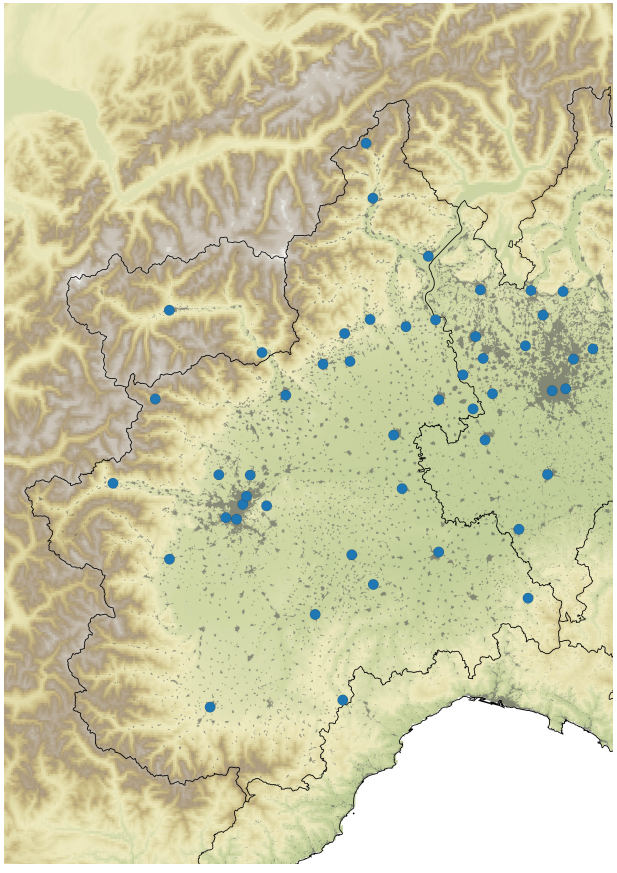


Figure: Monitoring sites for PM_{10}.

ARPA Piemonte monitoring network in 2015

- PM_{10} : daily concentration in $\mu g/m^3$, 50 sites $\times$ 365 observations
- NO_2 : hourly concentration in $\mu g/m^3$, 53 sites $\times$ 8760 observations
- O_3 : daily maximum of 8h moving average in $\mu g/m^3$, 30 sites $\times$ 365 observations

Data from ARPA CTM modelling system (**A**)

Chemistry Transport Model (CTM) **FARM** (Flexible Air quality Regional Model) with photochemistry and aerosols module:

three-dimensional deterministic modelling system capable to simulate air pollutant emission, transport, diffusion and chemical reactions

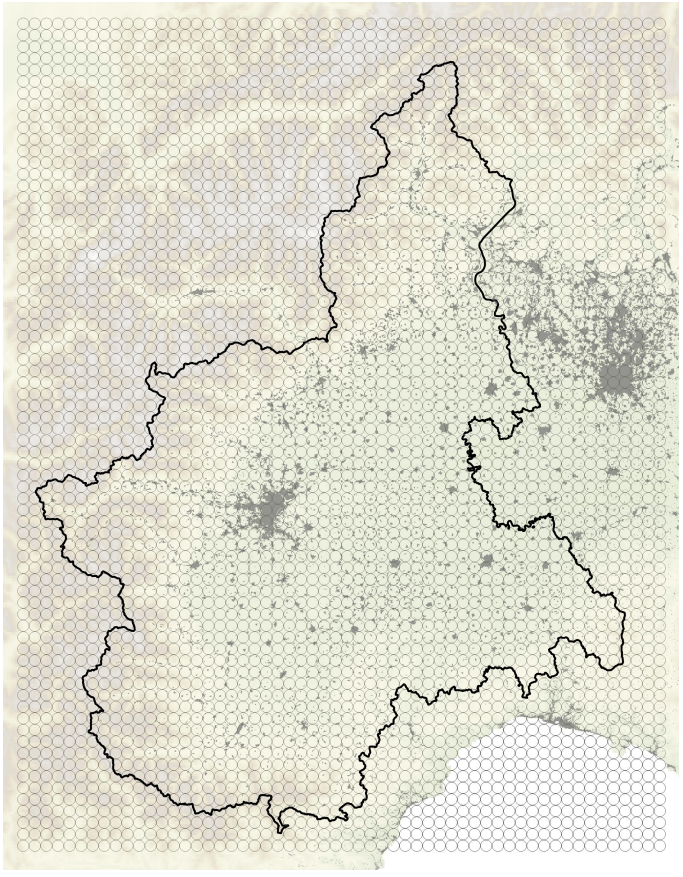


Figure: Spatial grid: 56x72 square cells with a 4 km resolution.

B: Spatial KED day by day - ***now adopted***

At time $t \in T$ and location $s \in D \subseteq \mathbb{R}^2$,

$$Y(s) = \mu(s) + \epsilon(s) \quad \text{with} \quad \mu(s) = \alpha + \beta X(s)$$

where $X(s) =$ FARM output (at a fixed time).

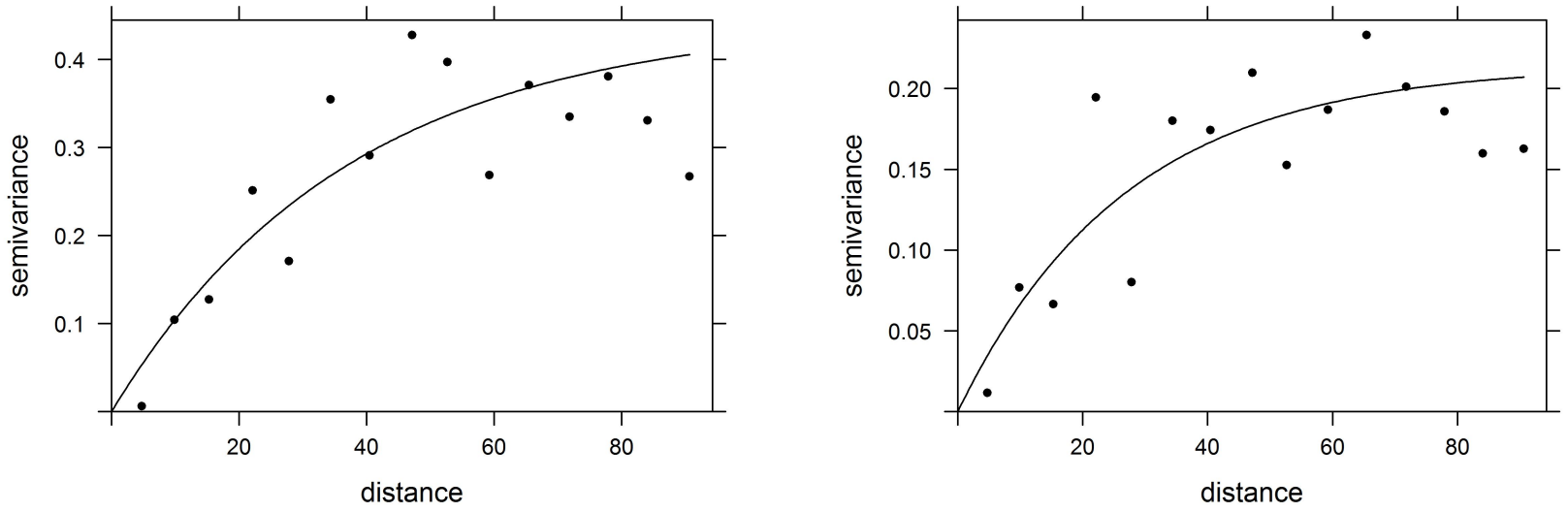


Figure: Residual Variogram when $Y = \log(PM_{10})$ in January 1st (left) and January 2nd (right).

Functional Kriging with External Drift (FKED) (1)

Ignaccolo *et al.* (2014) *SERRA*

Let $\Upsilon_s = \{Y_s(t); t \in T\}$ be a *functional random variable* observed at location $s \in D \subseteq \mathbb{R}^d$, whose realization is a function of $t \in T$ - that is a functional data - where T is a compact subset of $\mathbb{R}$

Assume that we observe a sample of curves Υ_{s_i}, for $s_i \in D$, $i = 1, \ldots, n$, that take values in a separable Hilbert space of square integrable functions

The set $\{\Upsilon_s, s \in D\}$ constitutes a *functional random field* or a *spatial functional process*, that can be non-stationary and whose elements are supposed to follow the model

$$\Upsilon_s = \mu_s + \epsilon_s \tag{1}$$

The term μ_s is interpreted as a drift describing a spatial trend while ϵ_s represents a residual random field that is zero-mean, second-order stationary and isotropic, so that

i) $\mathbb{E}(\Upsilon_s) = \mu_s$, $s \in D$;

ii) $\mathbb{E}(\epsilon_s) = 0$, $s \in D$;

iii) $Cov(\epsilon_{s_i}, \epsilon_{s_j}) = C(h)$, $\forall s_i, s_j \in D$ with $h = \|s_i - s_j\|$.

Functional Kriging with External Drift (FKED) (2)

At the generic site s_i, $i = 1, \ldots, n$, and at point t model (1) can be rewritten as a functional concurrent linear model

$$Y_{s_i}(t) = \mu_{s_i}(t) + \epsilon_{s_i}(t)$$

with the drift

$$\mu_{s_i}(t) = \alpha(t) + \sum_p \gamma_p(t) C_{p,i} + \sum_q \beta_q(t) X_{q,i}(t)$$

- $\alpha(t)$ is a functional intercept
- $C_{p,i}$ is the p-th scalar covariate at site s_i
- $X_{q,i}$ is the q-th functional covariate at site s_i
- $\gamma_p(t)$ and $\beta_q(t)$ are the covariate coefficients
- $\epsilon_{s_i}(t)$ represents the residual spatial functional process $\{\epsilon_s(t), t \in T, s \in D\}$ at the site s_i

Functional Kriging with External Drift (FKED) (3)

1. functional regression model with functional response and scalar and functional covariates fitted to estimate drift coefficients and obtain functional residuals

$$e_{s_i}(t) = Y_{s_i}(t) - \hat{\mu}_{s_i}(t) = Y_{s_i}(t) - \left[\hat{\alpha}(t) + \sum_j \hat{\gamma}_j(t) C_{j,i} + \sum_l \hat{\beta}_l(t) X_{l,i}(t)\right]$$

**spatial correlation taken into account using an iterative algorithm whose convergence is determined based on AIC*

2. residual curve prediction at the unmonitored site s_0 obtained by ordinary kriging for functional data (OKFD)

$$\hat{e}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i e_{s_i}(t)$$

with λ_i depending on the *trace-variogram*

3. prediction at the unmonitored site s_0 obtained by adding the two terms

$$\hat{Y}_{s_0}(t) = \hat{\mu}_{s_0}(t) + \hat{e}_{s_0}(t)$$

C: FKED for Data Fusion

$$\tilde{Y}_{s_i}(t) = \mu_{s_i}(t) + \epsilon_{s_i}(t) \quad \text{with} \quad \mu_{s_i}(t) = \alpha(t) + \beta(t)\tilde{X}_{s_i}(t)$$

where $\tilde{X}_{s_i}(t) =$ 'smoothed' FARM output (as functional datum) at site s_i

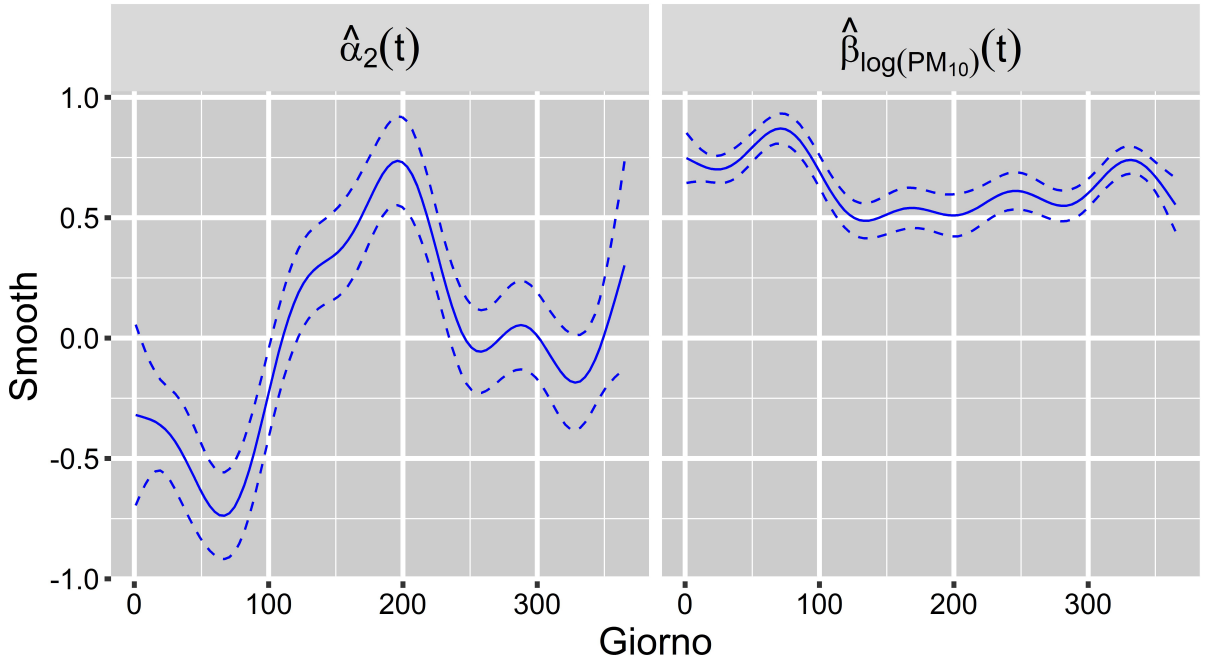


Figure: Functional coefficients for the $\log(PM_{10})$ case.

Bootstrap based uncertainty bands for FKED prediction

For the methodology see **Franco-Villoria & Ignaccolo (2017) *Spatial Stats***

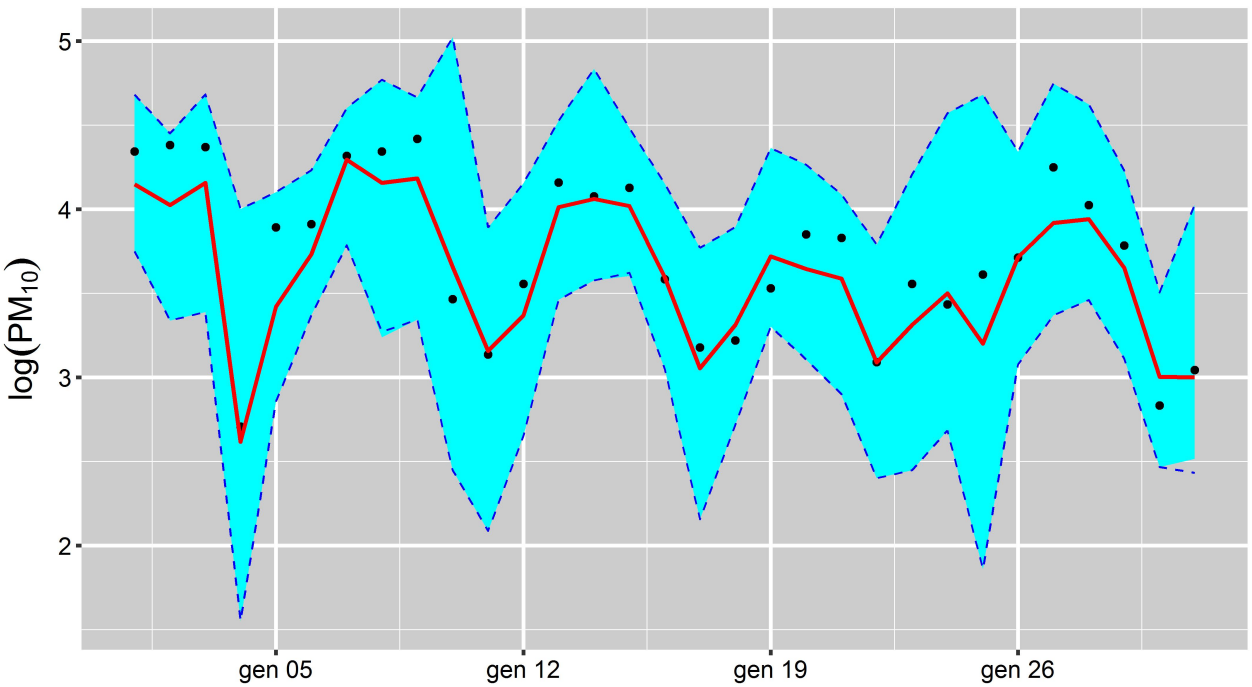


Figure: Bootstrap based uncertainty bands for $\log(PM_{10})$. Beinasco (TRM) - Aldo Mei station, January 2015.

Note that we have uncertainty band width changing along the domain

D: Spatio-Temporal KED

$$Y_{s_i}(t) = \mu_{s_i}(t) + \epsilon_{s_i}(t) \quad \text{with} \quad \mu_{s_i}(t) = \alpha(t) + \beta(t)X_{s_i}(t)$$

where $X_{s_i}(t) =$ FARM output at site s_i and time t

We have *time-varying coefficient in the drift* and Product-Sum covariance

$$C_{\text{ps}}(h,u) = kC_{\text{s}}(h)C_{\text{t}}(u) + C_{\text{s}}(h) + C_{\text{t}}(u) \quad \text{with} \quad k > 0$$

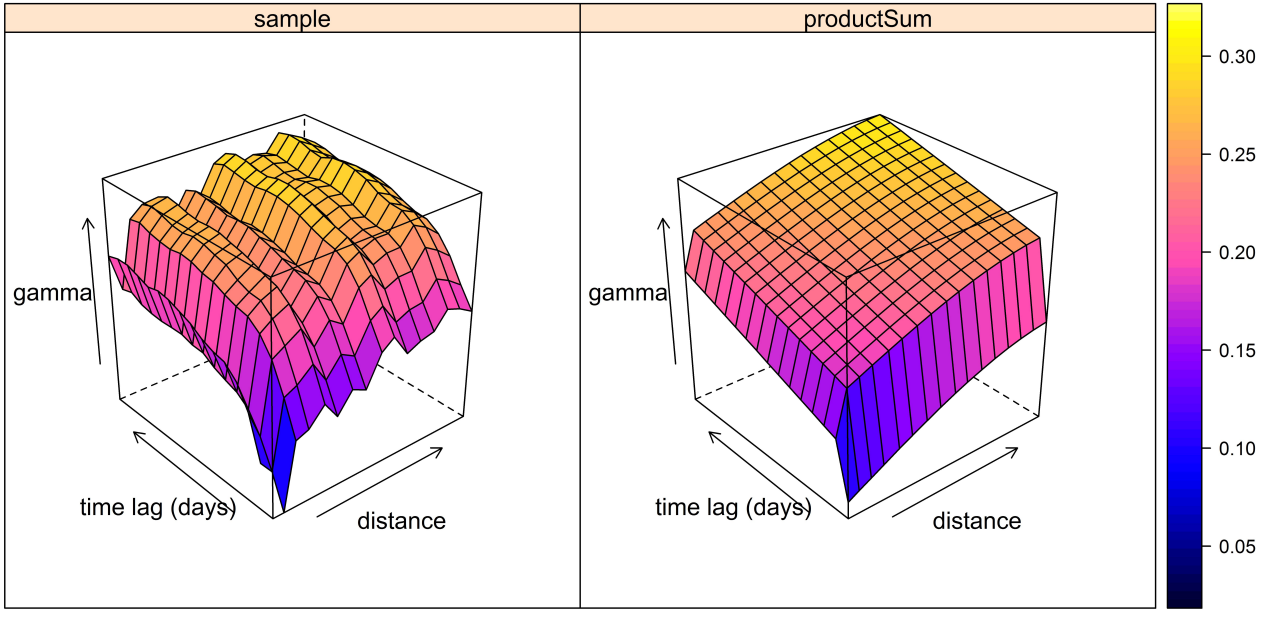


Figure: Residuals spatio-temporal variogram; log(PM_{10}) case.

Back-transformation for log-normal kriging

For air pollutants a log transformation is used:

$$Y_s(t) = \log Z_s(t), \quad s \in D, t \in T$$

With a common bias correction we have

$$\hat{Z}_{s_0}(t) = \exp\left(\hat{\mu}_{s_0}(t) + \hat{e}_{s_0}(t) + \frac{SE^2_{s_0}(t) + \hat{\sigma}^2_{s_0}(t)}{2} - m_{\hat{e}}\right)$$

where at s_0

- $\hat{\mu}_{s_0}(t)$ is the estimated trend,
- $\hat{e}_{s_0}(t)$ is the predicted residual via kriging,
- $SE_{s_0}(t)$ is the trend standar error,
- $\hat{\sigma}^2_{s_0}(t)$ is the estimated kriging variance,
- $m_{\hat{e}}$ is the Lagrange multiplier in the Y scale.

Which variance for FKED?

In case of the ordinary kriging for functional data (OKFD in **Giraldo et al. (2010) EES**) the *prediction trace-variance* is given by

$$\sigma^2(s_0) = \int_T V\big(\hat{e}_{s_0}(t) - e_{s_0}(t)\big)dt = \sum_{i=1}^{n} \upsilon(\| s_i - s_0 \|) - \mu,$$

based on the *trace-semivariogram*

$$\upsilon(h) = \int_T \frac{1}{2} Var\left(e_{s_i}(t) - e_{s_j}(t)\right) dt$$

estimated by

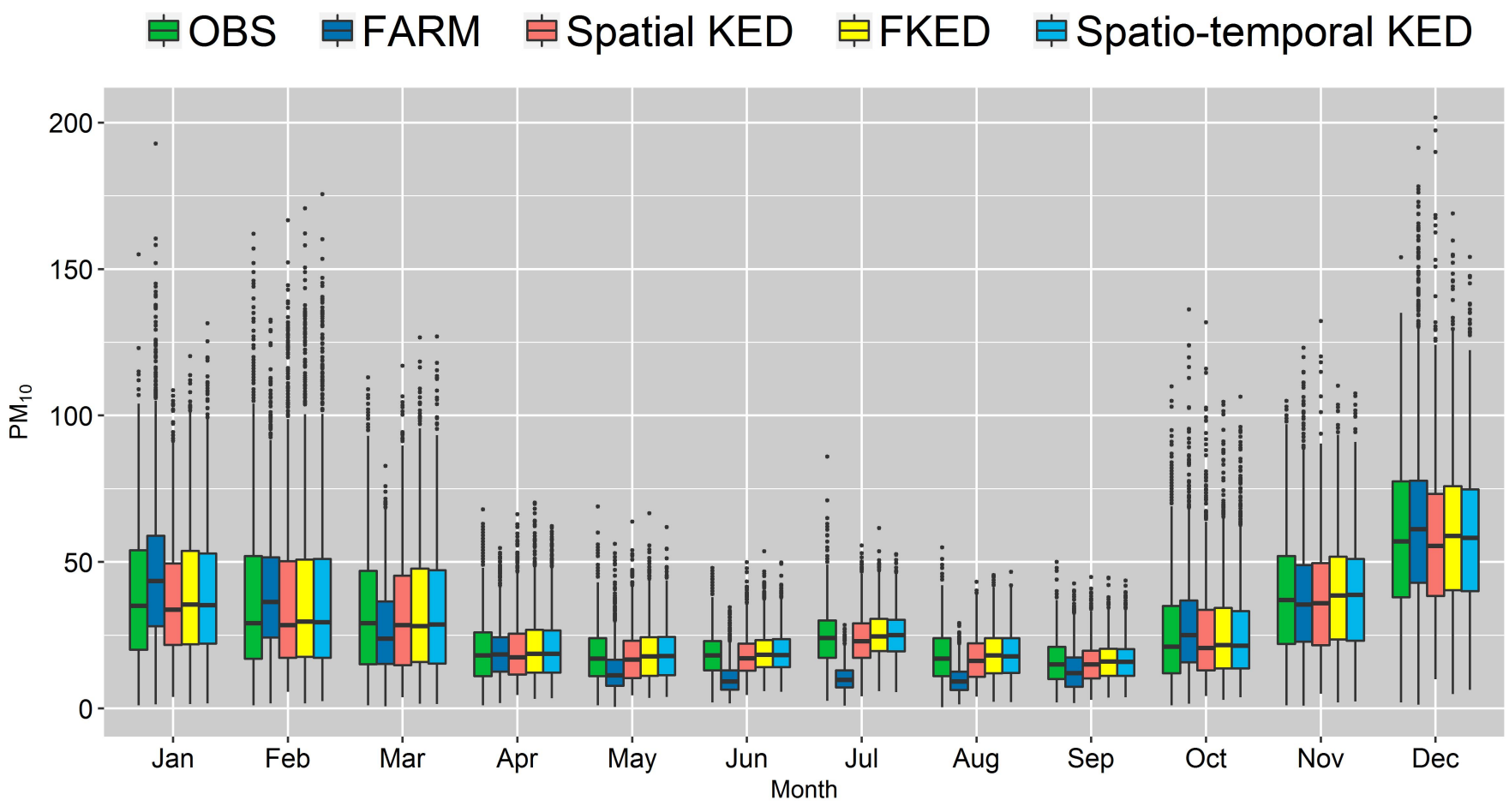
$$\hat{\upsilon}(h) = \frac{1}{2\,|N(h)|} \sum_{i,j \in N(h)} \int_T \left(e_{s_i}(t) - e_{s_j}(t)\right)^2 dt$$

where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$

Temporary strategy: to have a variance for each time t we use

$$\hat{\sigma}^2_{s_0}(t) = \frac{\hat{\sigma}^2(s_0)}{\text{n° of time points}}$$

PM_{10} in original scale – LOOCV on 50 sites



Performance indexes

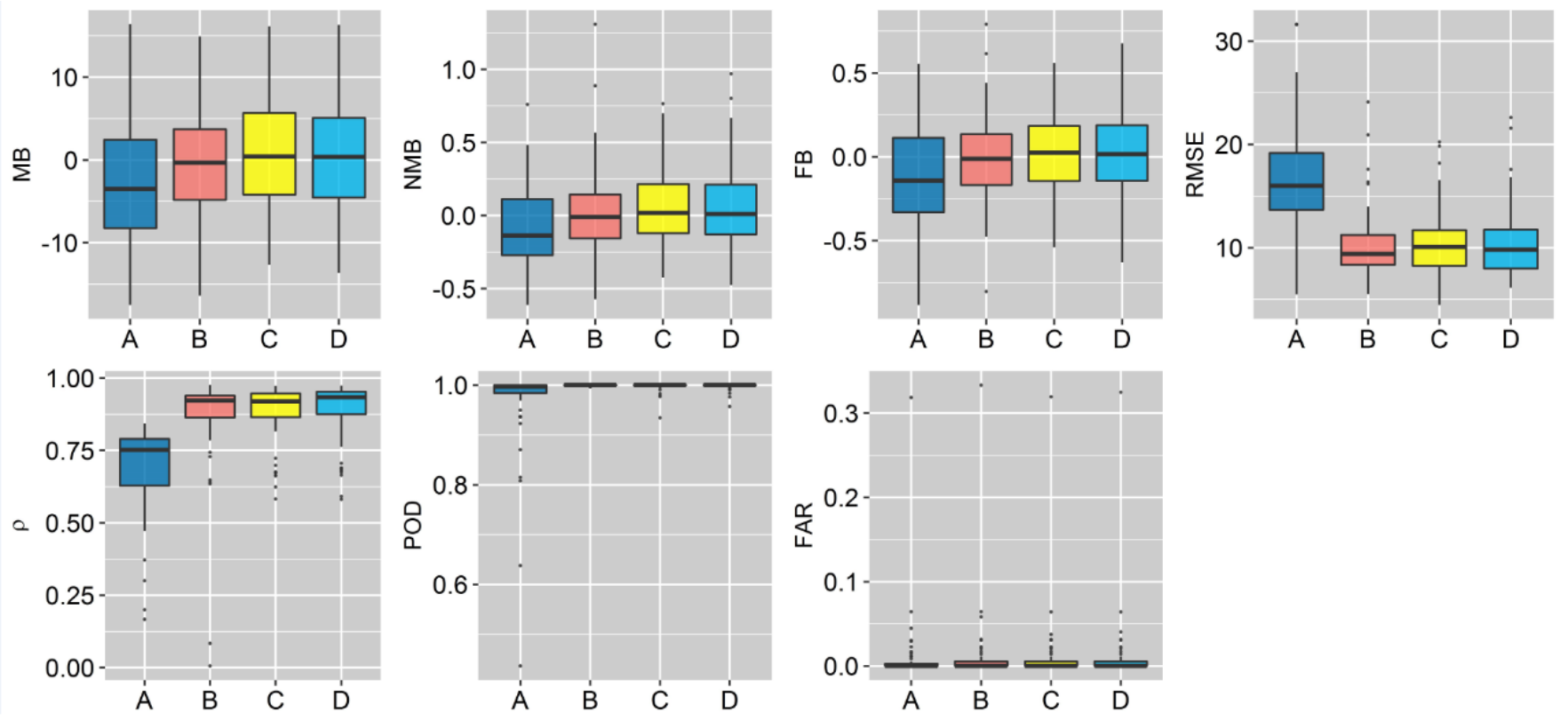
Observed values and predicted values are represented by z_j and $\hat{z}_j$.

- *Mean bias*: $MB = \frac{1}{N}\sum_{j=1}^{N}(\hat{z}_j - z_j)$.
- *Normalised mean bias*: $NMB = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{z}_j - z_j}{z_j}\right)$.
- *Fractional bias*: $FB = \frac{(\bar{\hat{z}} - \bar{z})}{0.5\left(\bar{\hat{z}} + \bar{z}\right)}$.
- *Root mean square error*: $RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{z}_j - z_j)^2}$.
- *Pearson correlation coefficient*: ρ

Looking at threshold exceedances, "Yes" = {z > Limit Value}:

- *Probability of detection*: $POD = \frac{hits}{hits + misses}$. Answers the question: What fraction of the observed "yes" events were correctly forecast?
- *False alarm ratio*: $FAR = \frac{false\ alarms}{hits + false\ alarms}$. Answers the question: What fraction of the predicted "yes" events actually did not occur (i.e., were false alarms)?

PM_{10} in original scale – LOOCV on 50 sites: performance



Conclusions so far...

Performance in log scale:

- PM_{10}: FKED
- NO_2: Spatial KED
- O_3: Additive Model (only drift)

Computational cost:

- Spatial KED: low
- FKED: low
- Space-Time KED: very high, if not prohibitive

Future: ...

Bootstrap based log-normal kriging (work in progress)

- Deriving a functional variance by bootstrap replicates
- Deriving an empirical bootstrap distribution of $\exp(Y)$
- Deriving a bias correction - multiplicative or additive - via bootstrap by extending **Rister and Lahiri (2013) Stat. Modelling**

Thank you! Grazie!

www.stephiproject.it

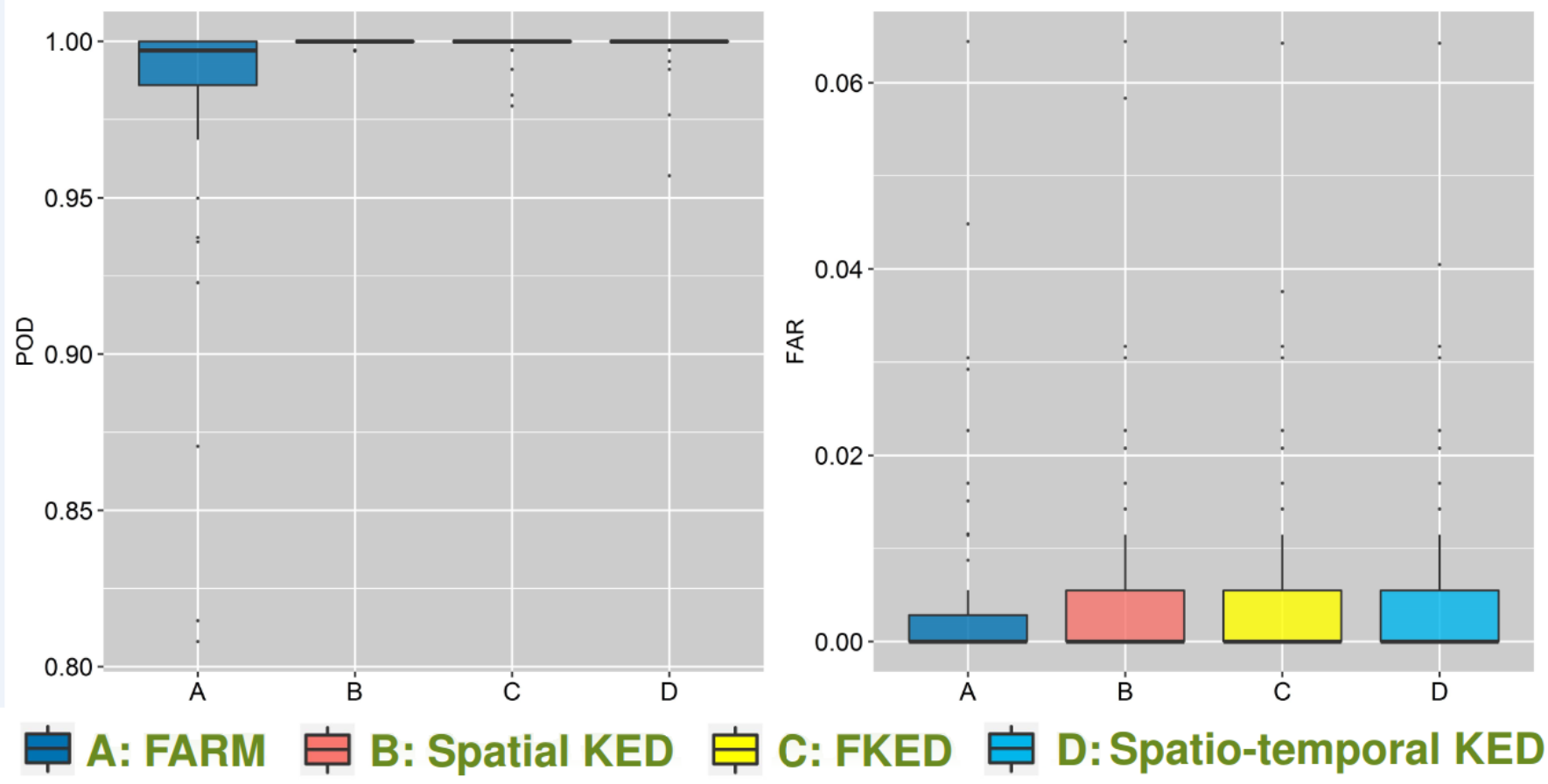
https://sites.google.com/site/ephastat/

⇝ Ignaccolo R, Mateu J, Giraldo R (2014) Kriging with external drift for functional data for air quality monitoring. *SERRA* **28**, 1171-1186

⇝ Franco-Villoria M, Ignaccolo R (2017) Bootstrap based uncertainty bands for prediction in functional kriging. *Spatial Statistics* **21**, 130-148

R code available!

FAR — POD


POD
1.00
0.95
0.90
0.85
0.80
A
B
C
D
FAR
0.06
0.04
0.02
0.00
A
B
C
D
A: FARM
B: Spatial KED
C: FKED
D: Spatio-temporal KED