C. Acal
chracal@ugr.es

Solving the multivariate functional ANOVA problem with application to environmental data from COVID-19 pandemic

Christian Acal[1], Ana M. Aguilera[1], Annalina Sarra[2], Adelia Evangelista[2], Tonio Di Battista[2], Sergio Palermi[3]

[1]University of Granada (Spain)
[2]University G. d' Annunzio (Pescara, Italy)
[3] Agency of Environmental Protection of Abruzzo (Pescara, Italy)

UNIVERSIDAD DE GRANADA

New Bridges between Mathematics and Data Science
Valladolid, Spain, November 8-11, 2021

Introduction & Motivation

C. Acal
chracal@ugr.es

Functional ANOVA problem from two different perspectives

1. FANOVA for repeated measures
2. Multivariate FANOVA for independent measures

Theoretical objectives

1. Extending the statistics available in the literature by considering the basis expansion of the curves
2. Generalizing the univariate principal component approach proposed in Aguilera et al. (2021) for multivariate functional data

Motivation

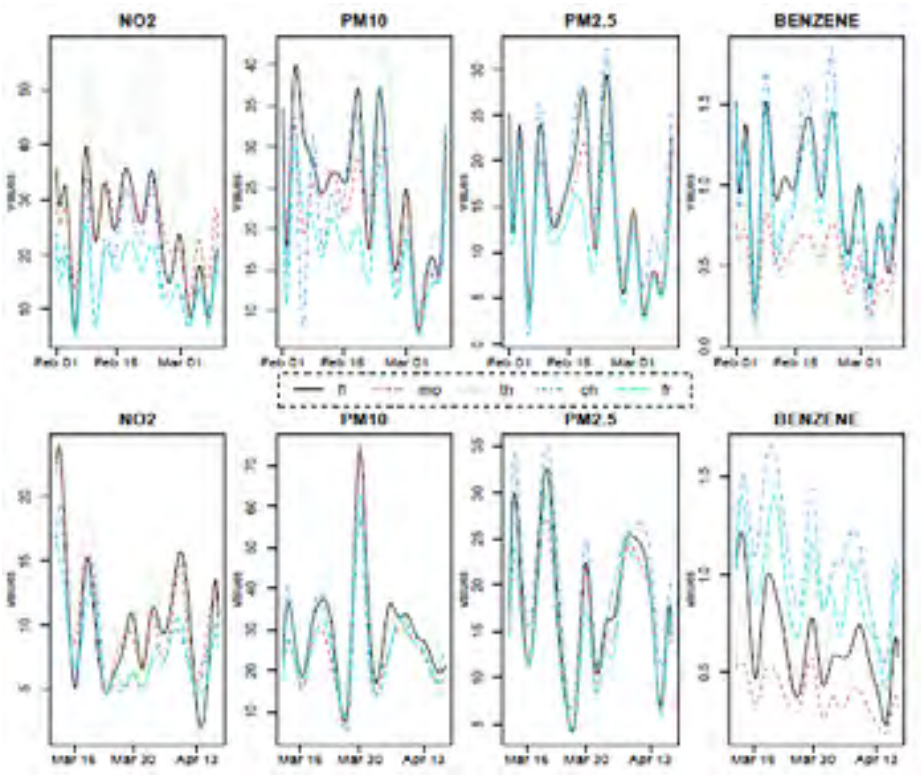
Analyzing the impact of quarantine policies on air quality in the Abruzzo Region (Italy)

Introduction & Motivation

Experimental data

C. Acal
chracal@ugr.es

- Daily average measurements about four air pollutants concentrations (NO2, PM10, PM2.5 and Benzene) in Abruzzo, Italy. Data measured in two different periods (pre and during lockdown) by five monitoring stations classified by their location (traffic and background stations)



C. Acal
chracal@ugr.es

Goals in the application

1. Studying whether the level of each pollutant decreased during the lock-down period

FANOVA for repeated measures

2. Studying the differences between the temporal evolution of four pollutants in terms of the location of the monitoring stations

Multivariate FANOVA for independent groups

Previous results in FDA

C. Acal
chracal@ugr.es

Theory about FPCA

Functional variable X with sample functions in $L^2[T]$

$$< f, g >= \int_T f(t)g(t)dt$$

Principal component score

$$\xi_j = \int_T (X(t) - \mu(t))f_j(t)dt$$

- $\mu(t)$ is the mean function
- f_j are the solutions to the eigenequation

$$C(f_j)(t) = \int_T C(t,s)f_j(s)ds = \lambda_j f_j(t)$$

- $C(t,s)$ is the covariance function
- $\lambda_j = Var[\xi_j]$

Karhunen-Loève expansion

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j f_j(t) \rightarrow X^q(t) = \mu(t) + \sum_{j=1}^{q} \xi_j f_j(t)$$

Previous results in FDA

C. Acal
chracal@ugr.es

Basis expansion of the sample curves (Ramsay and Silverman, 2002)

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t) = \mathbf{a}_i'\Phi(t), \quad i = 1, \ldots, n$$

Options for basis systems

1. Fourier basis for periodic data
2. B-spline basis for non-periodic smooth data with continuous derivatives up to certain order
3. Wavelet basis for data with a strong local behavior whose derivatives are not required

FPCA of $X(t) \leftrightarrow$ Multivariate PCA of matrix $A\Psi^{1/2}$ (Ocaña et al., 2007)

$A = (a_{ij})_{i=1,\ldots,n;j=1,\ldots,p}$ matrix of basis coefficients

$\Psi = \left(< \phi_j, \phi_k >_{L^2[0,1]}\right)_{j,k=1,\ldots,p}$ matrix of inner products

C. Acal
chracal@ugr.es

Problem: FANOVA for repeated measures

Checking if the mean of a functional variable observed in two different conditions or periods of time (repeated measures) is the same

Sample of curves

$X_{jr}(t)$ with $t \in \mathcal{T} = [a, b]$, $j = 1, ..., n$ and $r = 1, 2$

Linear model

$$X_{jr}(t) = \mu_r(t) + e_{jr}(t)$$

- $E[X_{jr}(t)] = \mu_r(t)$
- $e_{jr}(t)$ are independent random functions centered in mean

Aim

$$\begin{cases} H_0 : \mu_1(t) = \mu_2(t) \ \forall t \in [a, b] \\ H_1 : \mu_1(t) \neq \mu_2(t) \ \textit{for some } t \end{cases}$$

C. Acal
chracal@ugr.es

Martinez-Camblor and Corral (2011)

$$\mathcal{C}_n = n \int_T (\overline{X}_1(t) - \overline{X}_2(t))^2 dt$$

$\overline{X}_r(t) = n^{-1} \sum_{j=1}^{n} X_{jr}(t)$ is the mean function for each condition or period of time

Smaga (2020)

$$\mathcal{D}_n = n \int_T \frac{\left(\overline{X}_1(t) - \overline{X}_2(t)\right)^2}{\hat{K}(t,t)} \, dt$$

$$\mathcal{E}_n = sup_{t \in [a,b]} \left\{ \frac{n \left(\overline{X}_1(t) - \overline{X}_2(t)\right)^2}{\hat{K}(t,t)} \right\}$$

$\hat{K}(t,t) = \frac{\sum_{j=1}^{n} \left[(X_{j1}(t) - \overline{X}_1(t)) - (X_{j2}(t) - \overline{X}_2(t)) \right]^2}{n-1}$ is the unbiased estimator of the asymptotic covariance function

C. Acal
chracal@ugr.es

$\mathcal{C}_n$, $\mathcal{D}_n$ and $\mathcal{E}_n$ computed in terms of basis functions

Numerator

$$\begin{aligned}\left(\overline{X}_1(t)-\overline{X}_2(t)\right)^2 &= \left(\bar{\mathbf{a}}_1'\phi(t)-\bar{\mathbf{a}}_2'\phi(t)\right)^2\\ &= \left(\phi(t)'\bar{\mathbf{d}}\right)^2=\phi(t)'\bar{\mathbf{d}}\bar{\mathbf{d}}'\phi(t)\end{aligned}$$

Denominator

$$\begin{aligned}\hat{K}(t,t) &= Var(X_1(t))-2Cov(X_1(t),X_2(t))+Var(X_2(t))\\ &= \hat{C}_1(t,t)-2\hat{C}_{12}(t,t)+\hat{C}_2(t,t)\\ &= \phi(t)'(\hat{\Sigma}_1-2\hat{\Sigma}_{12}+\hat{\Sigma}_2)\phi(t)\end{aligned}$$

- $\bar{\mathbf{d}}=(\overline{d}_1,...,\overline{d}_p)'=\bar{\mathbf{a}}_1-\bar{\mathbf{a}}_2=(\overline{a}_{11},...,\overline{a}_{1p})'-(\overline{a}_{21},...,\overline{a}_{2p})'$
- $\overline{a}_{rk}=n^{-1}\sum_{j=1}^{n}a_{jrk}\ \ r=1,2;\ k=1,...,p$
- $\hat{\Sigma}_r$ is the sample covariance matriz of $A_r=(a_{jrk})$
- $\hat{\Sigma}_{12}$ is the sample cross-covariance matrix between A_1 and A_2
- $\overline{X}_r=n^{-1}\sum_{j=1}^{n}\mathbf{a}_{jr}'\phi(t)=\bar{\mathbf{a}}_r'\phi(t)$

C. Acal
chracal@ugr.es

Problem: Multivariate FANOVA for independent measures

Checking the equality of the multivariate dimensional mean functions for independent groups

Sample of curves

$X_{ijh}(t)$ with $i = 1, ..., g$, $j = 1, ..., n_i$ and $h = 1, ..., H$

Model

$\mathbf{X}_{ij}(t) = (X_{ij1}(t), ..., X_{ijH}(t))'$ are i.i.d. multivariate functional variables with

- mean vector $\boldsymbol{\mu}_i(t) = (\mu_{i1}(t), ..., \mu_{iH}(t))'$
- matrix covariance function $\mathbf{C}(t, s) = (C_{h,h'}(t, s))$, $t, s \in \mathcal{T}$
  - if $h = h'$, then $C_{h,h}$ is the covariance function
  - if $h \neq h'$, then $C_{h,h'}$ is the cross-covariance function

Aim

$$H_0 : \boldsymbol{\mu}_1(t) = ... = \boldsymbol{\mu}_g(t) \ \forall t \in [a, b]$$

against the alternative that its negation holds

C. Acal
chracal@ugr.es

Aguilera et al. (2021) - Univariate case

- Sample of curves independent and identically distributed

$$\{X_{hi}(t): \ i = 1, ..., n_h; \ h = 1, ..., m; \ t \in T\}$$

- FANOVA is equivalent to MANOVA of matrix $A_{(\sum_{h=1}^{m} n_h \times p)}$
- Problems
  1. Multivariate homogeneity tests do not perform well with high-dimensional vectors
  2. Number of basis functions needed for an accurate approximation of the sample curves is usually high
- **Functional principal component approach**: The problem is reduced to testing homogeneity on a small set of functional PCs
  1. One way ANOVA for each PC with Bonferroni's correction
  2. Non-parametric multivariate tests

C. Acal
chracal@ugr.es

Theory about Multivariate FPCA

Principal component scores

$$\xi_{ijm} = \int_{\mathcal{T}} (\mathbf{X}_{ij}(t) - \boldsymbol{\mu}(t))' \mathbf{f}_m(t) dt = \sum_{h=1}^{H} \int_{\mathcal{T}} (X_{ijh}(t) - \mu_h(t)) f_{mh}(t) dt$$

Karhunen-Loève expansion

$$\mathbf{X}_{ij}(t) = \boldsymbol{\mu}(t) + \sum_{m=1}^{\infty} \xi_{ijm} \mathbf{f}_m(t) \rightarrow \mathbf{X}_{ij}^{q}(t) = \boldsymbol{\mu}(t) + \sum_{m=1}^{q} \xi_{ijm} \mathbf{f}_m(t)$$

Multivariate basis expansion: $\mathbf{X}_{ij}(t) = \boldsymbol{\Phi}(t)\mathbf{a}_{ij}'$

$$\boldsymbol{\Phi}(t) = \begin{pmatrix} \phi_{11}(t) & \cdots & \phi_{1p_1}(t) & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \phi_{21}(t) & \cdots & \phi_{2p_2}(t) & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & \phi_{H1}(t) & \cdots & \phi_{Hp_H}(t) \end{pmatrix}$$

C. Acal
chracal@ugr.es

MFPCA of $\mathbf{X}(t) \leftrightarrow$ MPCA of $\mathbf{AW}^{1/2}$ (Jacques and Preda, 2014)

$$\xi_{ijm} = \mathbf{a}_{ij}' \mathbf{W} \mathbf{b}_m$$

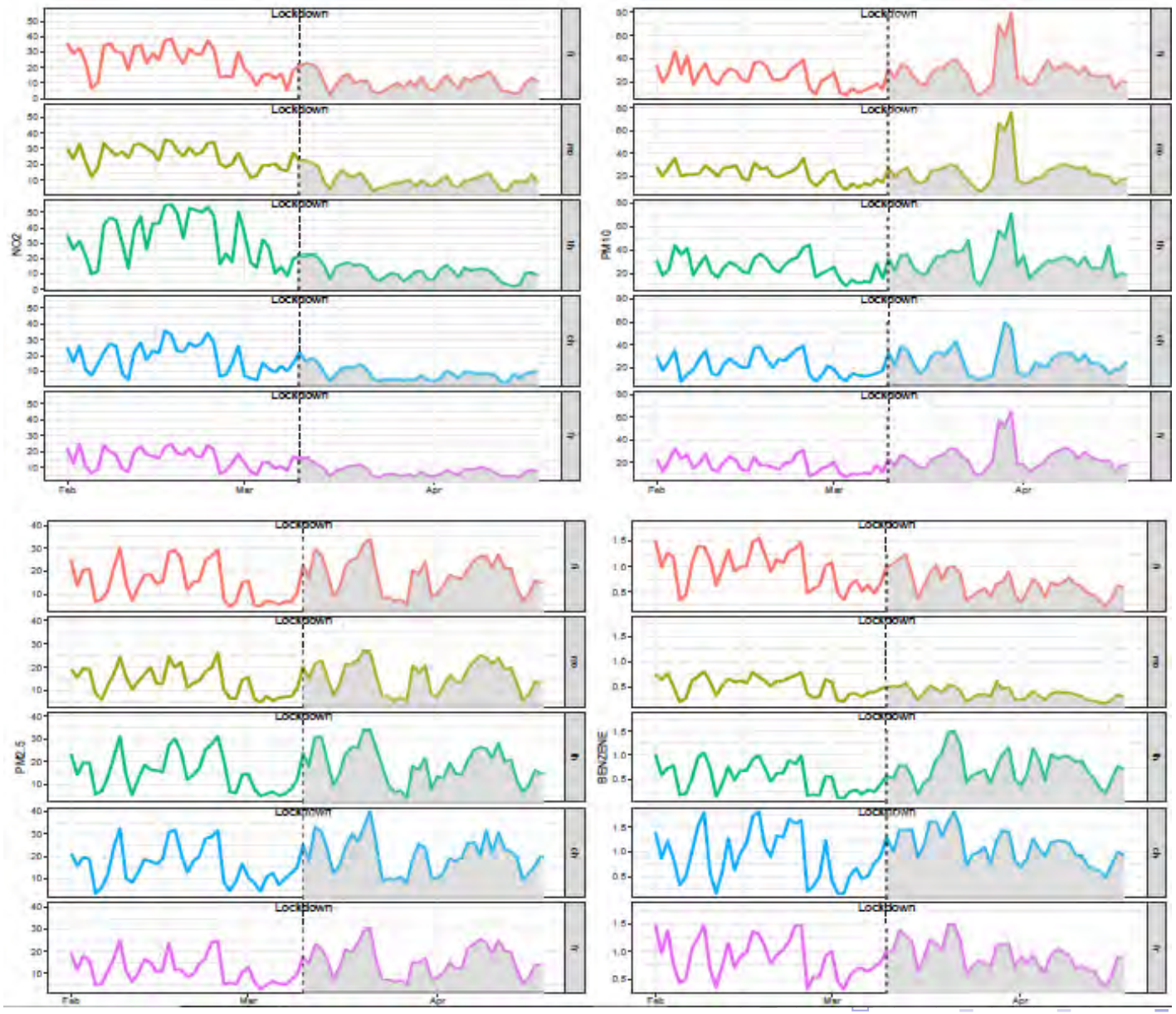
- $\mathbf{b}_m$ basis coefficients of $\mathbf{f}_m(t) = \mathbf{\Phi}(t)\mathbf{b}_m'$
- $\mathbf{W} = \int_{\mathcal{T}} \mathbf{\Phi}(t)' \mathbf{\Phi}(t) dt$

Solution

Testing multivariate homogeneity on the vectors of the most explicative principal components scores

Application to environmental data

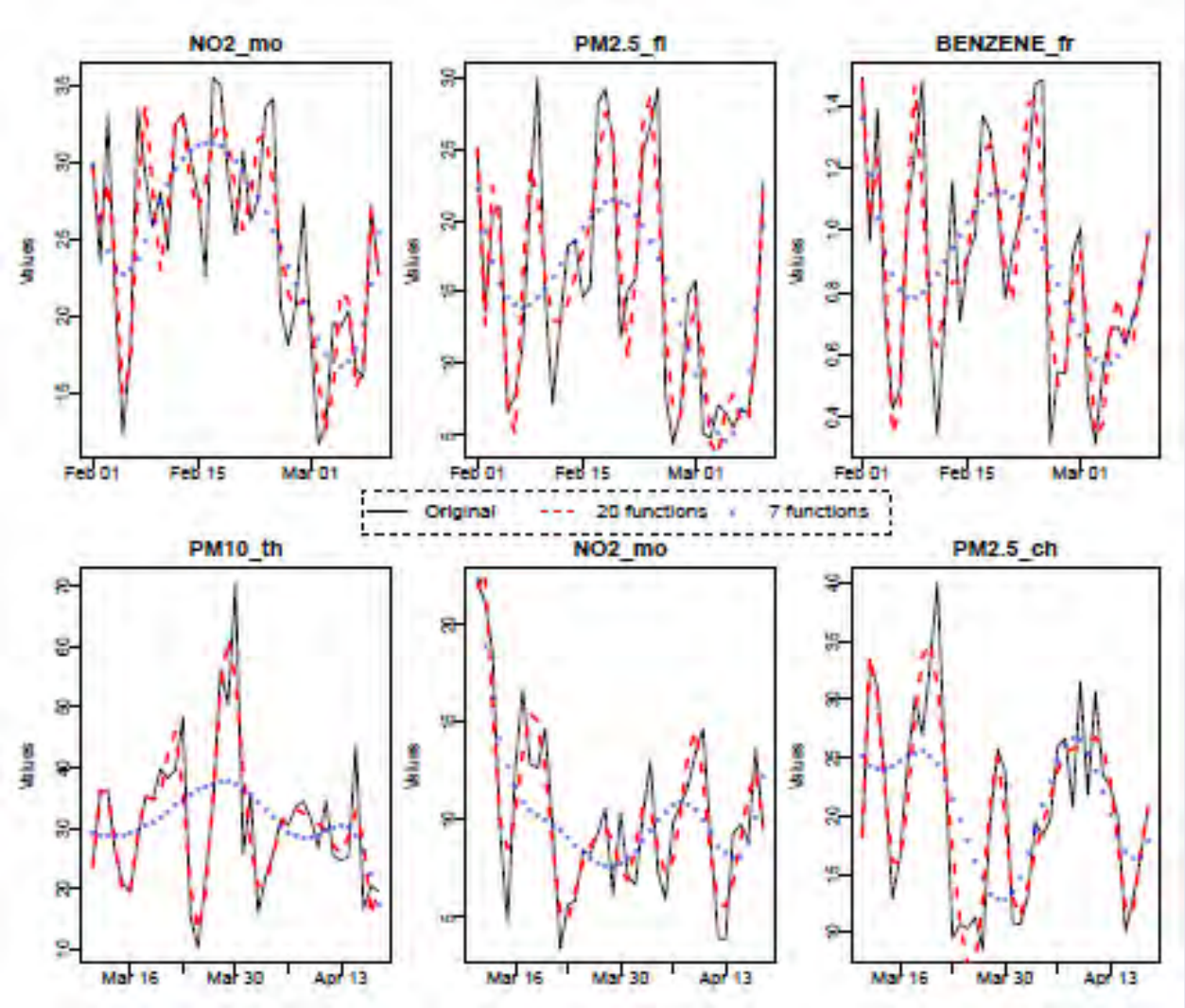
C. Acal
chracal@ugr.es



Application to environmental data

C. Acal
chracal@ugr.es

Functional reconstruction of pollutant curves

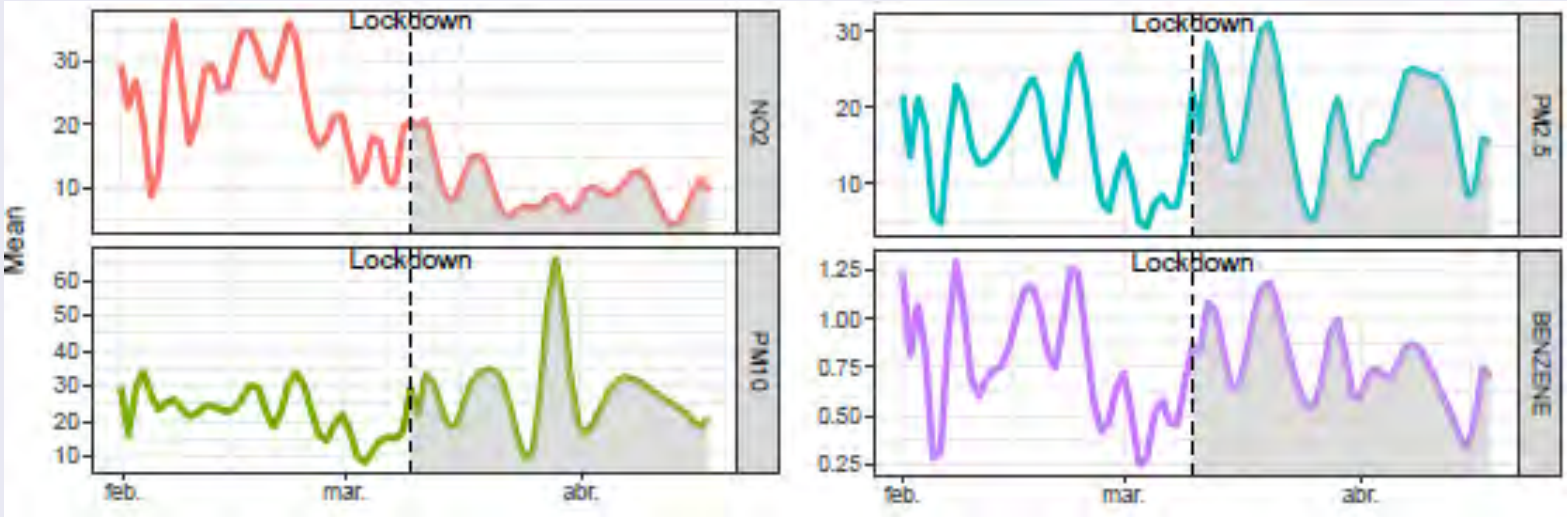


Application to environmental data

C. Acal
chracal@ugr.es

FANOVA for repeated measures results

Has the level of each pollutant changed during the lockdown period?



	$\mathcal{D}_n$	$\mathcal{E}_n$
NO_2	0.034	0.035
PM_{10}	0.000	0.034
$PM_{2,5}$	0.028	0.030
Benzene	0.049	0.070

(p-values obtained by means of a permutation test)

Application to environmental data

C. Acal
chracal@ugr.es

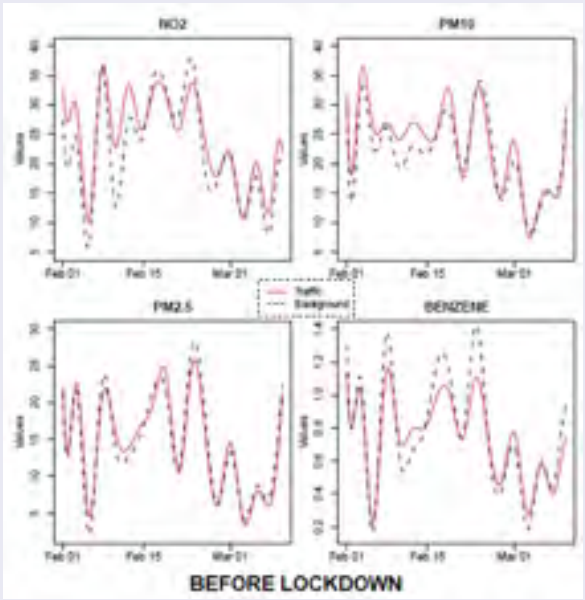
Multivariare FANOVA results for independent measures

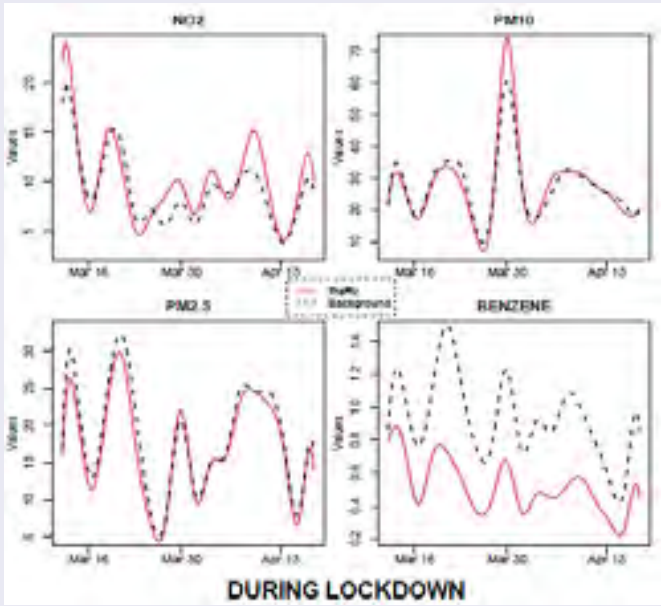
Are there differences between the temporal evolution of all pollutants in terms of the location of measuring stations?

Application to environmental data

C. Acal
chracal@ugr.es

Multivariare FANOVA results for independent measures





	BL	DL
All pollutants	0.000	0.302
NO_2	0.562	0.272
PM_{10}	0.000	0.306
$PM_{2,5}$	0.889	0.685
Benzene	0.186	0.000

- 4 PCs for multivariate and univariate analysis
- 99-100 % of total variability is explained
- Extension of Kruskal Wallis test (permutation version)

Conclusions

C. Acal
chracal@ugr.es

The current work addresses the functional ANOVA problem for two different theoretical frameworks

1. **FANOVA for repeated measures**:
   The statistics available in the literature are extended by considering the basis expansion of the curves
2. **Multivariate FANOVA for independent measures**:
   A novel approach based on multivariate FPCA has been introduced. Here, the problem is reduced to test multivariate homogeneity on the vectors of the most explicative PCs

Application

Analyzing the impact of quarantine policies on air quality in the Abruzzo Region (Italy)

1. The level of each pollutant changed during the lockdown period
2. Significant differences were found in terms of the location of the monitoring stations in relation to PM10 (before lockdown) and benzene (during lockdown)

Future directions

C. Acal
chracal@ugr.es

- Extending the functional homogeneity testing approaches based on FPCA for repeated measures
- Developing new homogeneity tests based on the PHD of PCs
- ....

Acknowledgments

- Proyecto A-FQM-66-UGR20 (Universidad de Granada. Programa Operativo FEDER Andalucía)
- Proyecto PID2020-113961GB-I00 (Ministerio de Ciencia e Innovación)
- IMAG - Unidad de excelencia María de Maeztu - CEX2020-001105-M

References

C. Acal
chracal@ugr.es

- **Acal et al. (2021)**
  Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. *Stochastic Environmental Research and Risk Assessment*, in press

- **Aguilera et al. (2021)**
  Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Mathematics and Computers in Simulation*, 186, 41-51.

- **Jacques and Preda (2014)**
  Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71, 92-106.

- **Ocaña et al. (2007)**
  Computational considerations in functional principal component analysis. *Computational Statistics*, 22 (3), 449-465.

C. Acal
chracal@ugr.es

Solving the multivariate functional ANOVA problem with application to environmental data from COVID-19 pandemic

Christian Acal[1], Ana M. Aguilera[1], Annalina Sarra[2], Adelia Evangelista[2], Tonio Di Battista[2], Sergio Palermi[3]

[1]University of Granada (Spain)
[2]University G. d' Annunzio (Pescara, Italy)
[3] Agency of Environmental Protection of Abruzzo (Pescara, Italy)

UNIVERSIDAD DE GRANADA

New Bridges between Mathematics and Data Science
Valladolid, Spain, November 8-11, 2021