# *Random forest for functional data*

**Tomasz Górecki**[1], Piotr Kokoszka[2], Felix Gnettner[2]

[1]Faculty of Mathematics and Computer Science,
Adam Mickiewicz University, Poznań, Poland
[2]Department of Statistics,
Colorado State University, USA

V Kongres Statystyki Polskiej
Warszawa 1-3.07.2025


# Contents

# Functional data

Functional data consists of observations that are functions rather than scalar or vector values. Each observation is typically a real-valued function $X(t)$ defined on a continuous domain $\mathcal{T}$, such as time or space.

Formally, functional data can be viewed as realizations of a stochastic process $\{X(t) : t \in \mathcal{T}\}$, where each $X(t)$ is smooth and lies in an infinite-dimensional function space, often $L^2(\mathcal{T})$.

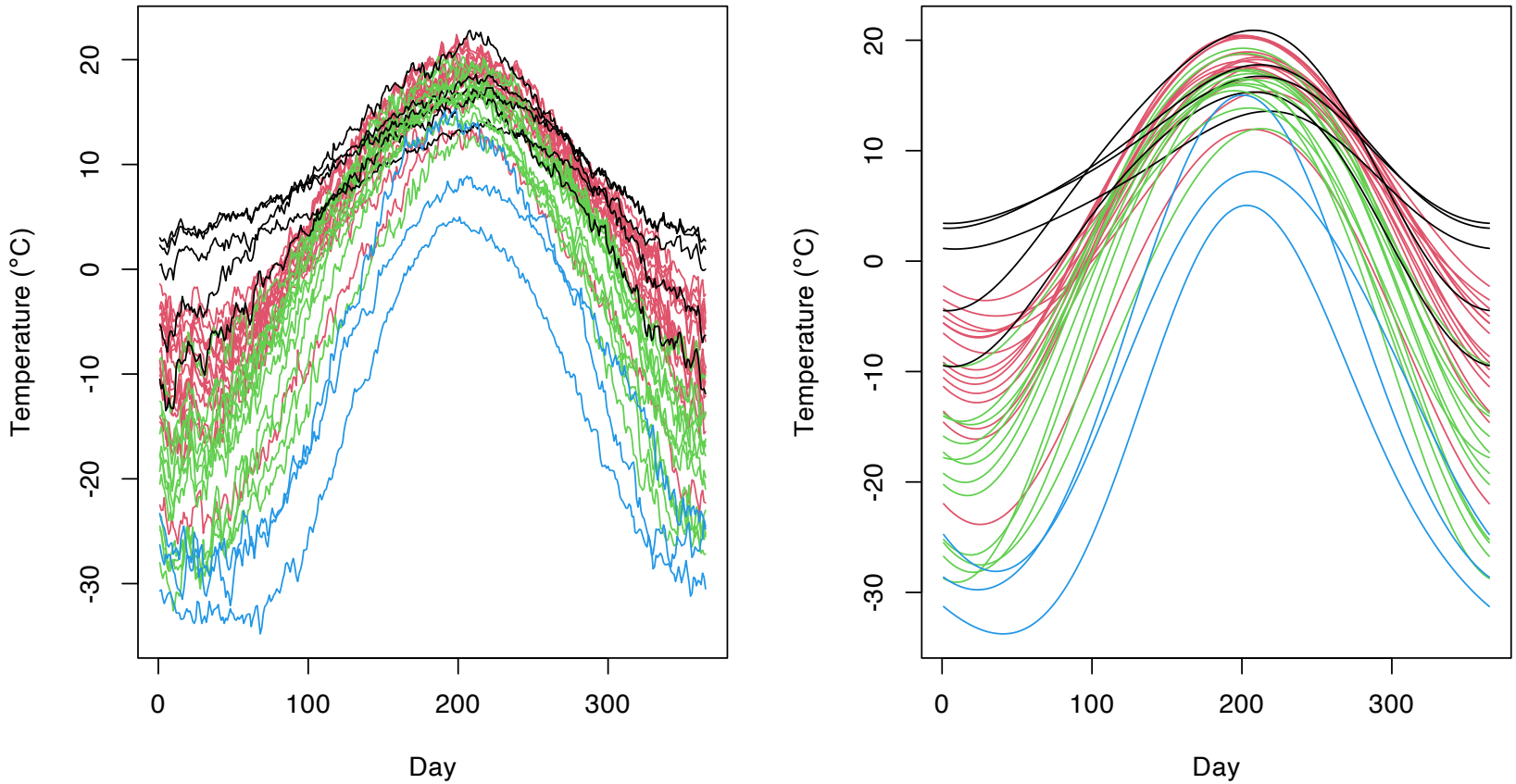
# Functional data

In practice, functional data are often observed at discrete time points and require smoothing or approximation before analysis. A common approach is to represent each observed function $X_i(t)$ as a linear combination of basis functions:

$$X_i(t) \approx \sum_{k=1}^{K} c_{ik}\phi_k(t),$$

where $\{\phi_k(t)\}_{k=1}^{K}$ is a set of known basis functions (e.g., B-splines, FOURIER basis) and $c_{ik}$ are the coefficients specific to the $i$-th observation. The choice of basis and number of components $K$ affects the smoothness and accuracy of the approximation.

# Functional data




Average daily temperature for each day of the year in Canadian weather stations (left – raw data, right – smoothed data (FOURIER basis, $K = 5$)).

# Functional regression for scalar response (FR)

In functional data regression, the goal is to model the relationship between a scalar response variable $Y$ and a functional predictor $X(t)$, where $t$ belongs to a compact interval $\mathcal{T}$. The functional predictor is typically a smooth function, such as a curve or a signal. The standard functional linear model assumes that the response depends linearly on the entire trajectory of $X(t)$:

$$Y = \alpha + \int_{\mathcal{T}} \beta(t)X(t)\,dt + \varepsilon,$$

where $\alpha$ is the intercept, $\beta(t)$ is the regression coefficient function, and $\varepsilon$ is a zero-mean random error. The objective is to estimate $\beta(t)$ based on a sample of observed pairs $(X_i(t), Y_i)$.

# Functional principal components regression (FPCR)

In functional principal components regression, each functional predictor is decomposed as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t),$$

where $\mu(t)$ is the mean function, $\phi_k(t)$ are orthonormal eigenfunctions, and $\xi_{ik}$ are the scores. The scalar response is modeled by regressing on these scores:

$$Y_i = \alpha + \sum_{k=1}^{K} \beta_k \xi_{ik} + \epsilon_i.$$

# Multiple linear regression (LM)

Multiple linear regression models a scalar response $Y_i$ using multiple predictors. The model is written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

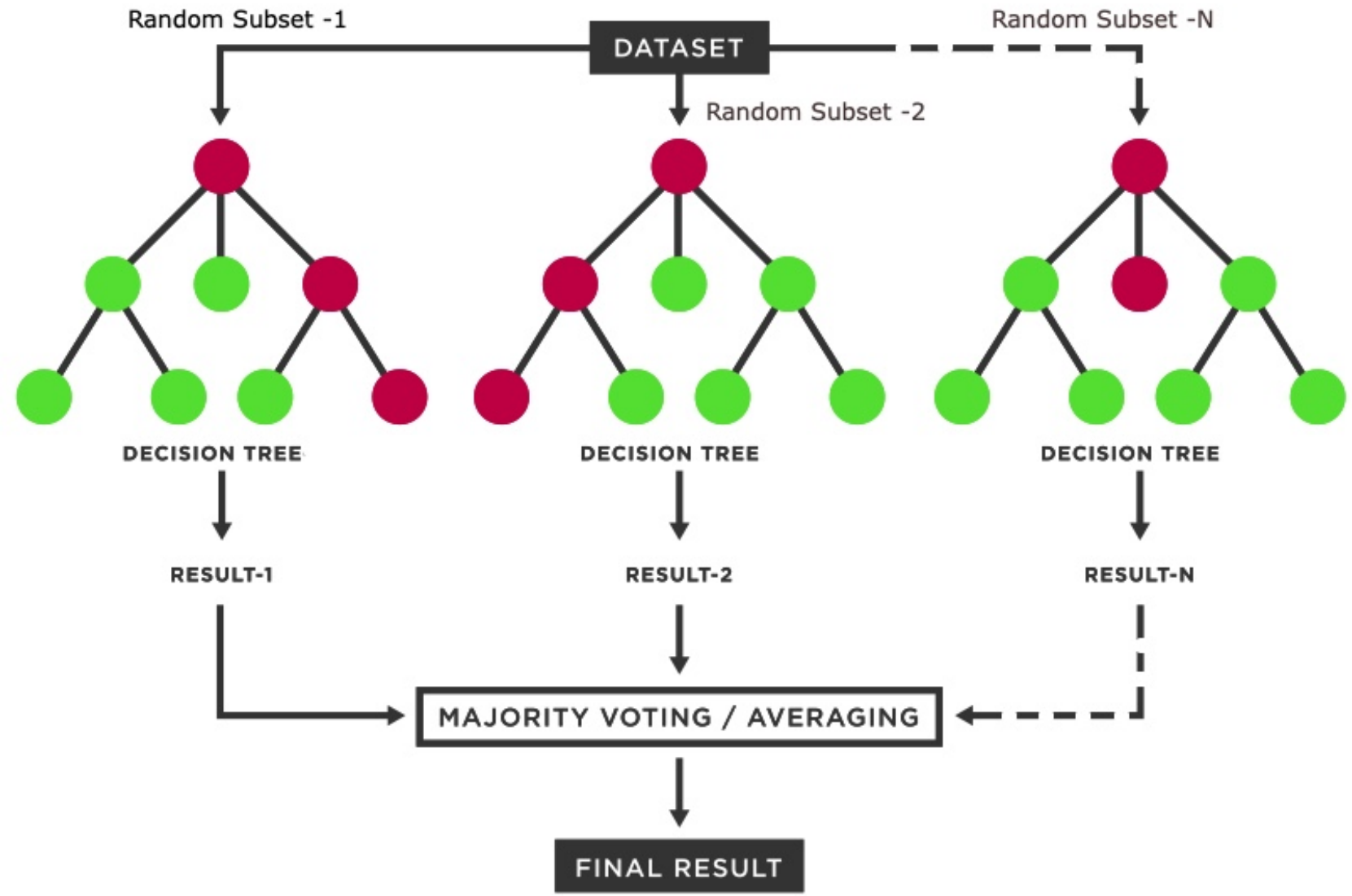
where $Y_i$ is the response for observation $i$, $\beta_0$ is the intercept, $\beta_1, \ldots, \beta_p$ are the coefficients, $x_{i1}, \ldots, x_{ip}$ are the predictor values, and $\varepsilon_i$ is the error term.

# (ordinary) Random forest (RF)

Random Forest is an ensemble learning method for classification and regression that builds a collection of decision trees and combines their predictions. It reduces variance and overfitting by averaging multiple trees trained on random subsets of data and features. The algorithm proceeds as follows:

- For each tree in the forest:
  - Draw a bootstrap sample from the training data.
  - At each node, select a random subset of features.
  - Split the node using the best feature among the subset.
  - Grow the tree to full depth without pruning.
- For regression, the prediction is the average of individual tree outputs.
- For classification, the prediction is made by majority vote.

# (ordinary) Random forest (RF)



# Functional random forest (FRF)

In this model, we apply FPCA for each bootstrap sample before training a tree. In the next step, we construct an ordinary random forest. In fact, we add a small computational overhead for each tree.

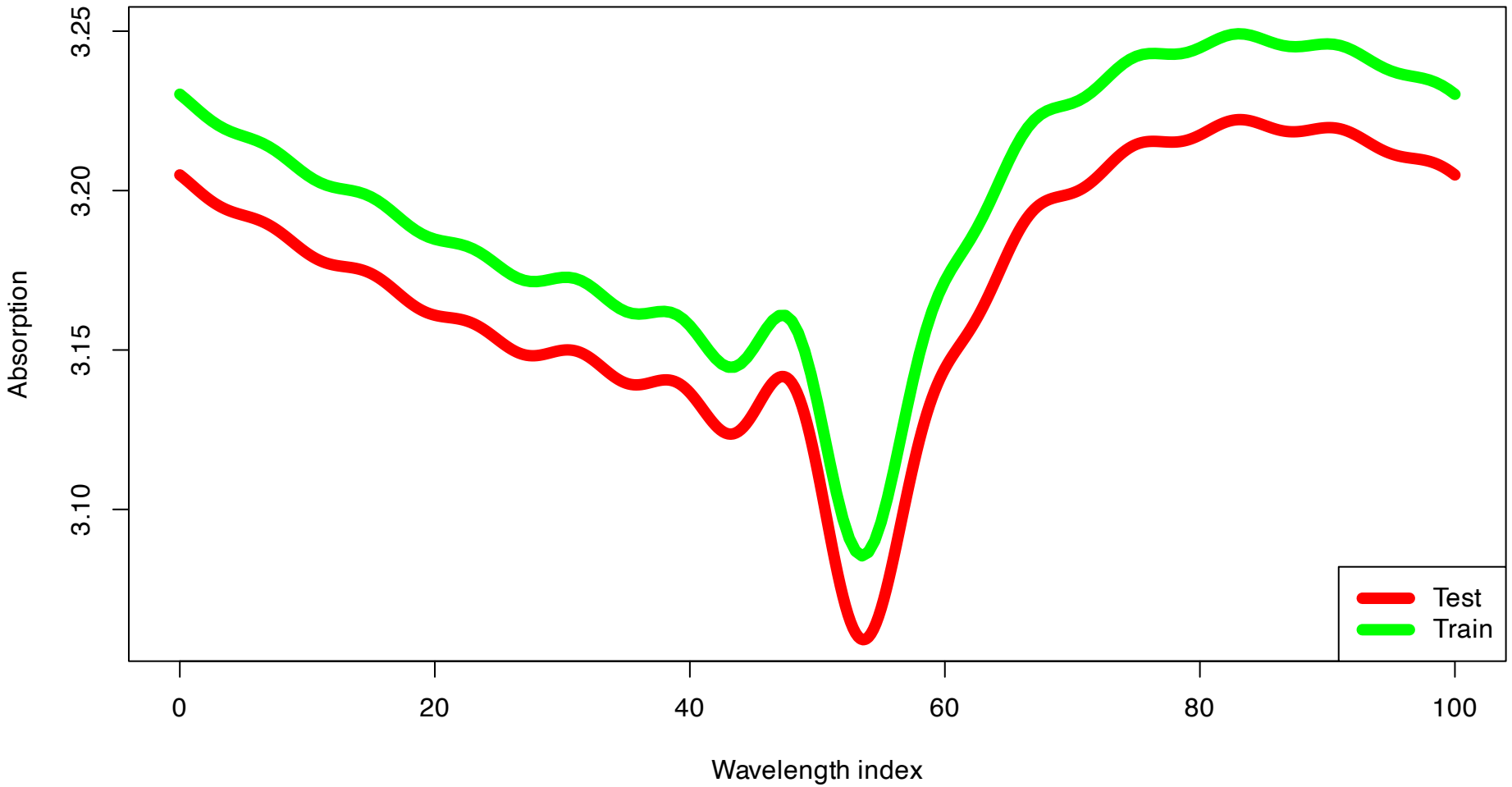
# Full functional random forest (FFRF)

In this model, we apply FPCA for each split in the random forest. It is a much more computationally challenging model because we add computational overhead for each split.

# Datasets

The `fat content` spectrometric data (Ferraty and Vieu, 2006) comprises food samples made of finely chopped pure meat with varying fat levels. Each sample provides a 100-channel spectrum of absorbances, calculated as $-\log 10$ of the measured transmittance, and a fat content value determined by chemical analysis. The objective is to predict fat content from the NIR spectrum. The dataset contains 215 samples, with 175 used for training and 40 reserved for testing. The original dataset is called *Tecator*, and each sample consists of finely chopped pure meat with varying moisture, fat, and protein contents.

# Datasets

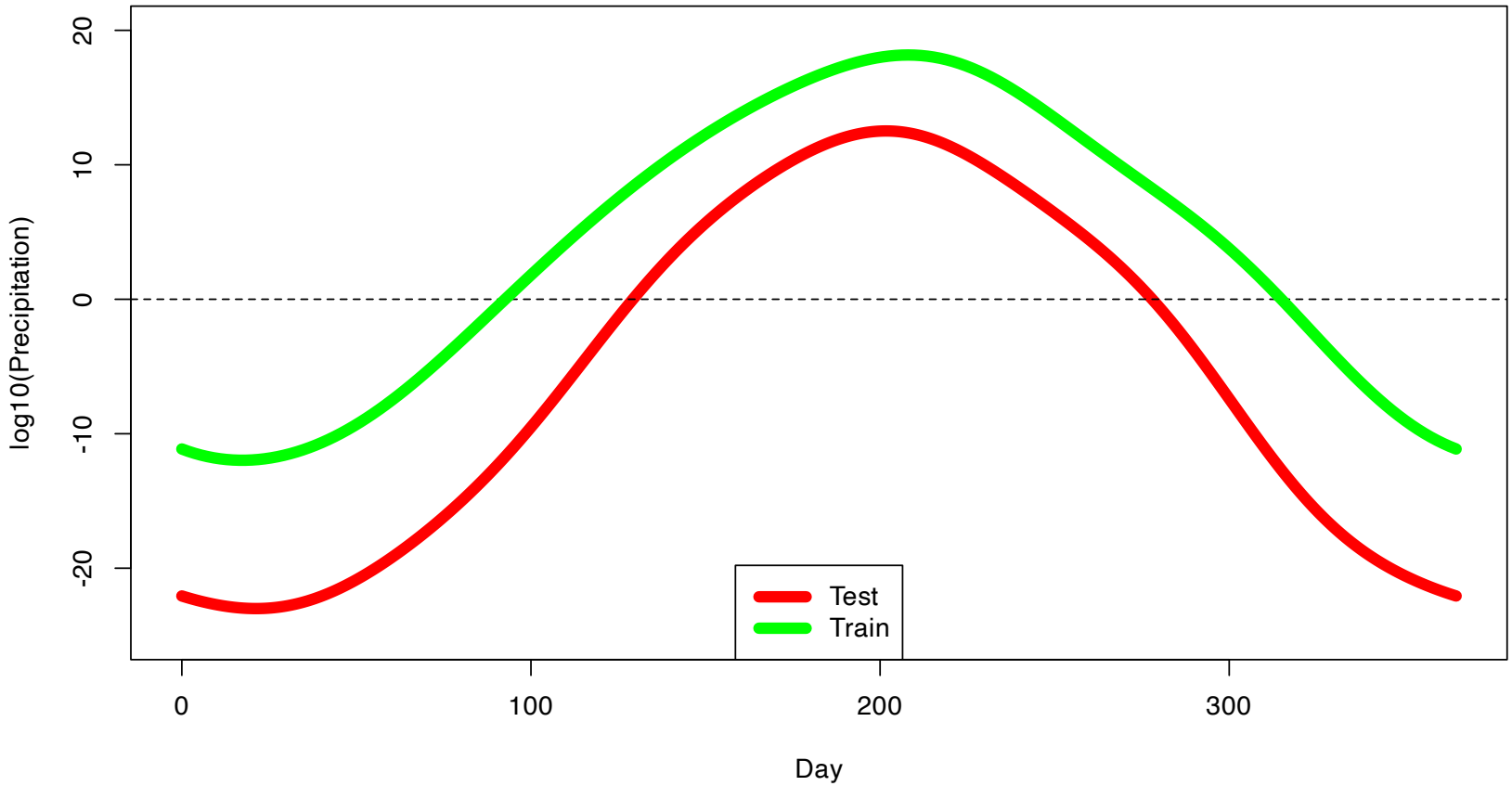



Functional means for `fat content` dataset

# Datasets

The `Canadian weather` dataset contains daily temperature and precipitation data at 35 different locations in Canada, averaged over the period from 1960 to 1994 (Ramsay and Silverman, 2006). We model the decimal logarithm of annual precipitation as a function of the temperature profile. We used 28 stations for training, and seven were reserved for testing.

# Datasets

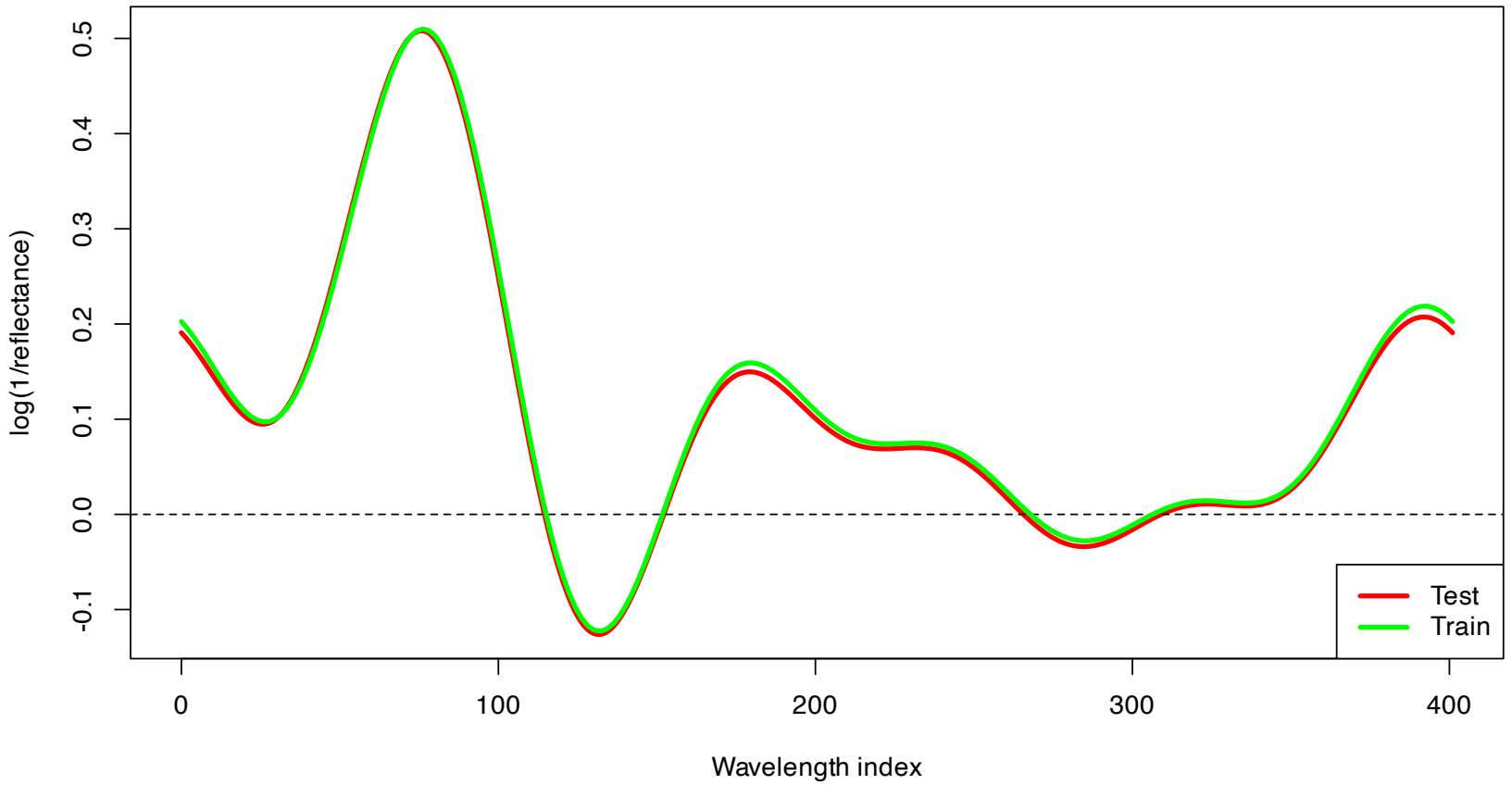



Functional means for `Canadian weather` dataset

# Datasets

Near-infrared reflectance spectra and octane numbers of 60 `gasoline` samples (Kalivas, 1997). Each NIR spectrum consists of log(1/reflectance) measurements at 401 wavelengths, in 2-nm intervals from 900 nm to 1700 nm. We used 48 samples for training and 12 for testing.

# Datasets




Functional means for `gasoline` dataset

# Implementation Details

For calculations, we used R and the following libraries:

- `caret`,
- `fda`,
- `fdapace`,
- `fda.usc`,
- `randomForest`.

# Results

*Table:* RMSE errors on test dataset

| Dataset | LM | FLM | FPCR | RF | FRF | FFRF |
|---|---|---|---|---|---|---|
| Fat content | 4.65 | 2.79 | 2.12 | 8.12 | 1.83 | **1.48** |
| Water content | 4.36 | 2.56 | 2.68 | 6.01 | 2.26 | **2.01** |
| Protein content | **0.87** | 1.12 | 1.22 | 2.15 | 1.14 | 1.02 |
| Canadian weather | 9.07 | 0.44 | 0.35 | 0.32 | 0.34 | **0.31** |
| Gasoline | 2.88 | 0.33 | **0.28** | 0.72 | 0.57 | 0.41 |

# Results

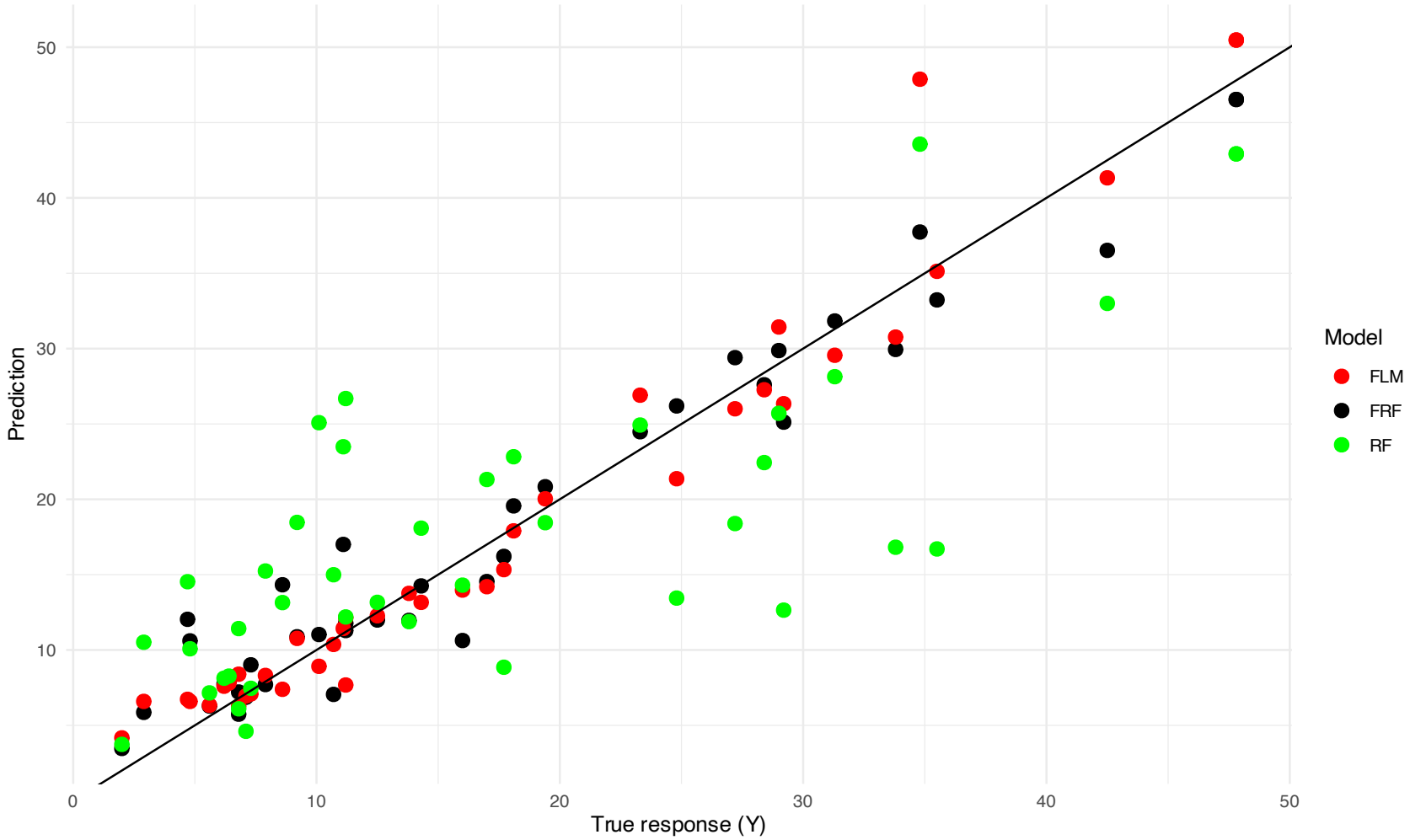



*Figure:* True response vs prediction for selected models for fat content dataset

# Main bibliography


- Breiman, L. (2001). Random forests. *Machine Learning* 45(1):5-32.
- Ferraty, F., Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer-Verlag, Berlin, Heidelberg.
- Kalivas, J.H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems* 37(2):255-259.
- Kokoszka, P., Horváth, L. (2012). *Inference for Functional Data with Applications*. Springer.
- Ramsay, J., Silverman, B. (2006). *Functional Data Analysis. Springer Series in Statistics*. Springer New York.

# Thanks