funHDDC, a new package for Clustering of multivariate functional data

https://cran.r-project.org/web/packages/funHDDC/index.html

Speaker : **Amandine Schmutz**[1,2,3,4]

Directors : Julien Jacques[2], Charles Bouveyron[5], Laurence Chèze[3] & Pauline Martin[1, 4]

1 Lim France, Leather in motion

2 Université Lumière Lyon 2

3 Université Claude Bernard Lyon 1

4 VetLab, EnvA, CWD

5 Université Côte d'Azur

Contents

Context

- **8,3 billion** of **smart devices** [Gartner study (2017)]

Context

- **8,3 billion** of **smart devices** [Gartner study (2017)]
  - ➔ **20,5 billion** from now until 2020

Context

- **8,3 billion** of **smart devices** [Gartner study (2017)]

  → **20,5 billion** from now until 2020
- Market divided between *companies* and *private individuals*

Context

- **8,3 billion** of **smart devices** [Gartner study (2017)]
  - ➔ **20,5 billion** from now until 2020
- Market divided between *companies* and *private individuals*
- Numerous *practical applications*: health, everyday life, hobbies…

Context

- **8,3 billion** of **smart devices** [Gartner study (2017)]
  - ➔ **20,5 billion** from now until 2020
- Market divided between *companies* and *private individuals*
- Numerous *practical applications*: health, everyday life, hobbies…

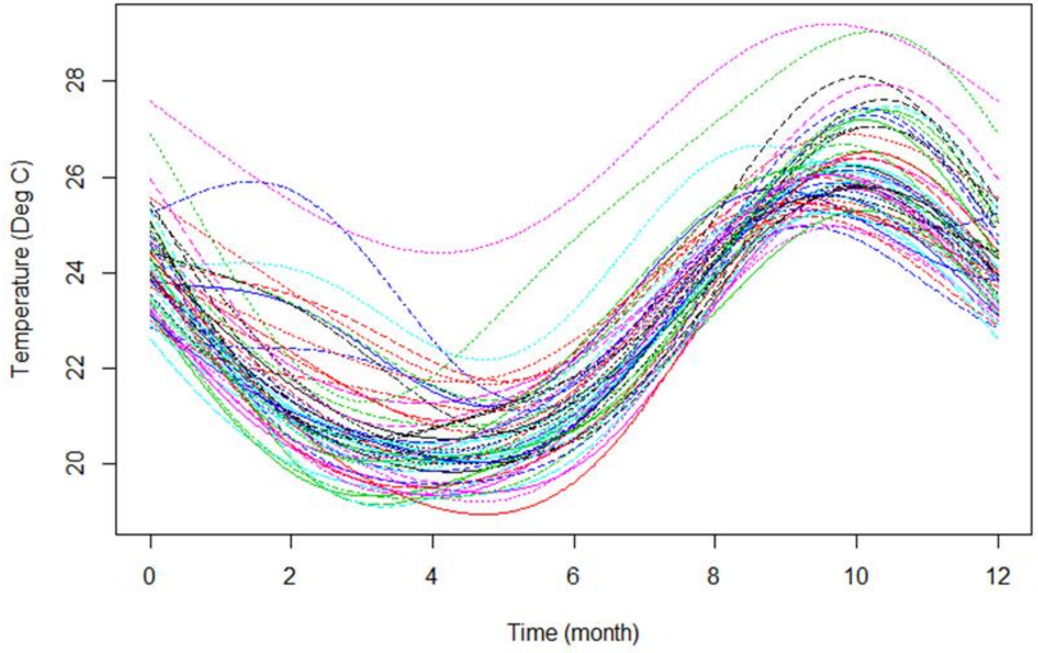
Collection of high frequency data

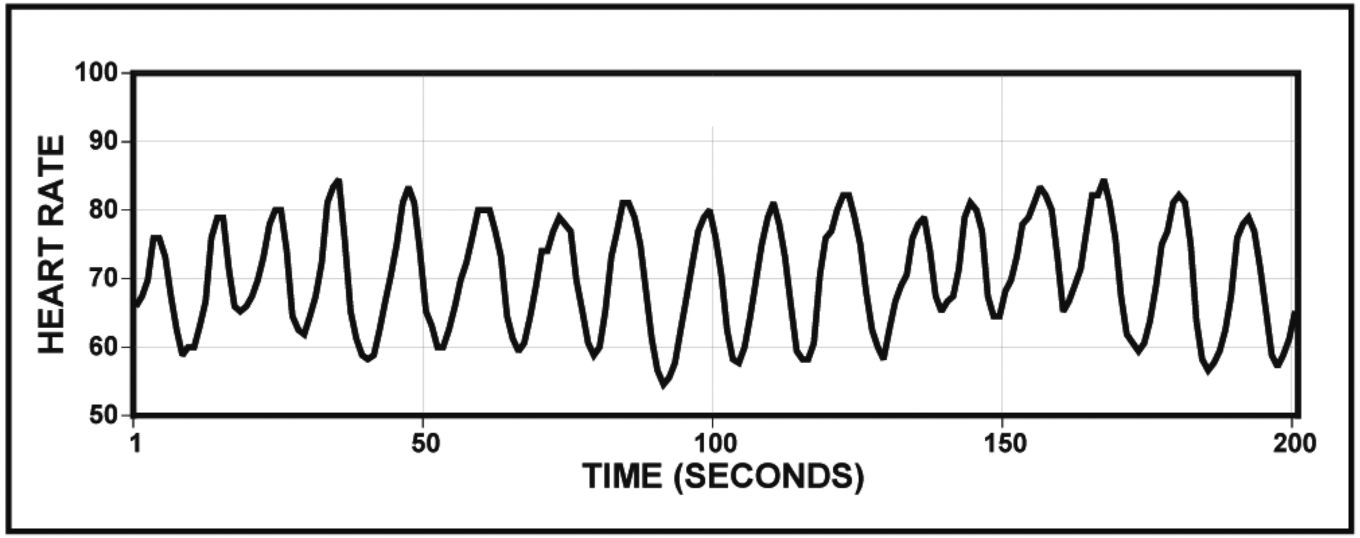
Functional data

Smart devices collect **continuous measurements**

❖Functional data

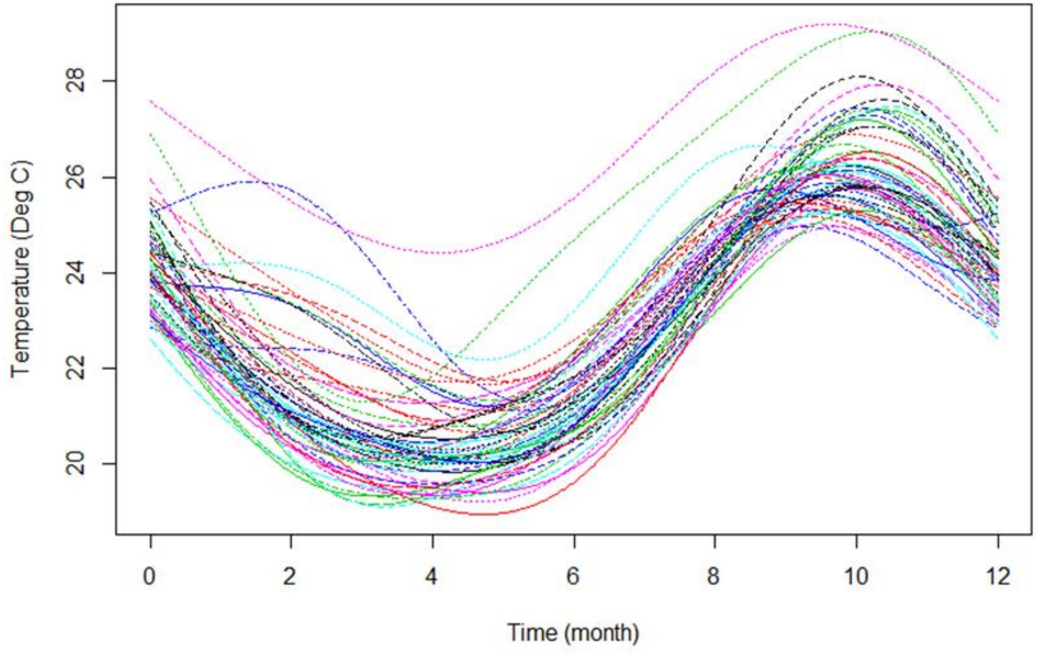
Smart devices collect **continuous measurements**

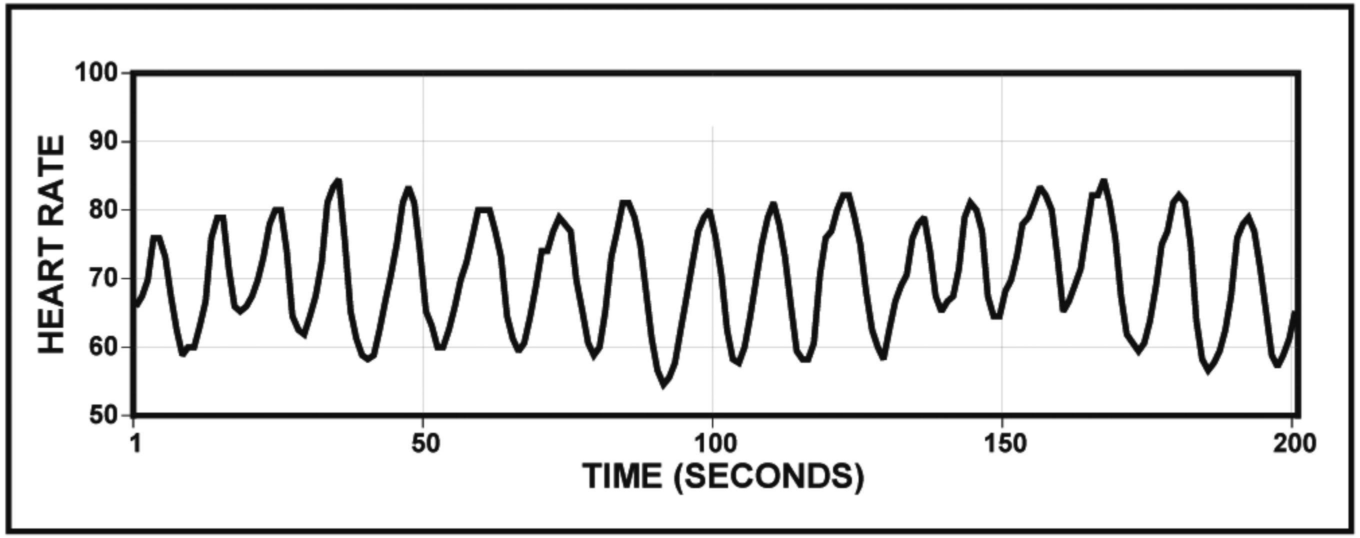




Functional data

Smart devices collect **continuous measurements**

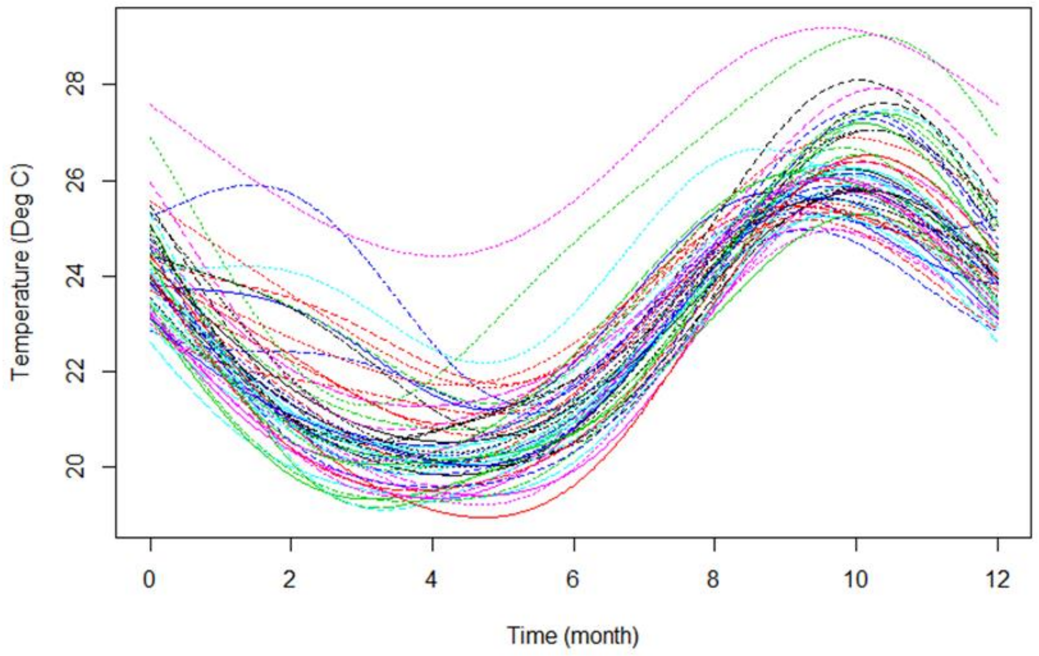


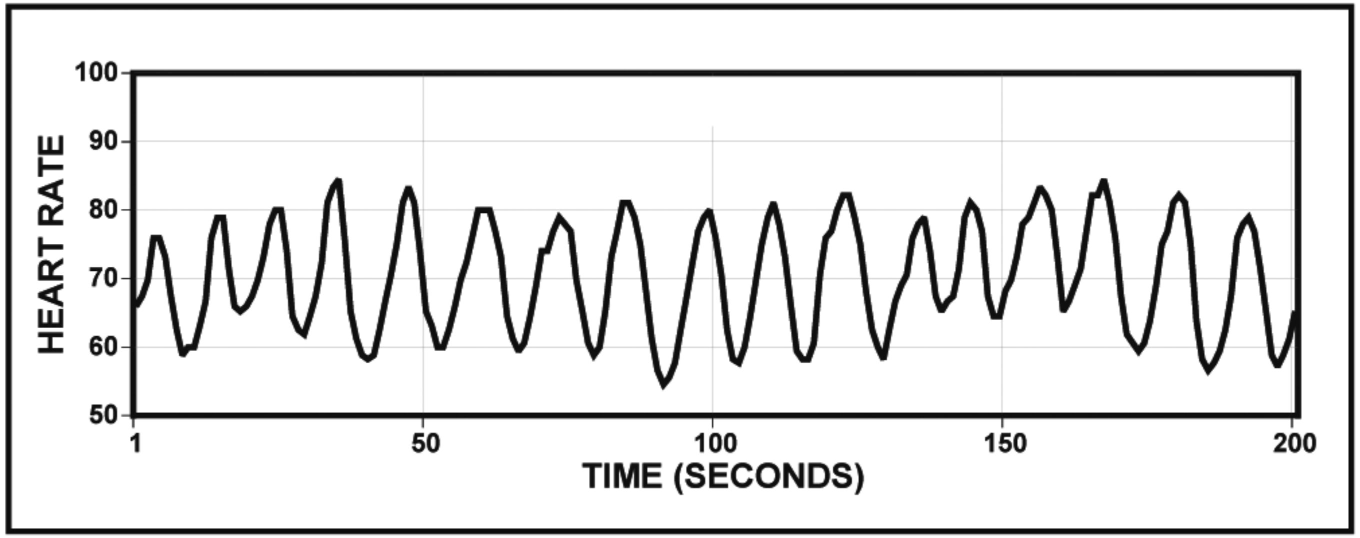


1 individual = 1 curve

Functional data

Smart devices collect **continuous measurements**





1 individual = 1 curve

➔ *Dependency kept* between points

Contents

Smart devices in sport

- Coming up of running smart watches (Garmin, Polar…)

Smart devices in sport

- Coming up of running smart watches (Garmin, Polar…)
- Smart racket for tennis players tennis (Babolat…)

Smart devices in sport

- Coming up of running smart watches (Garmin, Polar…)
- Smart racket for tennis players tennis (Babolat…)
- Cycling, swimming…

Smart devices in sport

- Coming up of running smart watches (Garmin, Polar…)
- Smart racket for tennis players tennis (Babolat…)
- Cycling, swimming…

Lack of smart devices for equestrian sports

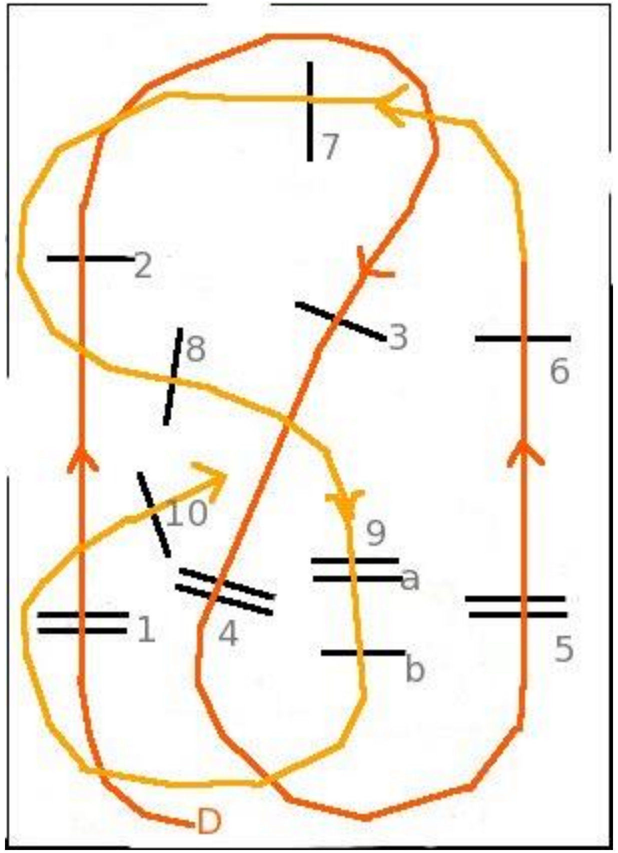
Smart devices in sport

- Coming up of running smart watches (Garmin, Polar…)
- Smart racket for tennis players tennis (Babolat…)
- Cycling, swimming…

Lack of smart devices for equestrian sports

❖Equestrian sports: Jumping WEG 2018 (Tryon USA)

- Timed course, 10-14 obstacles
- Check distances



❖Equestrian sports: Jumping WEG 2018 (Tryon USA)

- Timed course, 10-14 obstacles
- Check distances
- Up to 160 cm high, 450 cm wide

Smart saddle

- Training tool for equestrian sport

Smart saddle

- Training tool for equestrian sport

Accelerometer & gyroscope

Smart saddle

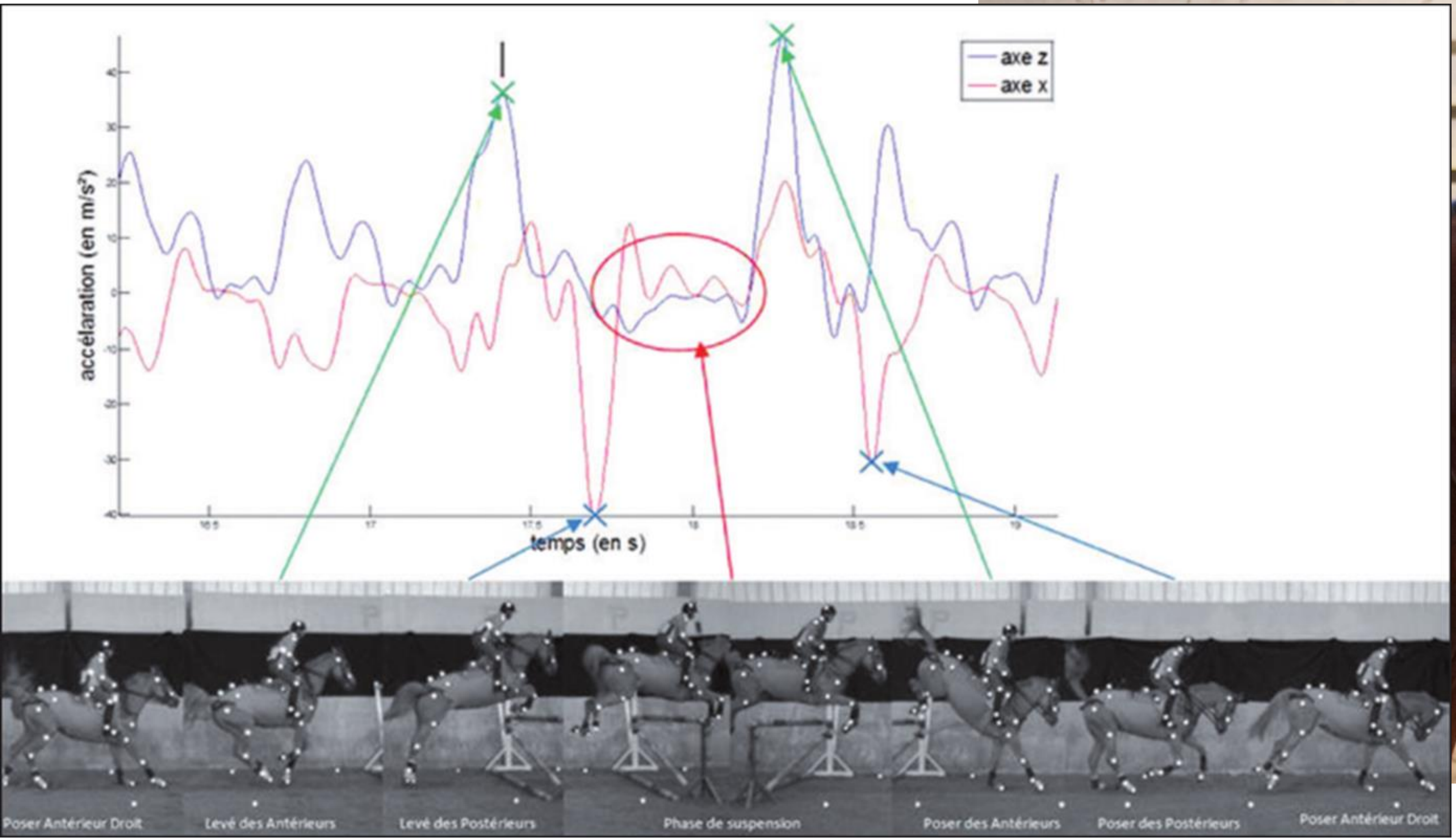
- Training tool for equestrian sport

Accelerometer & gyroscope

Bluetooth® antenna

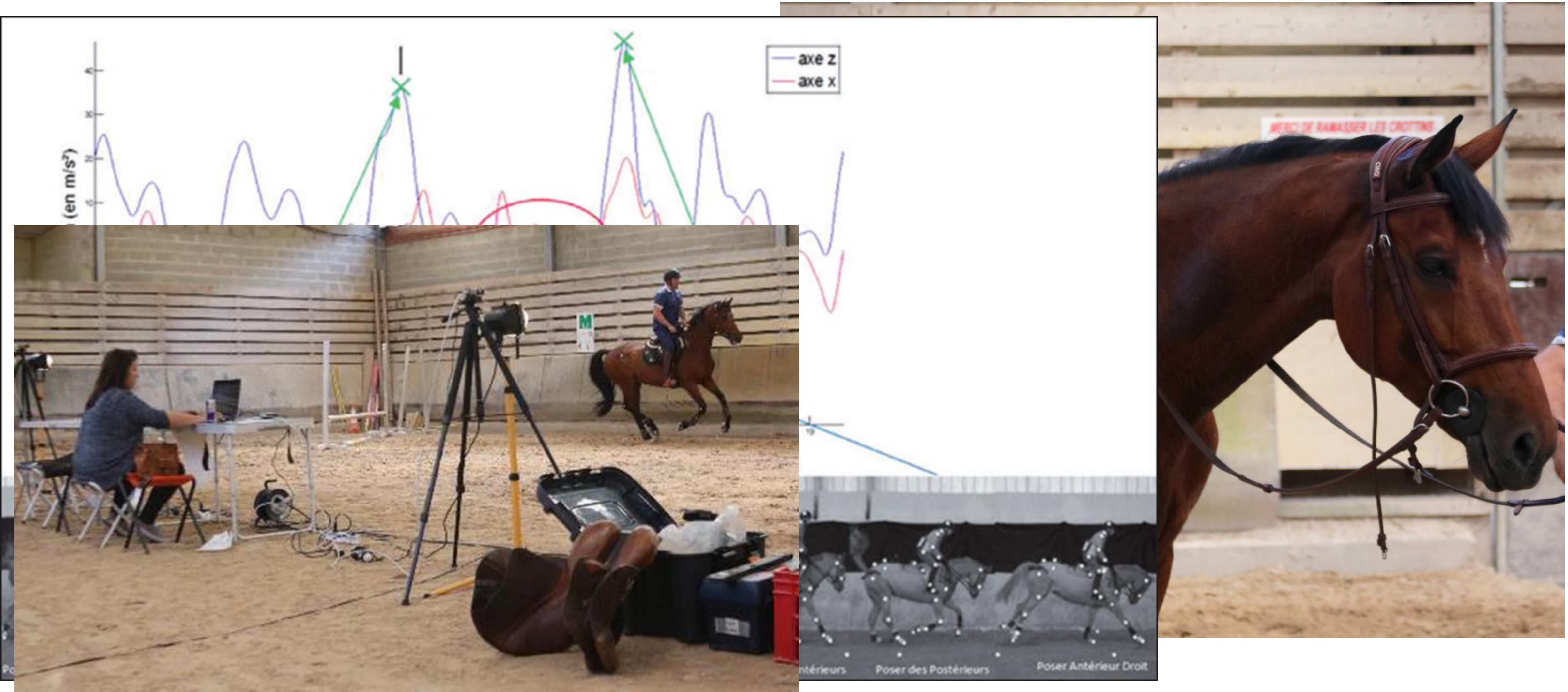
Smart saddle

- Training tool for equestrian sport



Smart saddle

- Training tool for equestrian sport



Contents

Clustering method

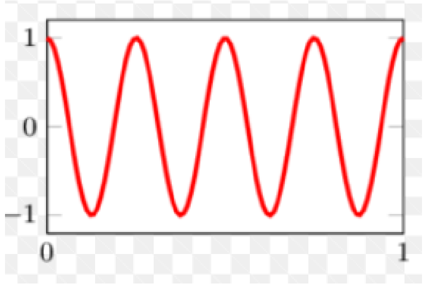
Objective : **Separate n p-variate curves in K clusters**

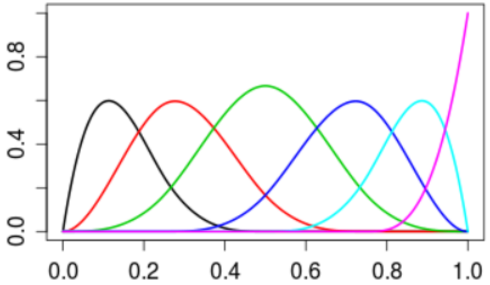
Clustering method

Objective : **Separate n p-variate curves in K clusters**

- Expression in a basis of functions

$$X_i^j(t) = \sum_{r=1}^{R_j} c_{ir}^j \Phi_r^j(t)$$





Clustering method

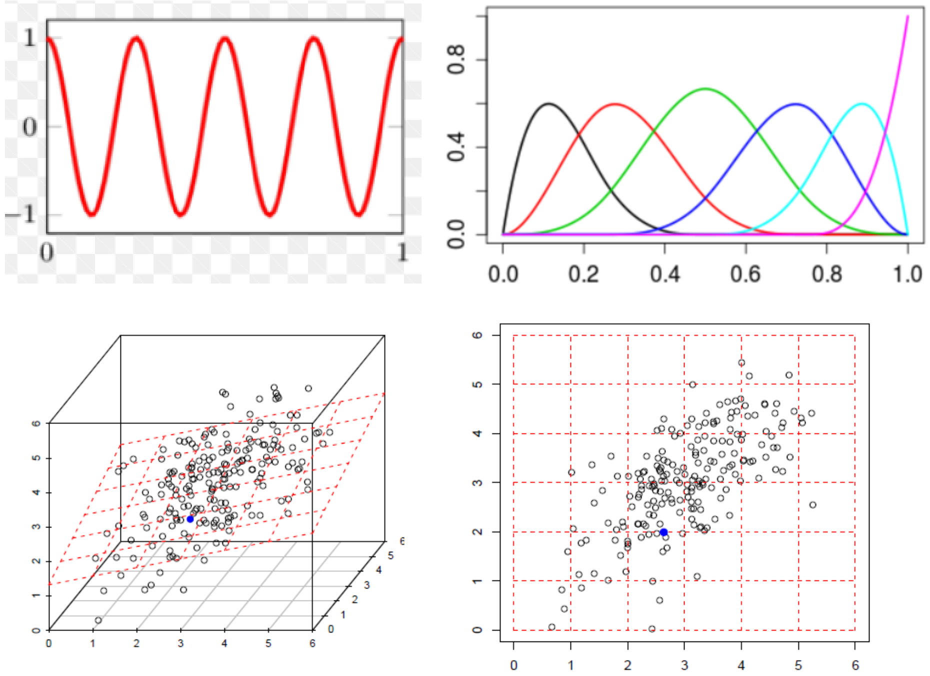
Objective : **Separate n p-variate curves in K clusters**

- Expression in a basis of functions

$$X_i^j(t) = \sum_{r=1}^{R_j} c_{ir}^j \Phi_r^j(t)$$

- Curves projections

$$X_i(t) = \mu_k(t) + \sum_{j=1}^{R} \delta_k \psi_{kj}(t)$$

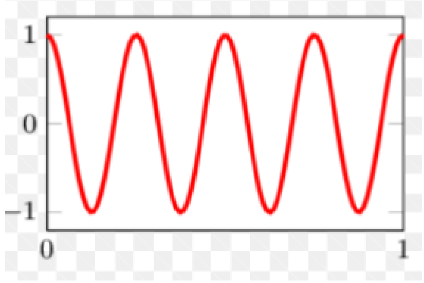


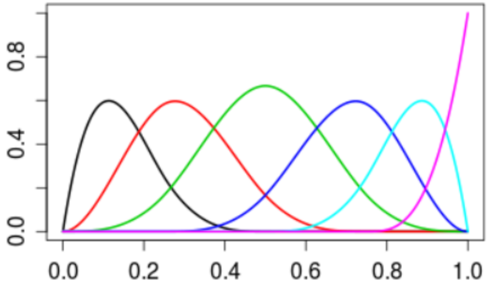
Clustering method

Objective : **Separate n p-variate curves in K clusters**

- Expression in a basis of functions

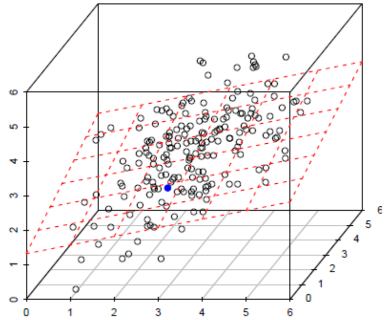
$$X_i^j(t) = \sum_{r=1}^{R_j} c_{ir}^j \Phi_r^j(t)$$

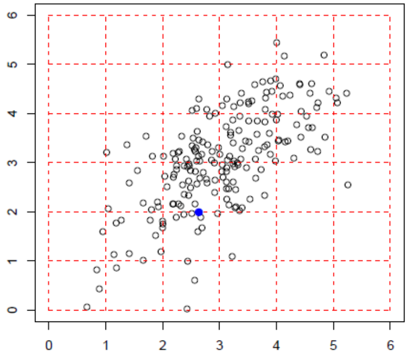




- Curves projections

$$X_i(t) = \mu_k(t) + \sum_{j=1}^{R} \delta_k \psi_{kj}(t)$$





- Mixture model

$$p(\delta) = \sum_{k=1}^{K} \pi_k \, N(\delta; \mu_k, \Delta_k)$$

R package funHDDC

funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces

```
funHDDC(data, K, init, …)
```

R package funHDDC

funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces

```
funHDDC(data, K, init, …)
```

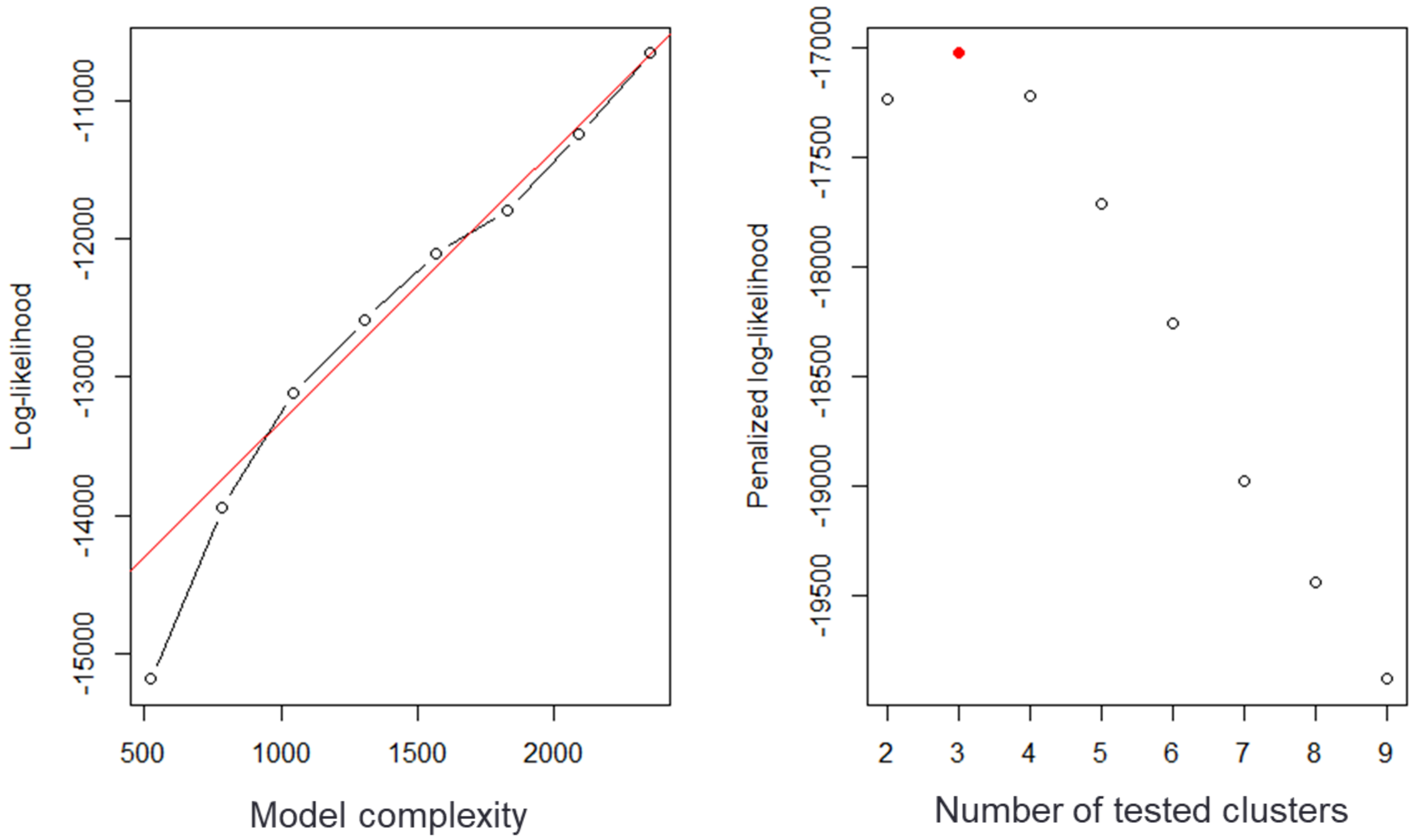
```
slopeheuristic(mod)
```

R package funHDDC

`funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces`

```
funHDDC(data, K, init, …)
```

```
slopeheuristic(mod)
```



R package funHDDC

funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces

```
funHDDC(data, K, init, …)
```

```
slopeheuristic(mod)
```

```
mfpca(data)
```

R package funHDDC

funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces

```
funHDDC(data, K, init, …)

slopeheuristic(mod)

mfpca(data)

plot.mfpca(x, nharm, threshold)
```

R package funHDDC

funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces

```
funHDDC(data, K, init, …)

slopeheuristic(mod)

mfpca(data)

plot.mfpca(x, nharm, threshold)

predict(mod, newdata)
```

Contents

Prediction of the horse speed

- Strides **clustering**: `funHDDC(list(az,gy),K=2)`

Prediction of the horse speed

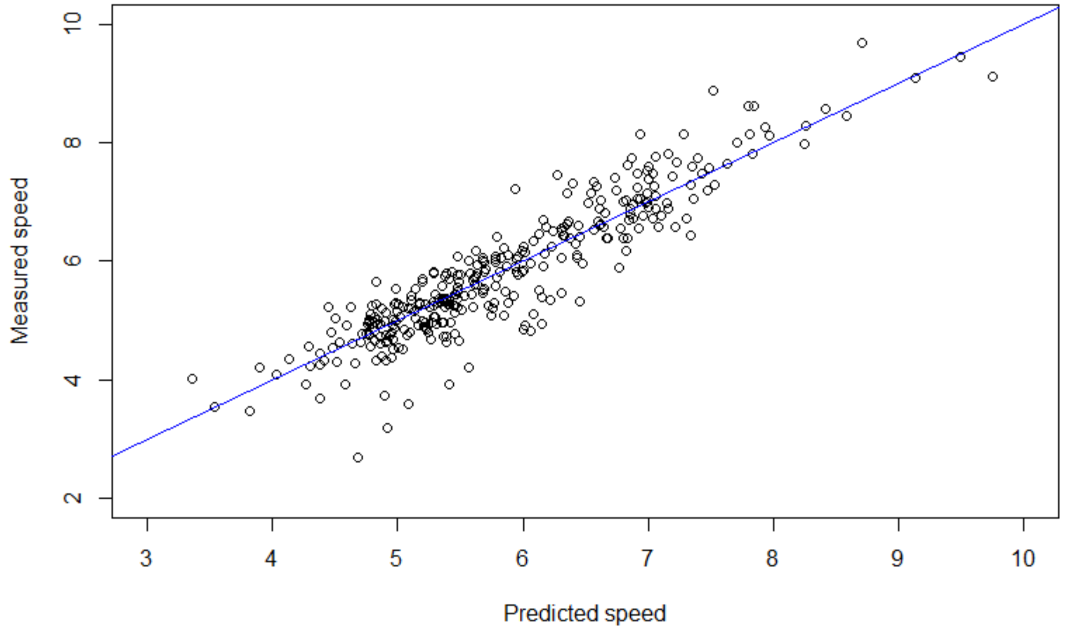
- Strides **clustering**: `funHDDC(list(az,gy),K=2)`
- **SVM** per cluster for speed prediction

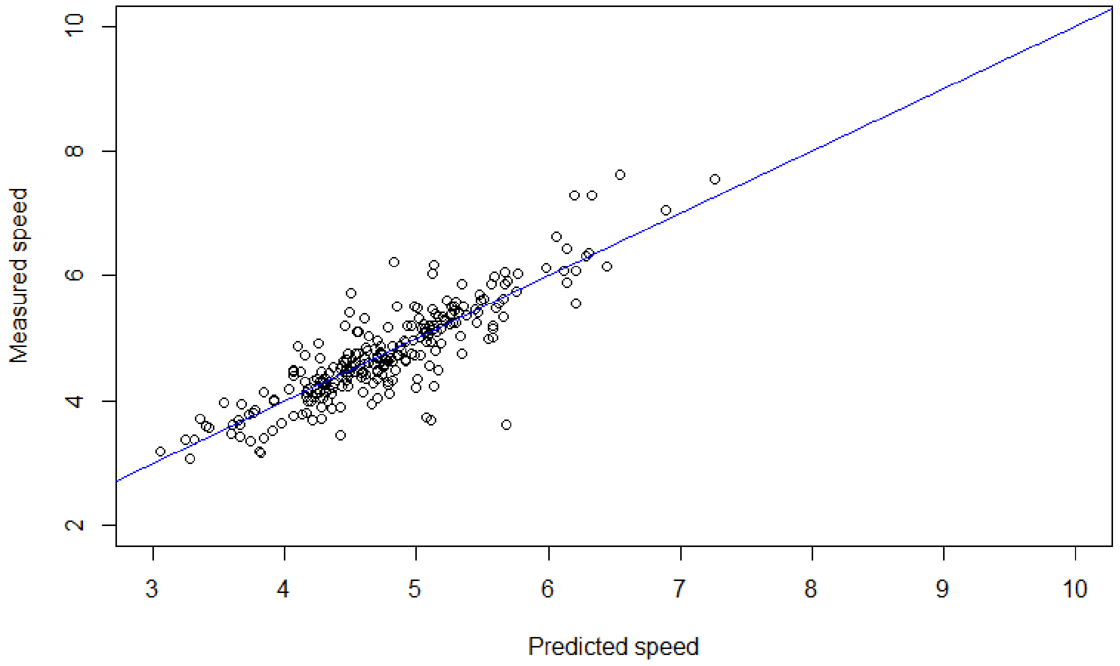
Prediction of the horse speed

- Strides **clustering**: `funHDDC(list(az,gy),K=2)`
- **SVM** per cluster for speed prediction
- Computation of the percentage of error *above 0,6 m/s*
  - Training dataset (80%), Test dataset (20%)

Prediction of the horse speed

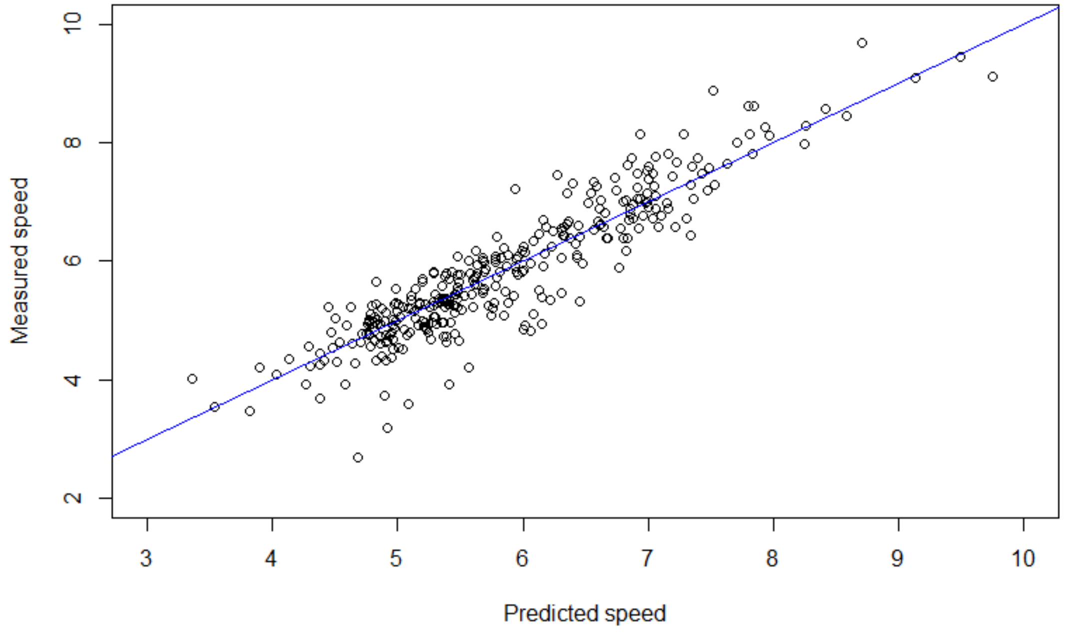
- Strides **clustering**: `funHDDC(list(az,gy),K=2)`
- **SVM** per cluster for speed prediction
- Computation of the percentage of error *above 0,6 m/s*
  - Training dataset (80%), Test dataset (20%)

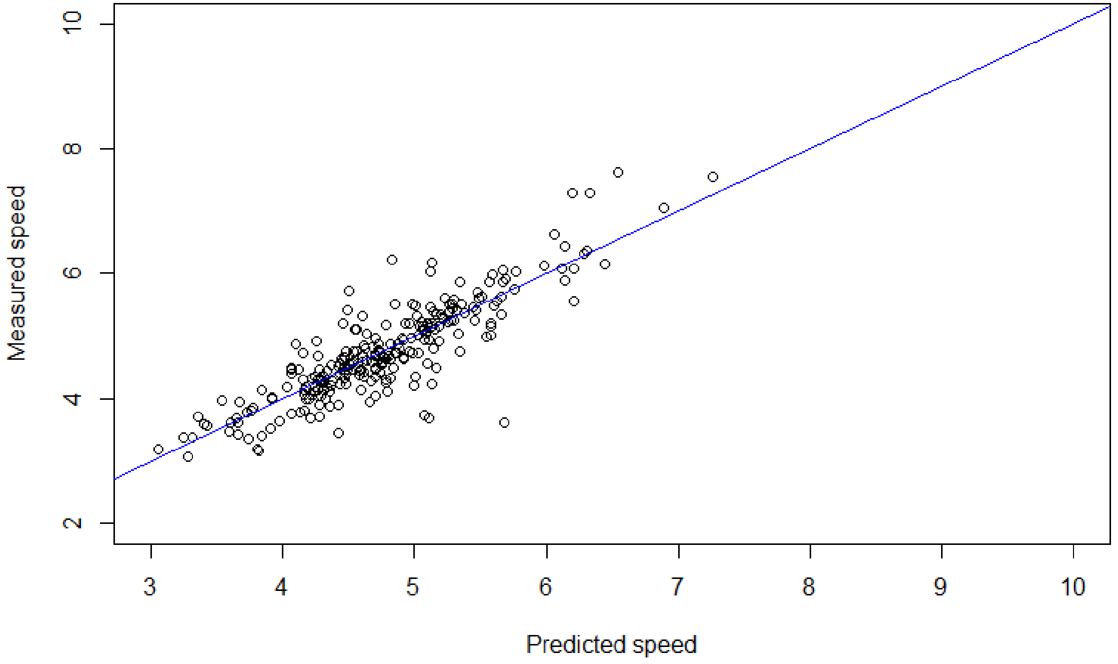




Prediction of the horse speed

- Strides **clustering**: `funHDDC(list(az,gy),K=2)`
- **SVM** per cluster for speed prediction
- Computation of the percentage of error *above 0,6 m/s*
  - Training dataset (80%), Test dataset (20%)





11,6% of errors above 0,6 m/s

Automation in a smartphone app

Objective: **Automate** calculations to provide a **tool** to help riders for their **training**

❖Automation in a smartphone app

Objective: **Automate** calculations to provide a **tool** to help riders for their **training**

- Use of *predict* function:

Automation in a smartphone app

Objective: **Automate** calculations to provide a **tool** to help riders for their **training**

- Use of *predict* function:

```
model←funHDDC(list(az_tot,gy_tot),K=2,model='AkjBkQkDk')
```

Automation in a smartphone app

Objective: **Automate** calculations to provide a **tool** to help riders for their **training**

- Use of *predict* function:

```
model←funHDDC(list(az_tot,gy_tot),K=2,model=‘AkjBkQkDk’)
prediction← predict(model,list(new_az,new_gy))
```

Automation in a smartphone app

Objective: **Automate** calculations to provide a **tool** to help riders for their **training**

- Use of *predict* function:

```
model←funHDDC(list(az_tot,gy_tot),K=2,model='AkjBkQkDk')
prediction← predict(model,list(new_az,new_gy))
```

- **SVM** per cluster for speed prediction

Weather stations Canada

- 35 cities distributed on all territory
- Temperature and pluviometry for 1 year

Weather stations Canada

- 35 cities distributed on all territory
- Temperature and pluviometry for 1 year

```
res1← funHDDC(list(temp,pluvio), K=2:8,
      model=‘AkjBkQkDk’)
```

Weather stations Canada

- 35 cities distributed on all territory
- Temperature and pluviometry for 1 year

```
res1← funHDDC(list(temp,pluvio), K=2:8,
model=‘AkjBkQkDk’)
```

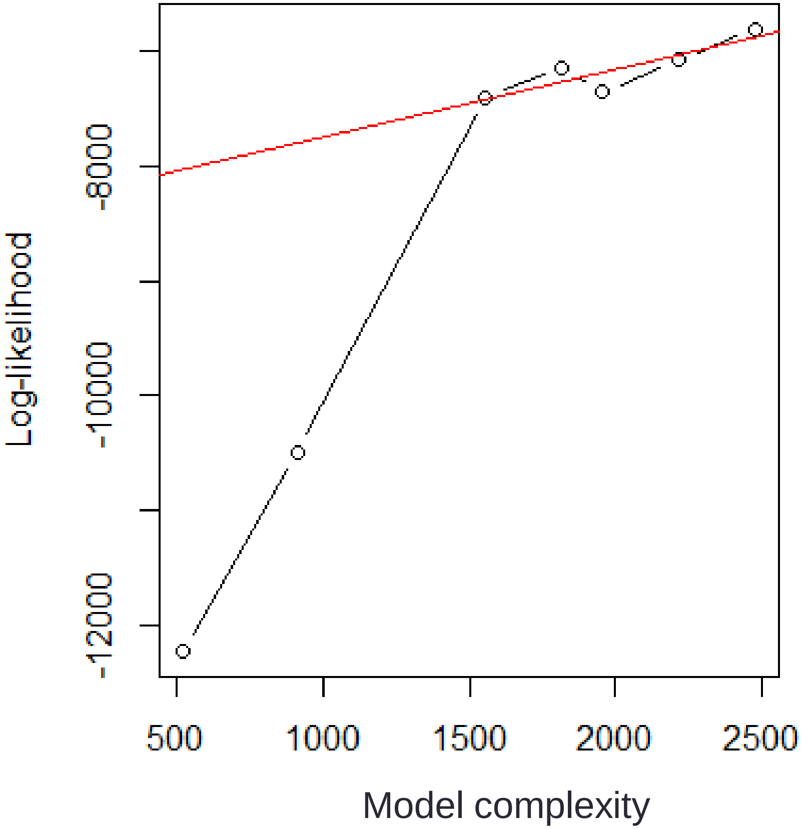
```
slopeheuristic(res1)
```

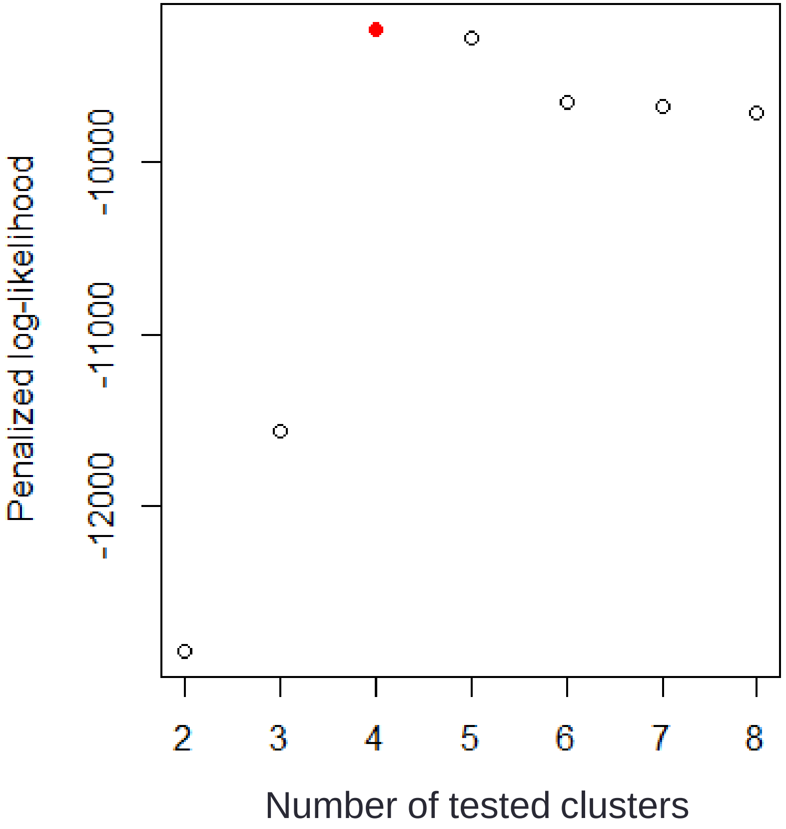
Weather stations Canada

- 35 cities distributed on all territory
- Temperature and pluviometry for 1 year

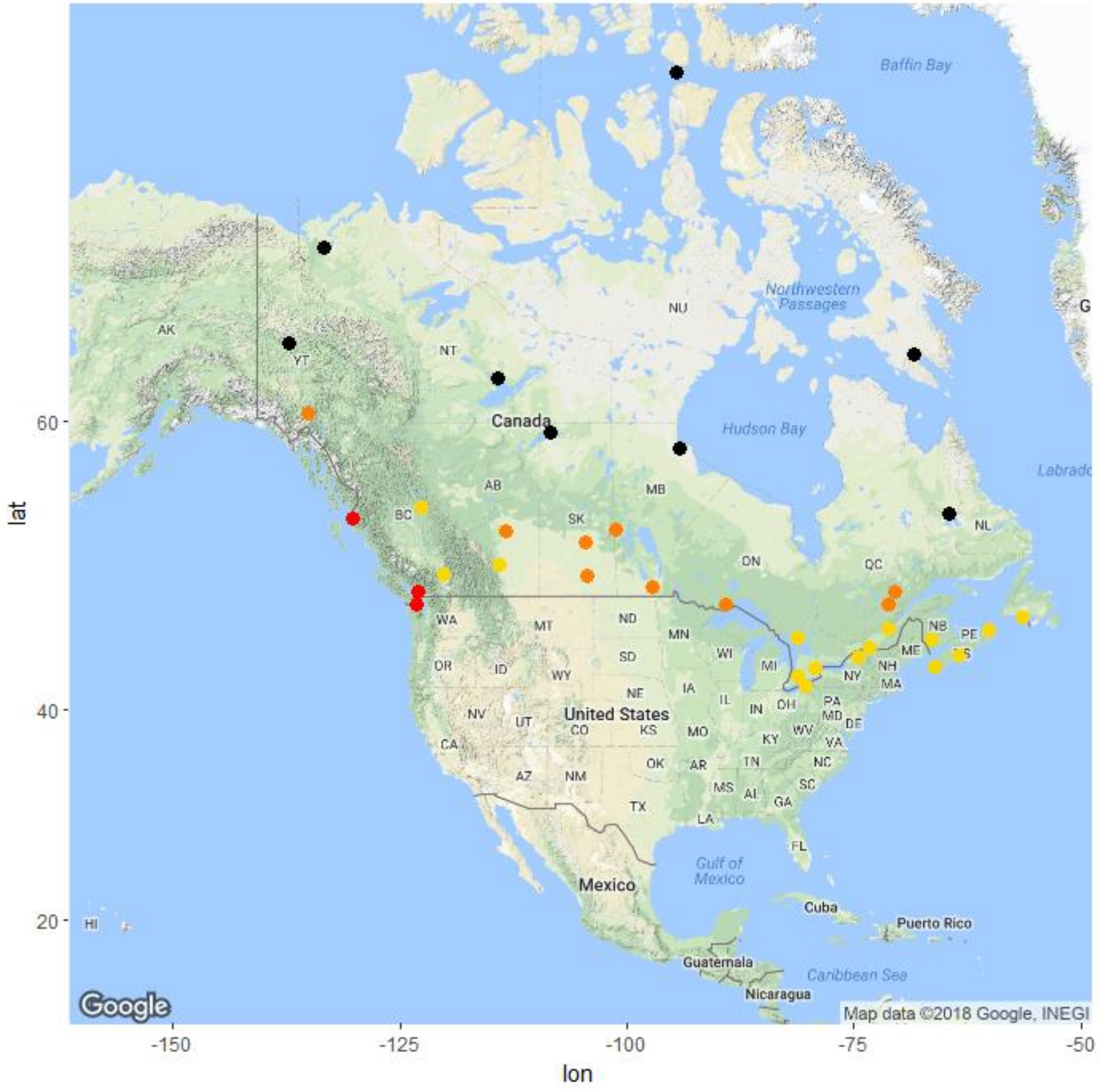
```
res1← fur

slopeheurist
```

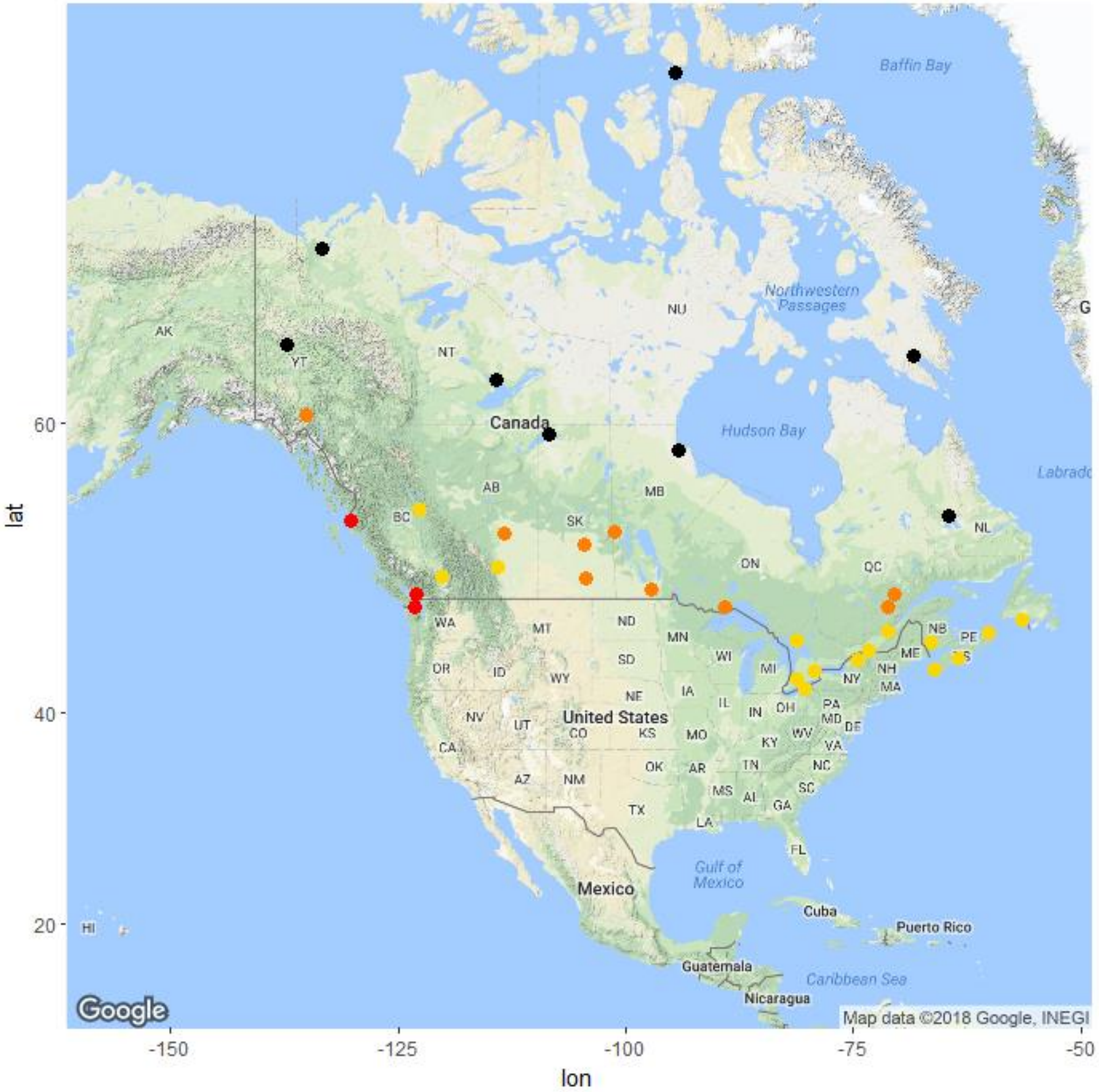


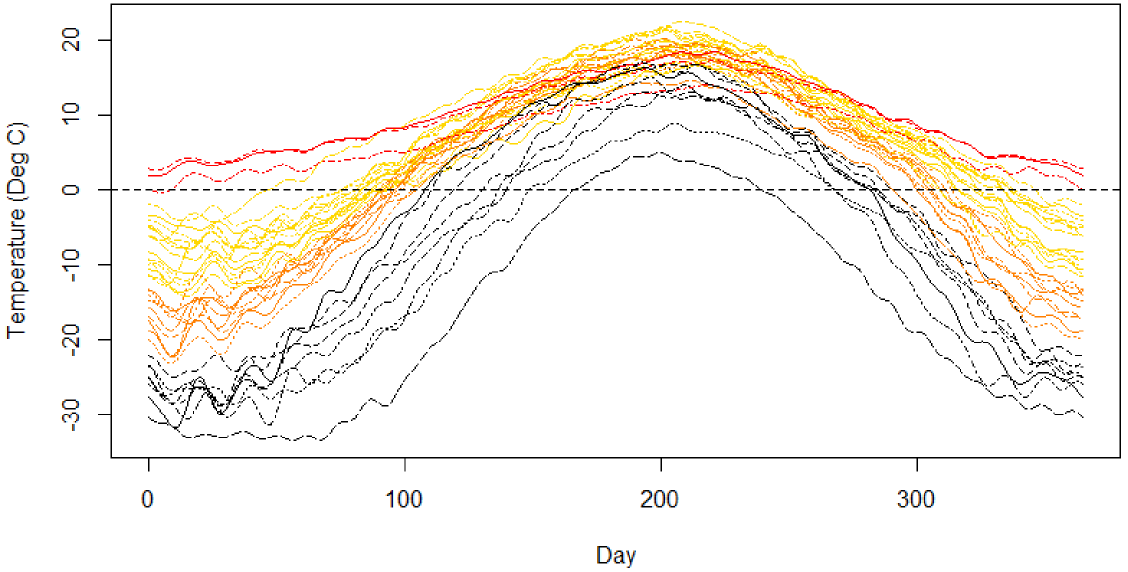


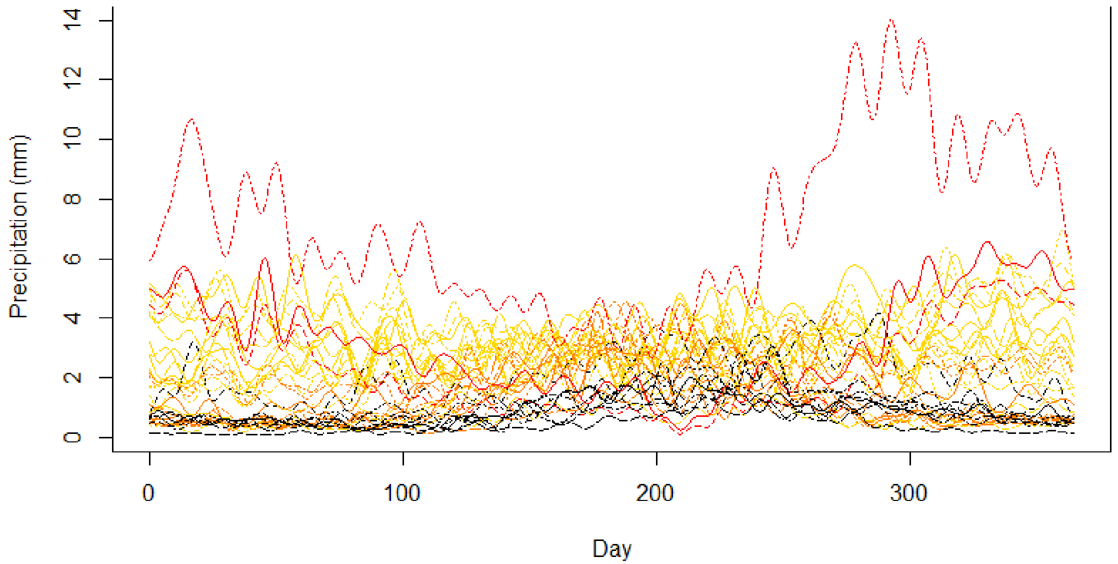
Weather stations Canada



Weather stations Canada







Weather stations Canada

- Main sources of variation

```
res.pca ← mfpca(list(daytempfd,dayprecfd))
```

Weather stations Canada

- Main sources of variation

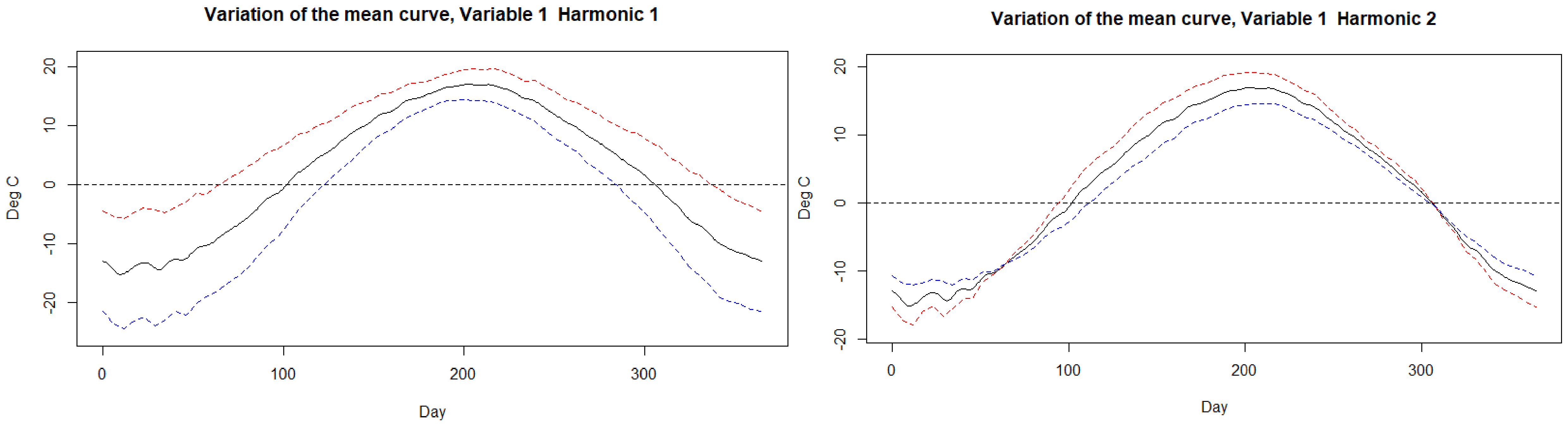
```
res.pca ← mfpca(list(daytempfd,dayprecfd))
plot(res.pca)
```

Weather stations Canada

- Main sources of variation

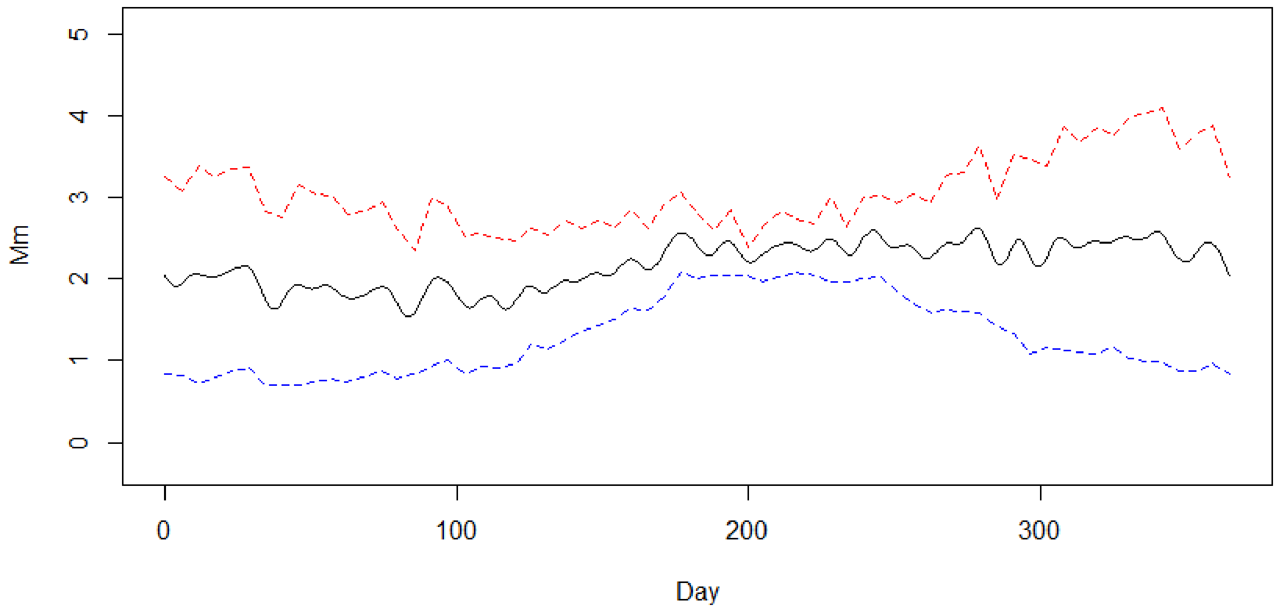
```
res.pca ← mfpca(list(daytempfd,dayprecfd))
plot(res.pca)
```

Temperature variation



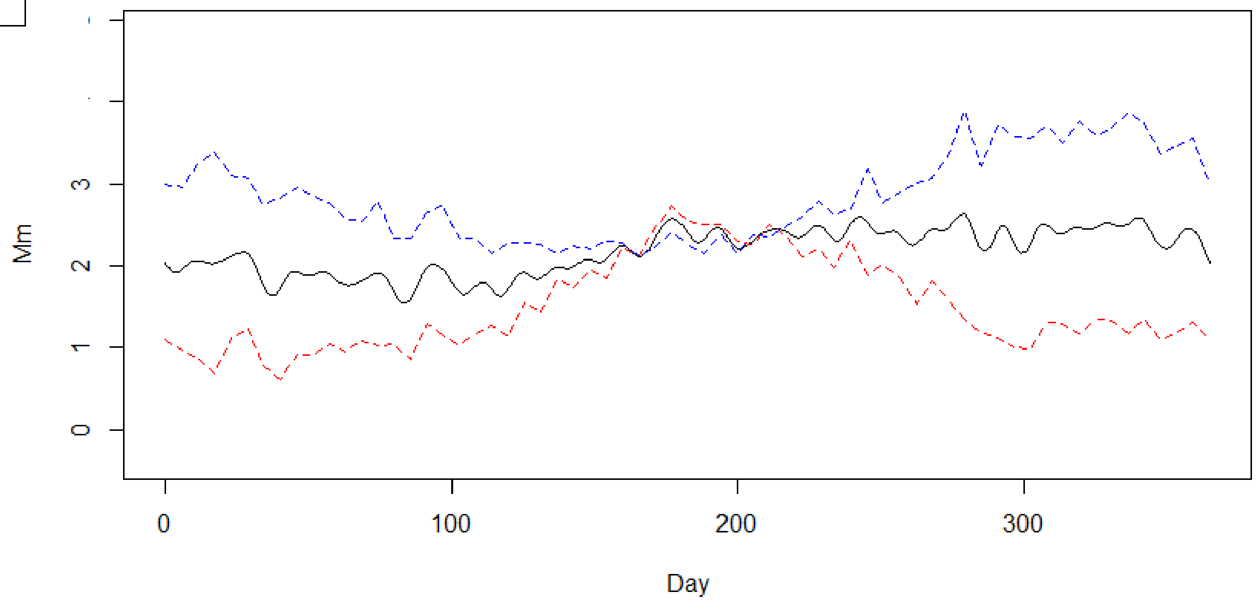
Weather stations Canada

Variation of the mean curve, Variable 2 Harmonic 1



Pluviometry variation

Variation of the mean curve, Variable 2 Harmonic 2



→ Amplitude variation

Contents

Conclusion

- **New model** which allows *univariate* and *multivariate* functional clustering
  - Paper & Simulations available on HAL

- **Designing** an R package available on:

  https://cran.r-project.org/web/packages/funHDDC/index.html

- Coming:

Extension to co-clustering