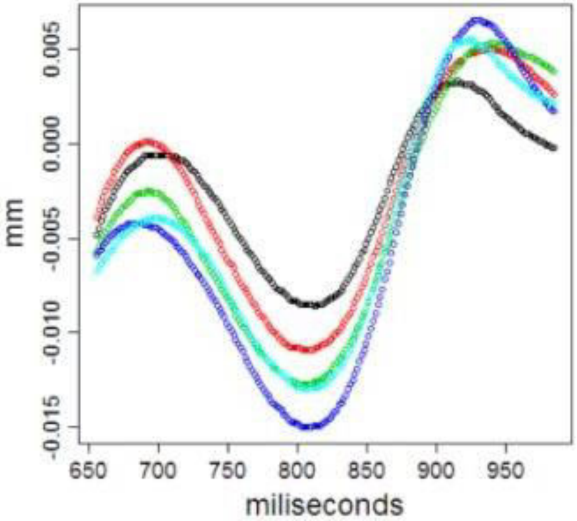
Model selection with functional data analysis.

Ba Amady
Dan Coviello
Ahmad Jaw

Functional data

What is functional data?

What are the most obvious features of these data?

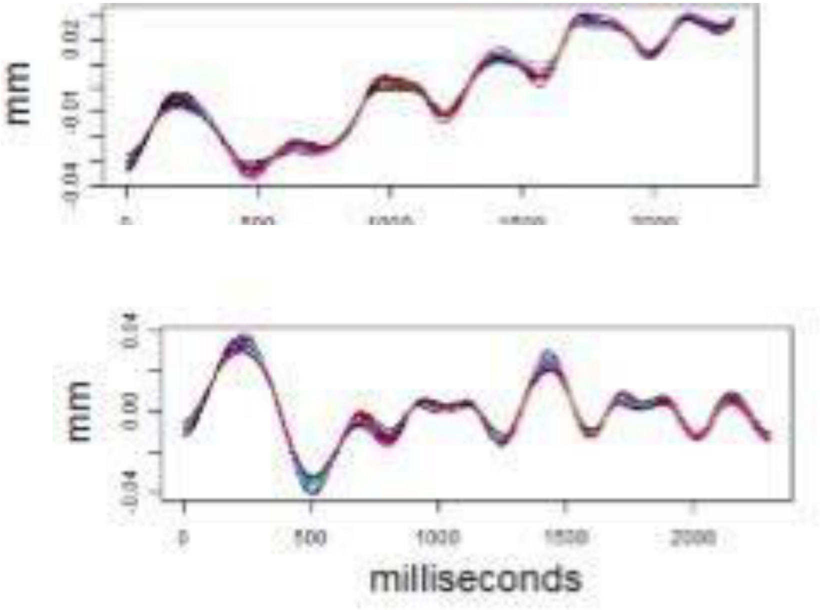


- quantity
- frequency (resolution)
- similarity
- smoothness

Example

20 replication 1401 observations with replication 2 dimensions

Immediate charisteristic



- High-frequency measurements
- Smooth, but complex, processes
- Repeated observations
- Multiple dimensions
- Let's plot 'y' against 'x

What is functional data?

Functional data is multivariate data with an ordering on the dimensions. (Müller, (2006))

Key assumption is *smoothness*:

$$y_{ij} = x_i(t_{ij}) + \epsilon_{ij}$$

with t in a continuum (usually time), and $x_i(t)$ smooth

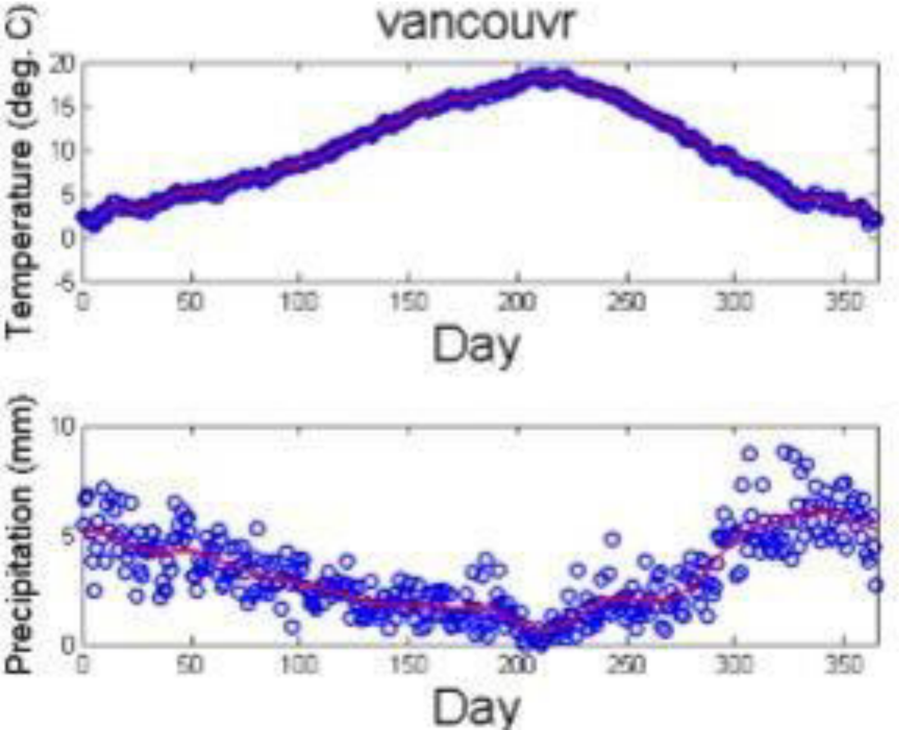
Functional data = the functions $x_i(t)$.

Highest quality data from monitoring equipment

- Optical tracking equipment (eg handwriting data, but also for physiology, motor control,...)
- Electrical measurements (EKG, EEG and others)
- Spectral measurements (astronomy, materials sciences)

Weather in Vancouver

Measure of climate daily precipitation and temperature in Vancouver BC averaged over 40 years



Model: Functional Reconstruction

The objective is to cluster $\{x_1, ..., x_n\}$ into K homogeneous clusters

Assuming that $\{x_1, ..., x_n\}$ are independent realization of $X = \{X(t)\}_{t\in[0,T]}$

Problem: we have only access to discrete observations

$x_{ij} = x_i(t_{is})$ at time $\{t_{is} : s = 1, \ldots, m_i\}$

Solution: smooth with a set of predefined basis $\{\psi_1, \ldots, \psi_p\}$

$$X(t) = \sum_{j=1}^{p} \gamma_j(X)\psi_j(t) \qquad x_i(t) = \sum_{j=1}^{p} \gamma_{ij}\psi_j(t)$$

Model

Assume there exists $Z = (Z_1, \ldots, Z_K) \in \{0,1\}^K$ we want to determine $z_i = (z_{i1}, \ldots, z_{iK})$

Let $F[0,T]$ be a latent subspace of $L_2[0,T]$ spanned by d basis functions, with $d < p$ and $d < K$

The $\{\varphi_j\}_{j=1,\ldots,d}$ is obtained from $\{\psi_j\}_{j=1,\ldots,p}$ through $\varphi_j = \sum_{\ell=1}^{p} u_{j\ell}\psi_\ell$ with $U = (u_{j\ell})$ orthogonal

Let $\{\lambda_1, \ldots, \lambda_n\}$ be the latent expansion coefficients on the bases $\{\varphi_j\}_{j=1,\ldots,d}$ and independent realizations of $\Lambda \in \mathbb{R}^d$

$$\boxed{\Gamma = U\Lambda + \varepsilon}$$

From the smoothing: $X(t) = \sum_{j=1}^{p} \gamma_j(X)\psi_j(t)$

Model (II)

Distributional assumptions

$$\Lambda_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k) \qquad \varepsilon \sim \mathcal{N}(0, \Xi)$$

The Marginal distribution of Γ is a mixture of Gaussians

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma; U\mu_k, U^t \Sigma_k U + \Xi)$$

Noise covariance matrix Ξ

$$\Delta_k = \mathrm{cov}(W^t \Gamma | Z = k)$$

$$\Delta_k = \begin{array}{cc} \left(\begin{array}{cc} \boxed{\Sigma_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{array}{ccc} \beta & & 0 \\ & \ddots & \\ 0 & & \beta \end{array}} \end{array} \right) & \begin{array}{l} \left.\vphantom{\Sigma_k}\right\} d \\ \left.\vphantom{\begin{array}{c}\beta\\ \ddots\\ \beta\end{array}}\right\} p-d \end{array} \end{array}$$

FEM Algorithm

Iterative algorithm like *EM* but for functional data

F step aims to determine the orientation matrix U of F

M step maximize, conditionally on U, the expectation of the log-likelihood

$$Q(\theta;\theta^{(q-1)}) = E\left[\ell(\theta;\mathbf{\Gamma}, z_1, ..., z_n)|\mathbf{\Gamma}, \theta^{(q-1)}\right]$$

$$\pi_k^{(q)} = n_k^{(q-1)}/n,$$
$$\mu_k^{(q)} = \frac{1}{n_k^{(q-1)}} \sum_{i=1}^n t_{ik}^{(q-1)} U^{(q)t} \gamma_i,$$
$$\Sigma_k^{(q)} = U^{(q)t} C_k^{(q)} U^{(q)},$$
$$\beta^{(q)} = \left(\text{trace}(C^{(q)}) - \sum_{j=1}^d u_j^{(q)t} C^{(q)} u_j^{(q)}\right) / (p-d)$$

FEM Algorithm

E step updates the posterior probabilities

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} \phi(\gamma_i, \theta_k^{(q)})}{\sum_{l=1}^{K} \pi_l^{(q)} \phi(\gamma_i | \theta_l^{(q)})}$$

Model Selection

One has to choose:

- Model Family
- Number of clusters K
- Discriminative basis functions

Criteria:

- AIC, BIC or Slope Criteria
- Model itself

Our dataset

CO2 Emissions Dataset (Data from The World Bank)

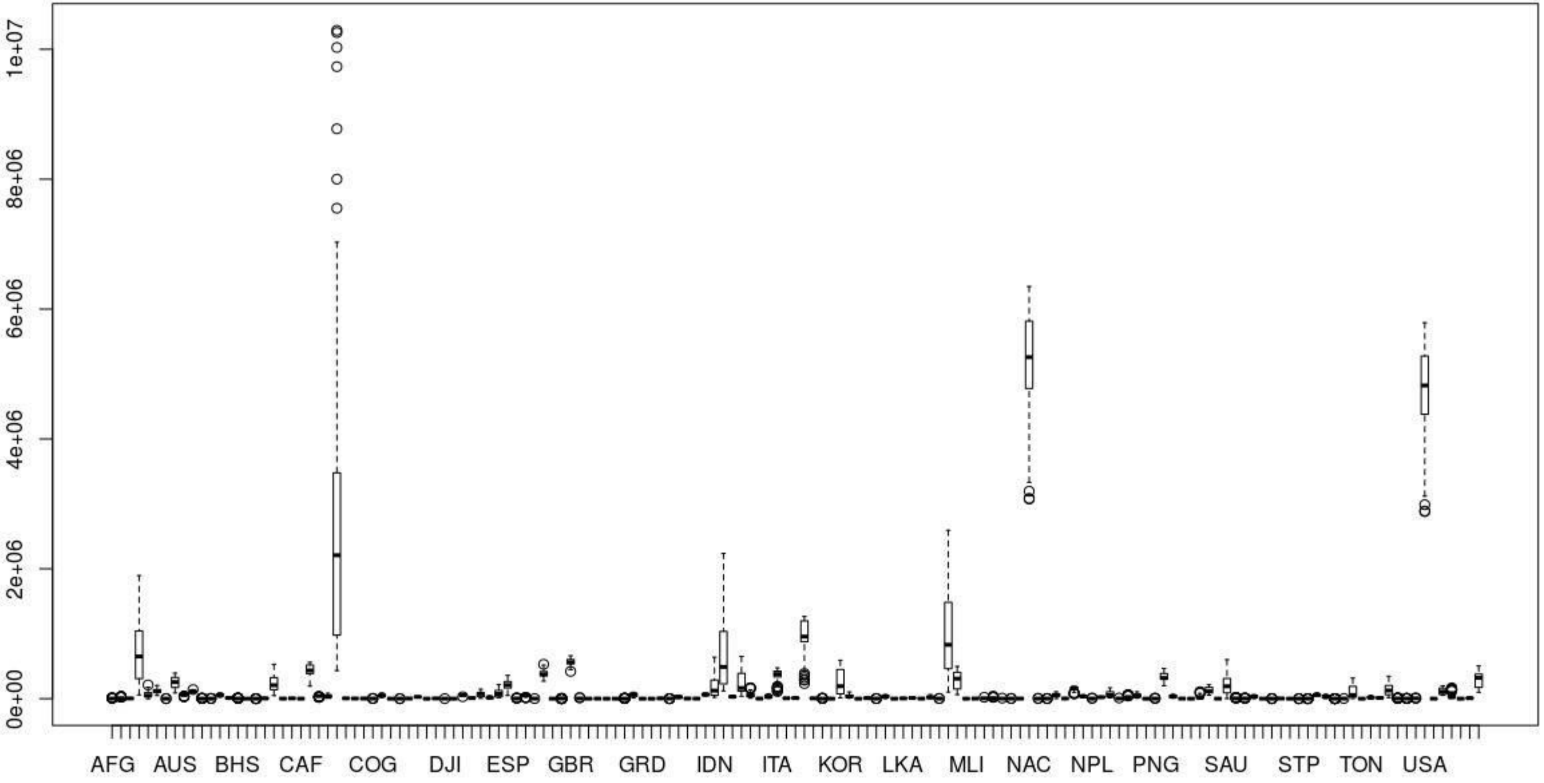
- Time Series Data
- CO2 emissions of the countries in the world
- Metric: tons per capita
- Aggregation Method: Weighted Average
- Periodicity: Annual

Dataset Size: 149 time series (countries)

Error Margin: 10% (the data is based on estimations)

Note: The trend in this time series data is more accurate than individual values.

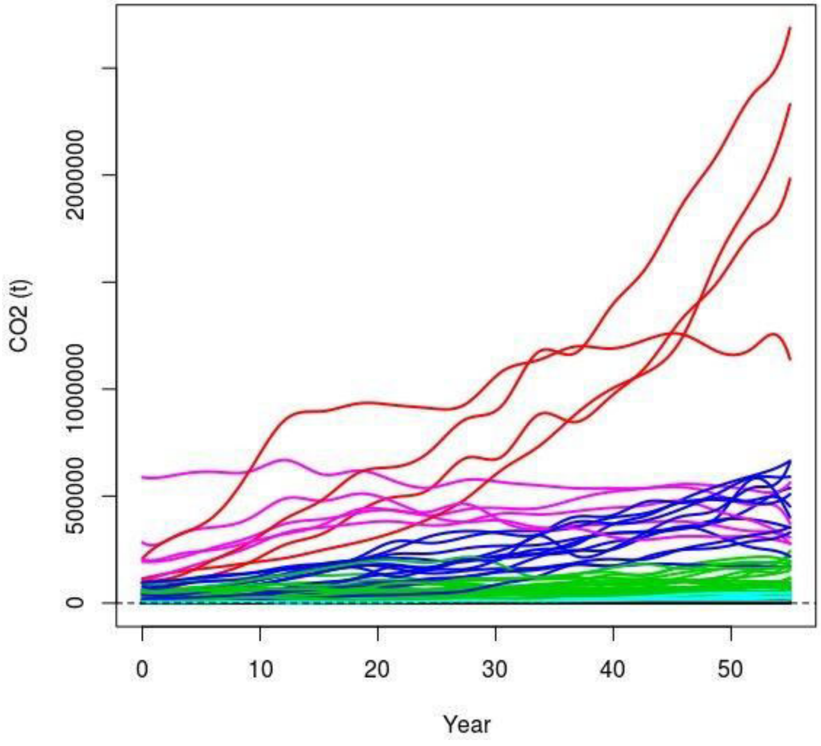
Data Inspection: Boxplot



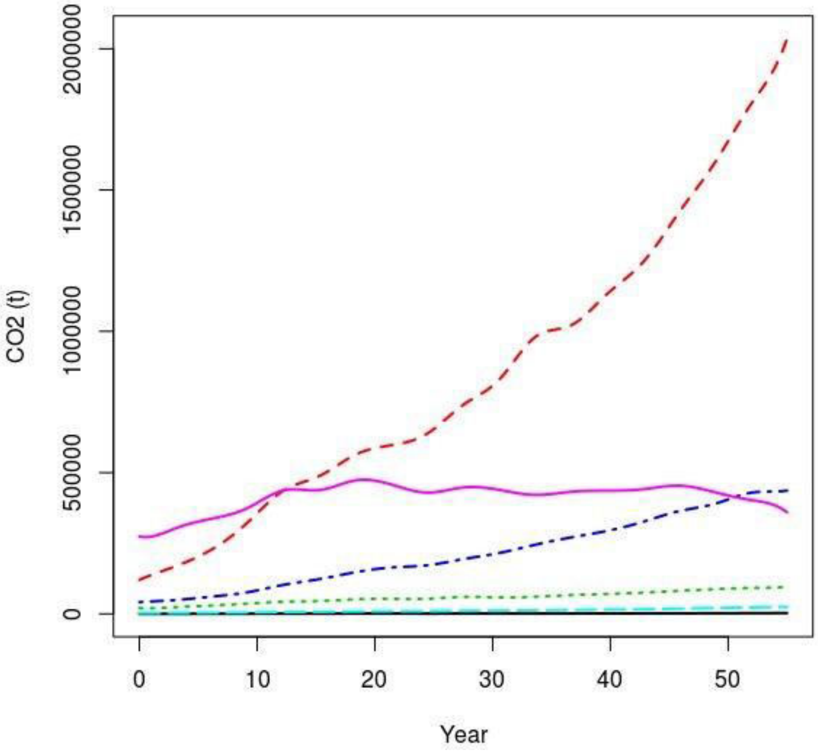
Model selection

Parameter	**Options**	**Choice**
Basis functions	B-spline, fourier	B-spline, 21 basis
K	2 - 20	6
Selection criterion	AIC, BIC, ICL, slope	BIC
Initialization type	Random, K-means, Hclust	Hclust
Model	...	$\alpha_k \beta_k$

Results

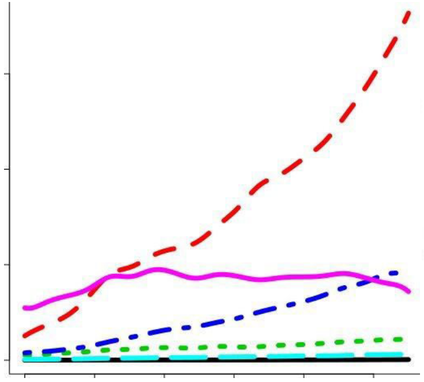


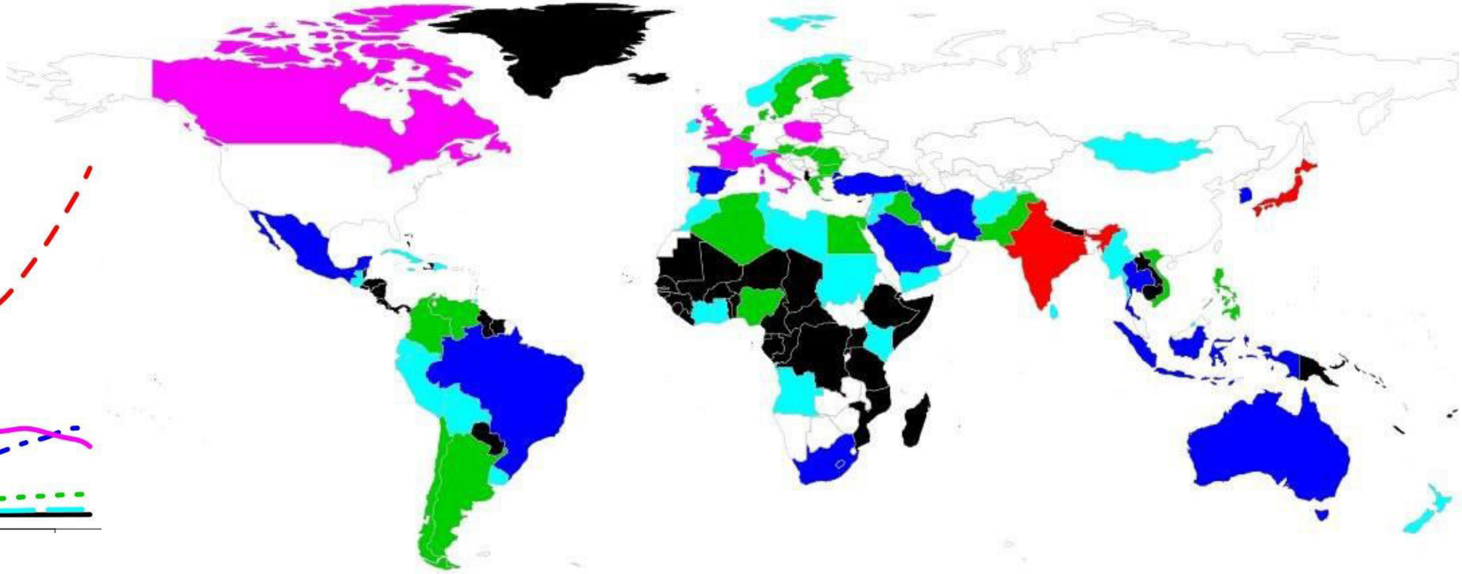
CO2 (t)
0
500000
1000000
2000000
0
10
20
30
40
50
Year



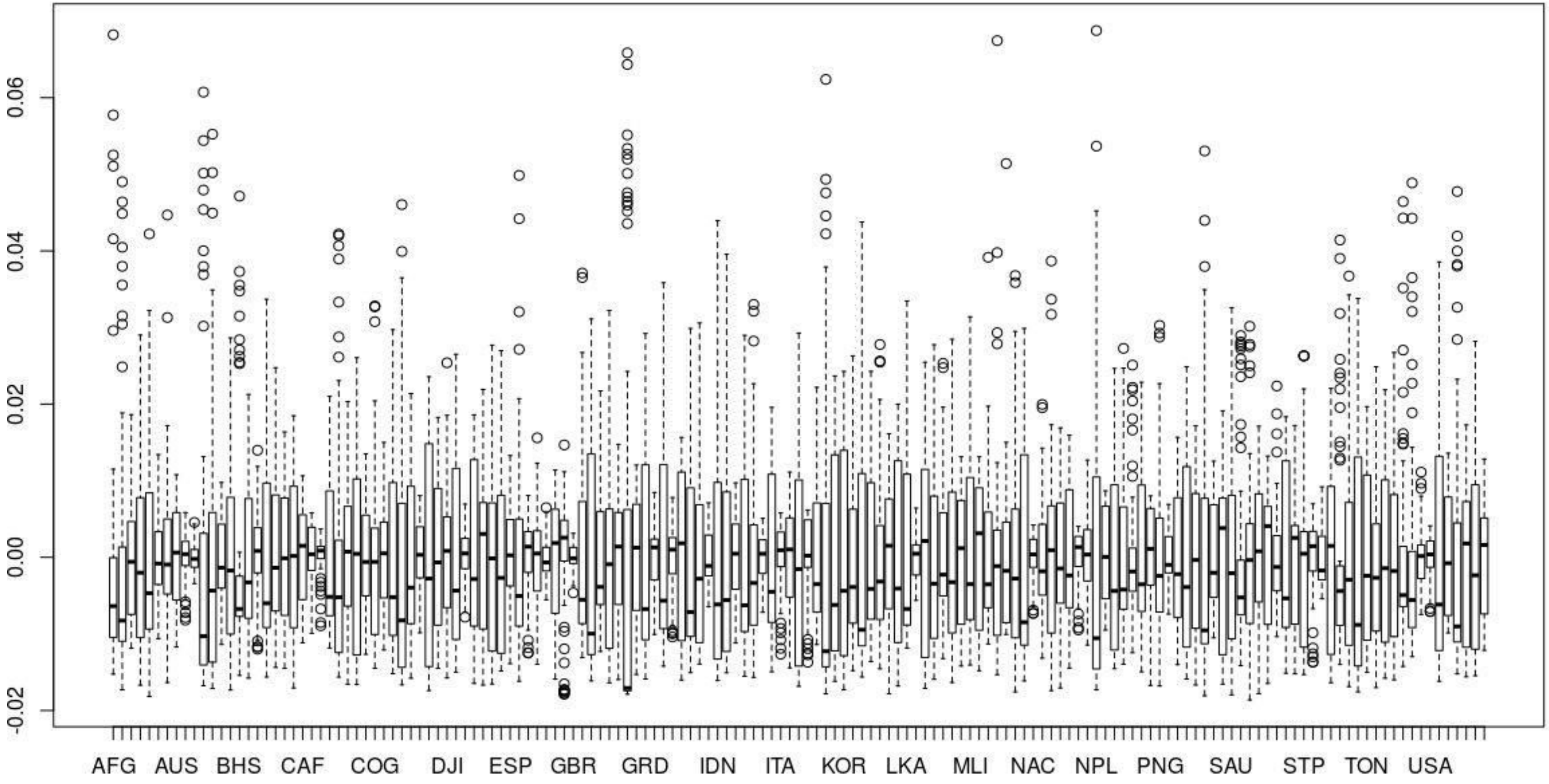
CO2 (t)
0
500000
1000000
1500000
2000000
0
10
20
30
40
50
Year

Results (II)





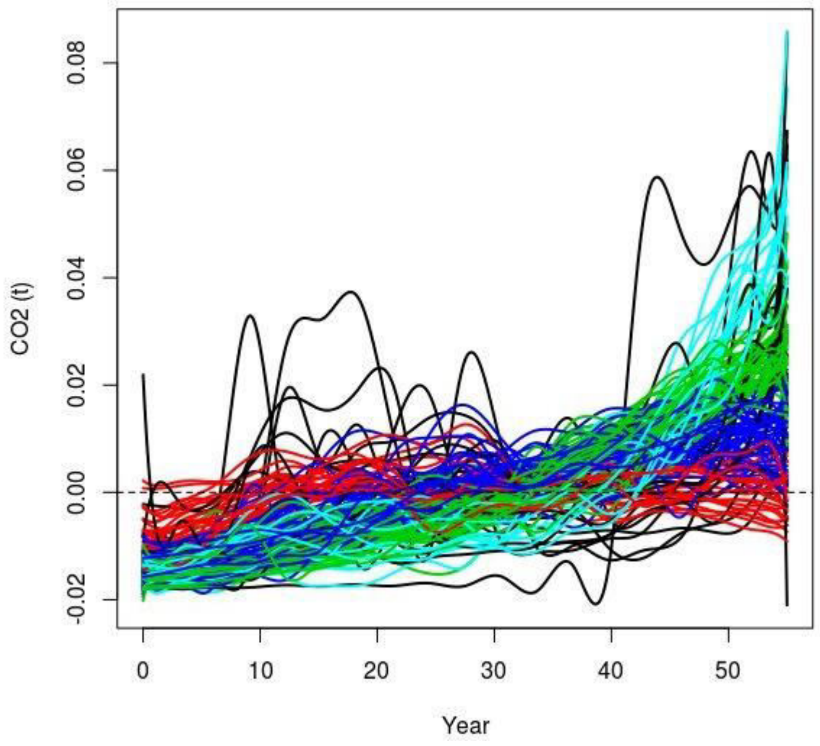
Variation: normalized data



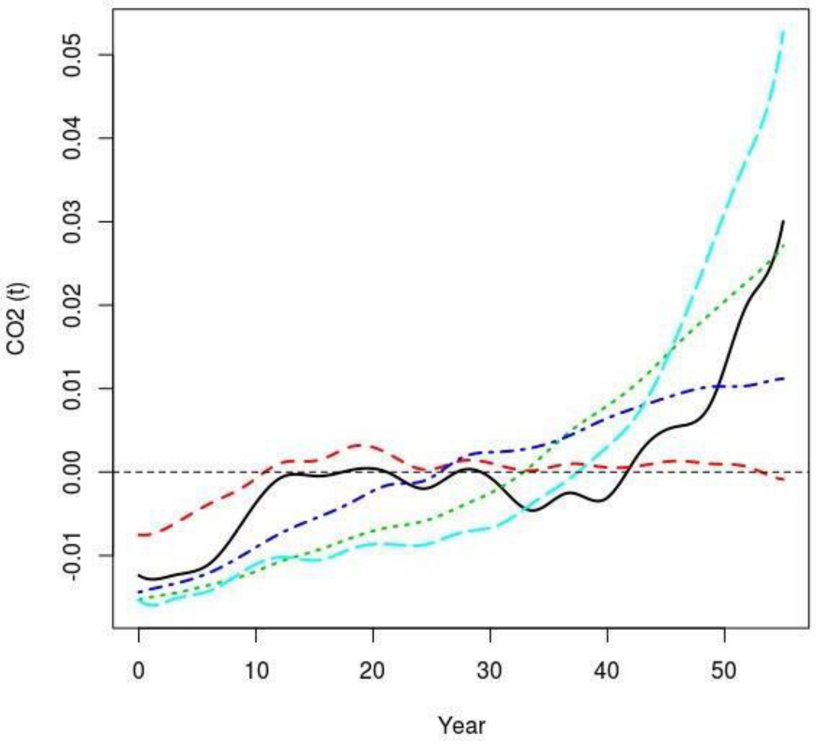
Model selection: normalized data

Parameter	Options	Choice
Basis functions	B-spline, fourier	B-spline, 21 basis
K	2 - 20	5
Selection criterion	AIC, BIC, ICL, slope	BIC
Initialization type	Random, K-means, Hclust	Hclust
Model	...	$\alpha_k \beta_k$

Results

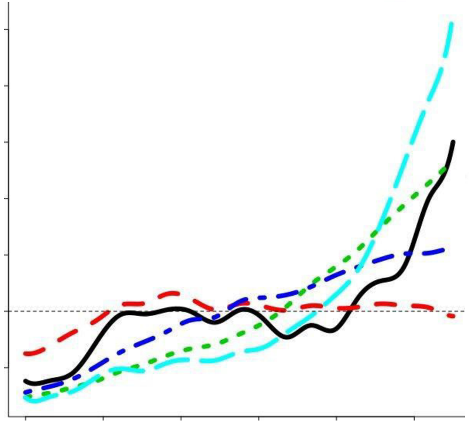


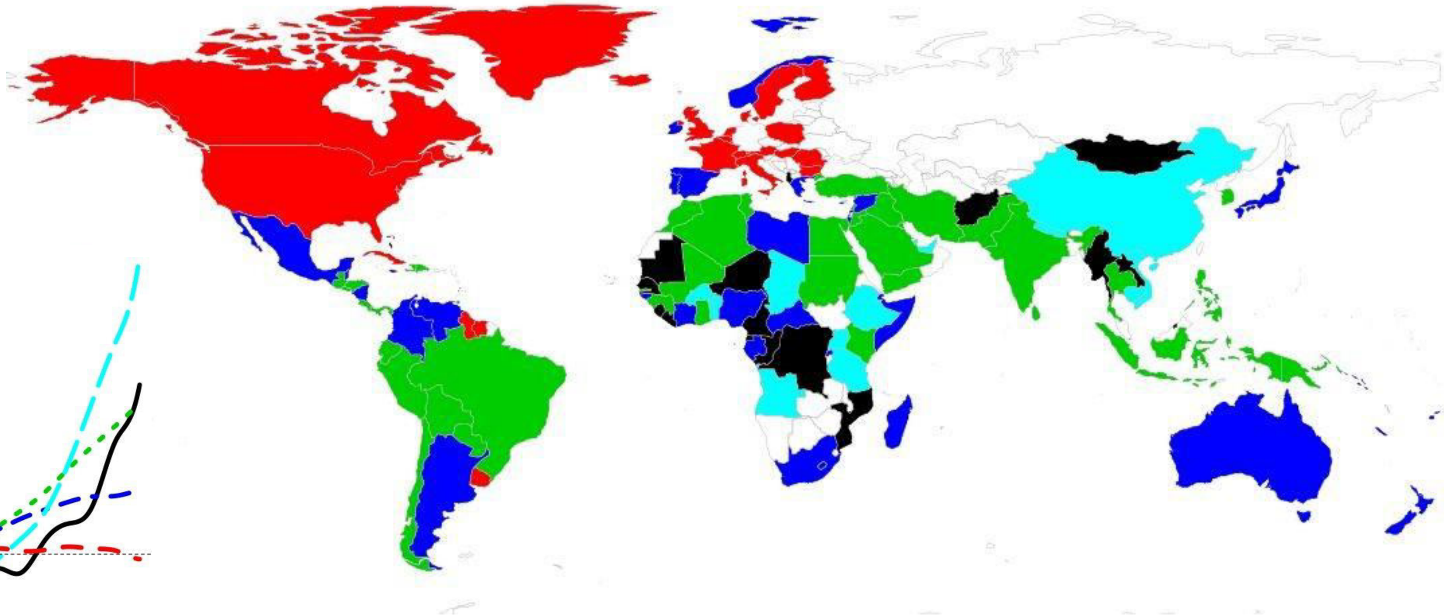
CO2 (t)
Year
0.08
0.06
0.04
0.02
0.00
-0.02
0
10
20
30
40
50



CO2 (t)
Year
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
0
10
20
30
40
50

Results visualization





Questions?