Prediction Regions for Functional-Valued Random Forests

Diego Serrano and Eduardo García-Portugués
Department of Statistics, Carlos III University of Madrid

2025 International Workshop on Functional and Operatorial Statistics IWFOS

Abstract

- Random Forests (RF) have been generalized to metric spaces [3].
- We present **prediction balls** [1, 2] to quantify the uncertainty in a RF prediction with metric-space-valued data, thus being applicable to functional data.
- Asymptotic coverage theory is presented in four probability coverage types.
- Prediction balls are illustrated in $W_2(\mathbb{R})$ and with real data on $\mathbb{S}^2$.

Metric space data analysis

- Increasingly complex data types: **functions**, directions, shapes, covariance matrices, ...
- Instead of using on **vector space** properties, analyze data as elements in a **metric space**:

> - **High generality**: only a distance function is required.
> - Exploiting **specific properties** of each space can enhance flexibility and efficiency.
> - **Setting**: regression problem with $(X, Y) \in (\mathcal{X}, d_{\mathcal{X}}) \times (\mathcal{Y}, d_{\mathcal{Y}})$.

- The **Fréchet mean** and **variance** adapt the mean and variance to metric spaces:

$$y_{\oplus} := \underset{y \in (\mathcal{Y}, d_{\mathcal{Y}})}{\arg\min}\ \mathsf{E}\left(d_{\mathcal{Y}}(Y, y)^2\right), \quad V_{\oplus} := \mathsf{E}\left(d_{\mathcal{Y}}(Y, y_{\oplus})^2\right).$$

- Given a sample $\mathcal{L}_n := \{(X_i, Y_i)\}_{i=1}^n$, the **empirical Fréchet mean** and **variance** are

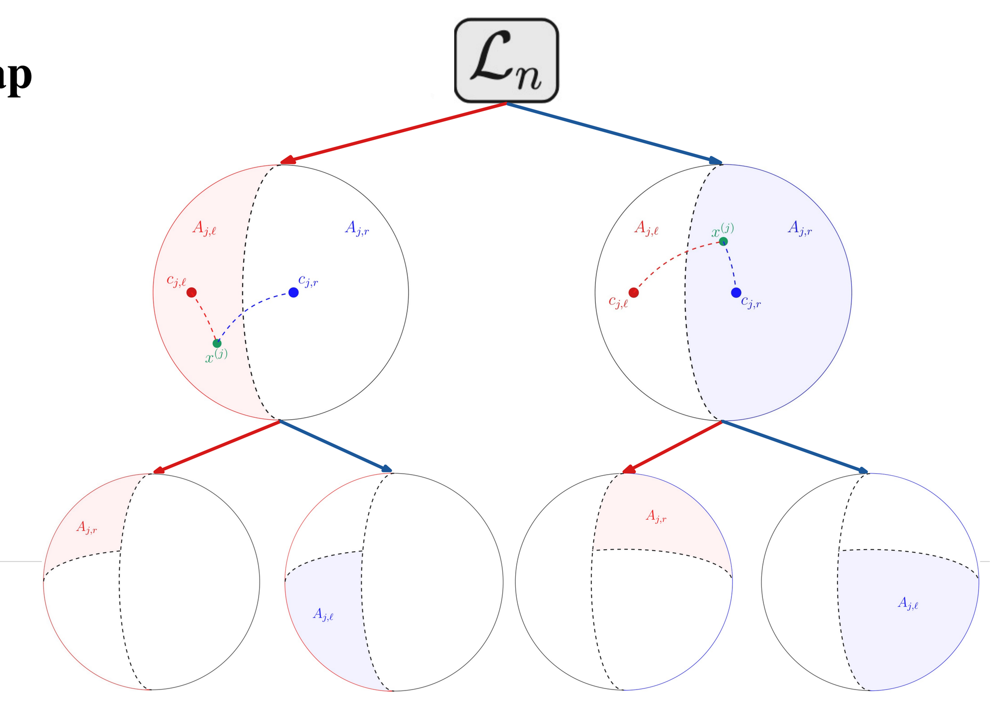
$$\hat{y}_{\oplus} := \underset{y \in (\mathcal{Y}, d_{\mathcal{Y}})}{\arg\min}\ \frac{1}{n}\sum_{i=1}^n d_{\mathcal{Y}}(Y_i, y)^2, \quad \widehat{V}_{\oplus} := \frac{1}{n}\sum_{i=1}^n d_{\mathcal{Y}}(Y_i, \hat{y}_{\oplus})^2.$$

- **Fréchet regression** generalizes the standard Euclidean regression to metric spaces:

$$m_{\oplus}(x) := \underset{y \in (\mathcal{Y}, d_{\mathcal{Y}})}{\arg\min}\ M_{\oplus}(x, y), \quad M_{\oplus}(x, y) := \mathsf{E}\left(d_{\mathcal{Y}}(Y, y)^2 \mid X = x\right).$$

Random forests in metric spaces

- Each tree is built with a **bootstrap resample with replacement** $\mathcal{L}_n^{*b}$.
- The goal is to estimte $m_{\oplus}(x)$.
- Split by centroids $(c_{j,\ell}, c_{j,r})$. Let:
  i. $d_l := d_{\mathcal{X}_j}(x^{(j)}, c_{j,\ell})$ and $A_{j,\ell} := \{(x, y) \in A : d_l \leq d_r\}$.
  ii. $d_r := d_{\mathcal{X}_j}(x^{(j)}, c_{j,r})$ and $A_{j,r} := \{(x, y) \in A : d_l > d_r\}$.



- The **CART** criterion measures split quality (decrease in variance after the split):

$$H_j(A, c_{j,\ell}, c_{j,r}) := \widehat{V}_{\oplus}(A) - \frac{|A_{j,\ell}|}{|A|}\widehat{V}_{\oplus}(A_{j,\ell}) - \frac{|A_{j,r}|}{|A|}\widehat{V}_{\oplus}(A_{j,r}).$$

- Let τ_x denote the terminal node for x, the **Fréchet tree prediction** for x is

$$\widehat{m}_T(x) := \underset{y \in (\mathcal{Y}, d_{\mathcal{Y}})}{\arg\min}\ \frac{1}{|\tau_x|}\sum_{i=1}^n d_{\mathcal{Y}}(Y_i, y)^2\, 1_{\{(X_i, Y_i) \in \tau_x\}}.$$

- The **Fréchet Random Forest** (**FRF**) is the Fréchet mean of $\widehat{m}_T(x)$'s (each from a different $\mathcal{L}_n^{*b}$).

$$\widehat{m}_{\mathrm{FRF}}(x) := \underset{y \in (\mathcal{Y}, d_{\mathcal{Y}})}{\arg\min}\ \frac{1}{B}\sum_{b=1}^B d_{\mathcal{Y}}(\widehat{m}_{T_b}(x), y)^2.$$

- We will use an improved version of FRFs [3], in which $M_n(x, y)$ is estimated through a weighted Fréchet mean, with weights generated by the Fréchet tree.

Estimating uncertainty

- For **Euclidean data**, prediction intervals using **Out-Of-Bag** (OOB) observations from a single **RF** have been developed [4]. We want to extend these ideas to metric spaces.

> **Advantage:** Leverage **RF structure** to use **full sample, no additional training cost**.

- For the resample $\mathcal{L}_n^{*b}$, we say that (X_i, Y_i) is OOB if $(X_i, Y_i) \in \mathcal{L}_n \setminus \mathcal{L}_n^{*b}$. We denote by $\widehat{Y}_{(i)}$ the **OOB prediction** of Y_i.
- The OOB radial errors $\widehat{R}_i^{\mathrm{oob}} := d_{\mathcal{Y}}(Y_i, \widehat{Y}_{(i)})$ estimate $d_{\mathcal{Y}}(Y_i, \widehat{Y}_i)$.

Definition (Prediction balls)

The OOB **prediction ball** for predictors $x \in \mathcal{X}$ with significance level $\alpha \in (0, 1)$ is

$$\mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(x, \mathcal{L}_n) := \left\{y \in \mathcal{Y} : d_{\mathcal{Y}}(\widehat{m}(x), y) < \widehat{R}_{[1-\alpha, n]}\right\},$$

where $\widehat{R}_{[1-\alpha, n]}$ denotes the $(1-\alpha)$-quantile of the ECDF based on $\widehat{R}_1^{\mathrm{oob}}, \ldots, \widehat{R}_n^{\mathrm{oob}}$.

Asymptotic properties

For $\alpha \in (0, 1)$, we considered **four probability coverage types**:

Type I	$\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n)\right\}$	Type II	$\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n) \mid \mathcal{L}_n\right\}$
Type III	$\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n) \mid X = x\right\}$	Type IV	$\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n) \mid \mathcal{L}_n, X = x\right\}$

Theorem (Coverage guarantees)

Under certain conditions [2], the OOB prediction ball has asymptotically correct coverage rate (Types I–IV) for any significance level $\alpha \in (0, 1)$; i.e., as $n \to \infty$:

I $\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n)\right\} \to 1 - \alpha$, **II** $\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n) \mid \mathcal{L}_n\right\} \xrightarrow{\mathsf{P}} 1 - \alpha$,

III $\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n) \mid X = x\right\} \to 1 - \alpha$, **IV** $\mathsf{P}\left\{Y \in \mathrm{PB}_{1-\alpha}^{\mathrm{oob}}(X, \mathcal{L}_n) \mid \mathcal{L}_n, X = x\right\} \xrightarrow{\mathsf{P}} 1 - \alpha$.

Numerical experiments in $W_2(\mathbb{R})$

- We study the **2-Wasserstein space** $\mathcal{W}_2(\mathbb{R})$ endowed with the 2-Wasserstein metric $d_{\mathcal{W}_2}$.
- Consider the regression function

$$x \in [0, 1] \mapsto m(x)(\cdot) = \mathsf{E}(Y(\cdot) \mid X = x) = \tfrac{1}{4} - \log(1 + x) + \left(\tfrac{1}{2} + x^2\right)\Phi^{-1}(\cdot),$$

where Φ^{-1} is the quantile function of a $\mathcal{N}(0, 1)$. To generate the response, set

$$Y(\cdot) = C - \log(1 + X) + (S + X^2)\Phi^{-1}(\cdot), \text{ with } C \sim \Gamma(\tfrac{1}{2}, \tfrac{1}{2}),\ X \sim U(0, 1),$$

and $S \sim \mathrm{Exp}(2)$ independent of X.

Figure 1: On the left panel, reported coverage (Types II and IV). On the central and right panels, example of a prediction ball for $X = 0.5$ and $\alpha = 0.01, 0.1$, respectively.

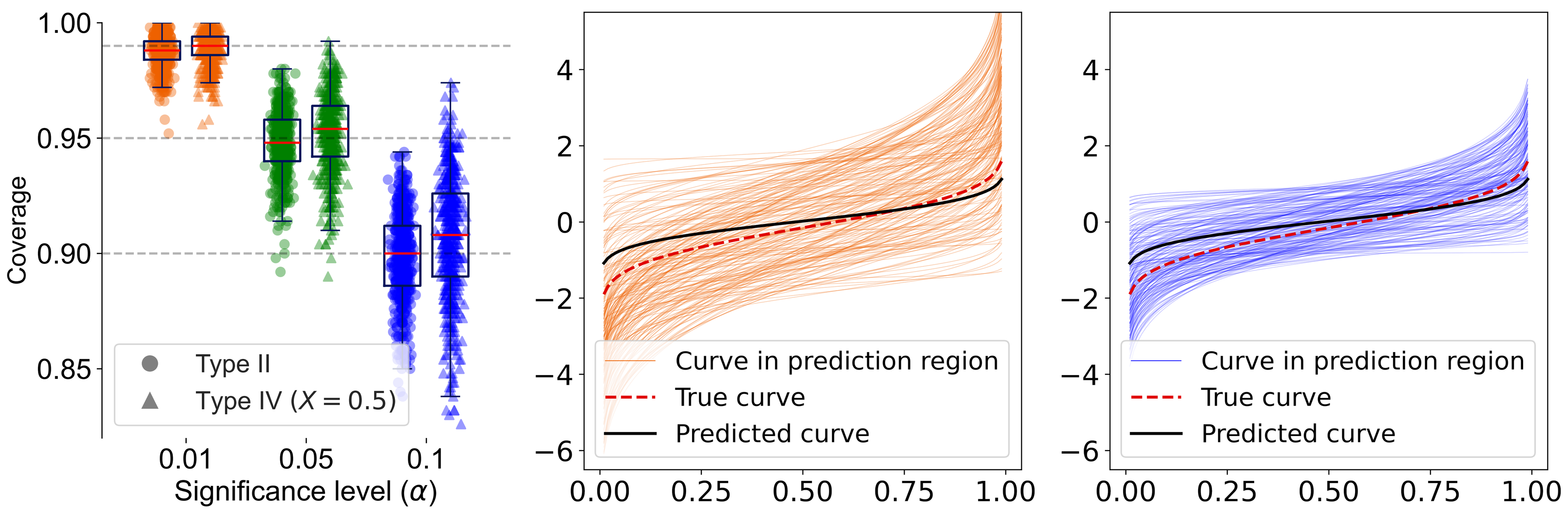


Table 1: Sample mean across 50 estimations of Types I and III coverages for prediction balls.

	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Type I	99.2	94.6	91.0
Type III	99.2	95.8	91.6

Prediction balls for sunspot dynamics

- The Sun's **differential rotation** creates regions of intense magnetic pressure (**sunspots**).
- Where does a sunspot "die" (last recorded observation), based on "birth" (first record)?
- Tough problem: no clear movement direction (Goal) **quantify prediction uncertainty**.
- Larger displacement along parallels than meridians. Non-isotropic distance?
- Consider a **spheroid** $S_{a,c}^2$, tune (a, c) (geometry) to minimize ball area (cross-validation).

> Map data from $\mathbb{S}^2$ to $S_{a,c}^2$ → Compute prediction balls on $S_{a,c}^2$ → Map balls to $\mathbb{S}^2$

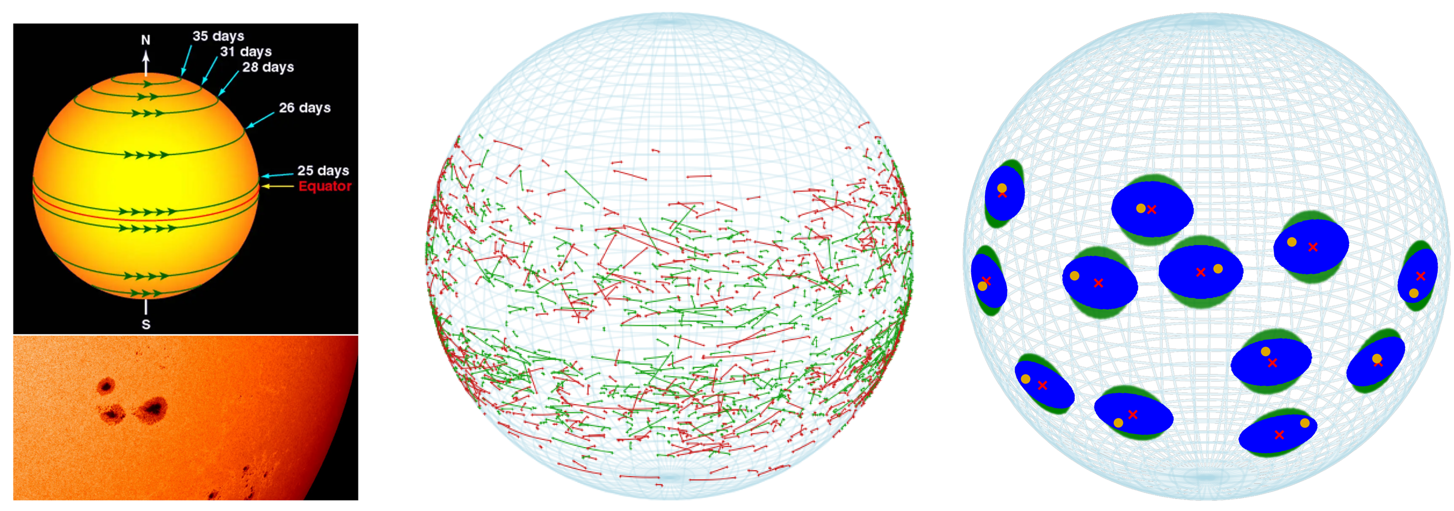


Figure 3: Left: differential rotation (top) and sunspots (bottom). Sources: www.nasa.gov and www.csiro.au. Center: displacements of sunspots (green, counterclockwise; red, clockwise). Right: 90% prediction balls. Green balls correspond to $\mathbb{S}^2$ and blue balls to $S_{0.6,1}$. Red cross: prediction; yellow dot: observed location.

Conclusions

- Prediction balls **estimate the uncertainty** in a RF prediction with metric data.
- **Specificity of RFs** (**OOB errors**) allows improvements over split-conformal methods.
- **Asymptotic theoretical** guarantees (four probability coverage types).
- Correct **finite sample** performance (numerical experiments in $W_2(\mathbb{R})$).

References

[1] Serrano, D. and García-Portugués, E. (2025). Prediction regions for functional-valued random forests. In: Aneiros, G., Bongiorno, E. G., Goia, A., and Hušková, M. (eds) *New Trends in Functional Statistics and Related Fields. IWFOS 2025. Contributions to Statistics*. Springer, Cham.
[2] Serrano, D. and García-Portugués, E. (2025). Random forest prediction balls. *Work in progress*.
[3] Qiu, R., Yu, Z., and Zhu, R. (2024). Random forest weighted local Fréchet regression with random objects. *J. Mach. Learn. Res.*, 25(107):1–69.
[4] Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2020). Random forest prediction intervals. *Am. Stat.*, 74(4):392–406.
[5] Capitaine, L., Bigot, J., Thiébaut, R., and Genuer, R. (2024), *Fréchet random forests for metric space valued regression with non Euclidean predictors*. J. Mach. Learn. Res., 25(355):1–41.