UNIVERSITY OF GRANADA

DOCTORATE PROGRAM IN MATHEMATICAL AND APPLIED STATISTICS

Doctoral Thesis

ADVANCES IN STOCHASTIC AND FUNCTIONAL MODELING OF HIGH DIMENSION DATA

CHRISTIAN JOSÉ ACAL GONZÁLEZ

Thesis supervised by
Prof. Ana María Aguilera del Pino
Prof. Juan Eloy Ruiz Castro

Granada 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Christian José Acal González
ISBN: 978-84-1306-942-5
URI: http://hdl.handle.net/10481/69648

A mis padres, a mi hermano y a Cristina

Agradecimientos

En primer lugar, me gustaría agradecer de todo corazón a los grandes artífices de este proyecto, a mis directores de tesis Ana María Aguilera del Pino y Juan Eloy Ruiz Castro. Gracias por confiar en mí, por inculcarme esta pasión por la investigación, por animarme incondicionalmente en todo momento y por la infinita paciencia que han tenido conmigo. Sin su esfuerzo y dedicación durante estos años, jamás hubiera sido capaz de poder alcanzar esta meta. Siempre estaré en deuda con ellos.

También a Manuel Escabias Machuca y a Juan Bautista Roldán, que aun no siendo mis directores, han colaborado estrechamente en esta tesis y siempre se han ofrecido a ayudarme en todo lo que he necesitado. Asimismo, me gustaría mostrar mi gratitud a María del Carmen Aguilera Morillo (Universitat Politécnica de Valencia, España) y a Tonio Di Battista (Università degli Studi G. d'Annunzio Chieti e Pescara, Italia) por brindarme la oportunidad de trabajar con ellos. Para mí ha sido todo un placer y un honor.

Reconocer también a todos los profesores y profesoras que me dieron clase durante el Grado en Estadística, y en general, a todo el Departamento de Estadística e Investigación Operativa de la Universidad de Granada por hacerme sentir como uno más desde el primer día. Sería muy injusto dar nombres porque de tod@s he aprendido y me han ayudado a alcanzar mis objetivos, pero me gustaría hacer mención especial, por un lado, a los integrantes de la unidad docente de Farmacia y, por otro lado, a la unidad docente de Bioestadística de Medicina, por su trato y confianza durante el período que he estado destinado allí para impartir clase.

Gracias también a todos mis compañeros del Grado con los que comencé esta aventura, sobre todo a Luis Jaime Jaenada, un magnífico estadístico y mejor amigo, y a mis compañeros del despacho de 'Los Becarios' con los que termino esta etapa: los ya Doctores Javier Álvarez y Beatriz Cobo, y especialmente, al futuro Doctor Ramón Ferri, con el que he compartido todo tipo de anécdotas, las mayorías de nuestra otra pasión, el fútbol.

Finalmente, dedico esta tesis a mis padres, a mi hermano y a Cristina. Gracias a mis padres por todo su amor y su enorme esfuerzo y sacrificio que han hecho por mí desde que era pequeño para que nunca me faltara de nada y poder alcanzar mis sueños. A mi hermano Alejandro por ser con quién he compartido toda mi vida, por esos momentos de risa infinitos y por ser mi mejor amigo. A Cristina por su apoyo y comprensión, por su cariño, por no dejarme caer nunca y por ser mi compañera de viaje tantos años. Gracias a los cuatro por ser los protagonistas de mi vida.

NOTA: Este trabajo de investigación se ha desarrollado al amparo de una beca para la Formación del Profesorado Universitario (FPU18/01779) concedida por el Ministerio de Ciencia, Innovación y Universidades, así como del proyecto MTM2017-88708-P en el que he sido partícipe.

RESUMEN

En muchos campos científicos, es habitual encontrar magnitudes caracterizadas por la evolución de una variable aleatoria a lo largo de algún continuo (proceso estocástico). A pesar de que los datos experimentales medidos sobre estas variables son claramente funciones (curvas, superficies o imágenes), históricamente su tratamiento ha sido a través del análisis multivariante o de series temporales, perdiéndose información importante. Por suerte, los grandes avances que ha experimentado el sector tecnológico en los últimos años, han facilitado el seguimiento y reconstrucción de las funciones de forma rápida y sin esfuerzo, siendo posible trabajar con las funciones completas. En este escenario, es altamente probable tener datos de alta dimensión, en los que el número de variables es mayor que el número de individuos muestreados. Este hecho hace que los métodos estadísticos tradicionales no sean adecuados. Dependiendo del propósito final, en esta tesis se abordan estos datos desde dos perspectivas estadísticas diferentes y complementarias: el Análisis de Datos Funcional (FDA) y el Análisis de la Fiabilidad (RA) basado en las distribuciones de probabilidad Tipo Fase (PH).

FDA surge ante la necesidad de construir métodos que permitan modelizar datos funcionales, cuyas observaciones suelen ser curvas dependiendo del tiempo u otro argumento continuo. En las últimas décadas, se viene realizando una intensa investigación en este campo, en el que se han generalizado la mayoría de las técnicas multivariantes, especialmente, métodos de reducción de la dimensión, clasificación y regresión. Destaca el Análisis de Componentes Principales (FPCA) porque reduce la dimensión y explica la estructura de variabilidad en términos de un número pequeño de variables incorreladas.

En el campo de la fiabilidad, uno de los objetivos es estudiar el comportamiento de sistemas complejos, cuyo funcionamiento está condicionado por varios factores incontrolables. En este sentido, RA intenta identificar la distribución de probabilidad de los datos para arrojar luz sobre la variabilidad que hay detrás del funcionamiento de los sistemas. Una posibilidad es considerar los procesos Markovianos y las distribuciones PH. Esta clase de distribuciones es capaz de aproximar cualquier distribución no negativa tanto como se desee gracias a su versatilidad, y permite modelar problemas complejos con resultados bien estructurados.

Las contribuciones metodológicas de esta tesis se desarrollan en base a problemas de gran interés impulsados por datos relacionados con las Memorias Resistivas de Acceso Aleatorio (RRAMs) y la pandemia de COVID-19. Las RRAM despiertan un gran interés porque son una de las principales fuentes de ingresos en la industria, mientras que para mitigar la propagación del virus, es crucial desarrollar modelos óptimos que ayuden a tomar buenas decisiones.

Un nuevo enfoque estadístico basado en las distribuciones PH es desarrollado para analizar la variabilidad de las RRAM, siendo ésta uno de los aspectos clave a resolver. Tras un exhaustivo estudio experimental se muestra que las distribuciones PH funcionan mejor que cualquier otra distribución y además, ayudan a conocer mejor el comportamiento interno de las RRAM.

Se construye un nuevo proceso estocástico de macro-estados considerando el desempeño interno de los mismos. El tiempo de permanencia en cada uno de estos macro-estado se distribuye mediante una PH. Se muestra como el comportamiento interno del proceso es Markoviano, pero tanto la homogeneidad como la Markovianidad desaparecen para el nuevo modelo de macro-estados. También se obtienen otras medidas asociadas al modelo. La nueva metodología permite modelar sistemas complejos de forma algorítmica, en particular, el ruido producido dentro de las RRAM.

FPCA basado en la expansión de Karhunen-Loève permite describir la evolución estocástica de las RRAM. Sin embargo, es esencial identificar la distribución de las componentes principales (pc's) para modelizar todo el proceso. Para ello, se introduce una nueva clase de distribuciones, llamada distribuciones Tipo-fase Lineal (LPH). A partir de esta metodología se demuestra que, si las pc's siguen una distribución LPH, el proceso es caracterizado por una distribución LPH en cada punto.

En relación a las pc's, a veces su interpretación no es inmediata y se necesita aplicar una rotación para facilitarla. En este sentido, se desarrollan dos nuevos enfoques de rotación Varimax funcional basado en la equivalencia entre el FPCA y PCA. El primer método consiste en rotar los autovectores, mientras que el segundo rota las cargas de las puntuaciones de las pc's estandarizadas. Estas rotaciones son aplicadas para interpretar la variabilidad de las curvas de positivos por COVID-19 en las comunidades autónomas españolas.

Además, se proponen dos nuevos enfoques paramétricos y no paramétricos para resolver el problema de la homogeneidad funcional, asumiendo la expansión básica de las curvas. Estos métodos consisten en aplicar los test de homogeneidad multivariante sobre el vector de coeficientes básicos y sobre el vector de las puntuaciones de las pc's. Esta metodología ayudará a analizar qué influencia tienen el material y el grosor empleado en los procesos de fabricación sobre el funcionamiento de las RRAM.

Para el caso de más de una variable de respuesta funcional, se extiende la metodología anterior basada en el FPCA multivariante para probar la homogeneidad. En particular, se usa para comprobar si existen diferencias significativas entre los niveles de varios contaminantes según la localización geográfica de las estaciones de monitoreo en la Región de Abruzzo, Italia. Además, se considera un enfoque de medidas repetidas para estudiar si el nivel de cada contaminante se redujo durante el confinamiento establecido por el Gobierno Italiano durante la pandemia del COVID-19.

Finalmente, se propone un modelo de regresión múltiple función-sobre-función en términos de las pc's para la imputación de datos faltantes en una variable de respuesta funcional. Se asume que todos los predictores funcionales son completamente observados. Este método permitirá la imputación de datos faltantes relacionados con el COVID-19.

El contenido de esta tesis está presentado como un compendio de siete publicaciones. Las versiones completas de los artículos están incluidas en los Apéndices.

SUMMARY

In many scientific fields, it is usual to find magnitudes characterized by the evolution of a random variable over some continuum (stochastic process). Despite the experimental data measured on these variables are functions (curves, surfaces or images), historically their treatment has been through multivariate or time-series analysis, losing key information. Luckily, the great advances experimented by the technology sector in last years, have made easier the monitoring and reconstruction of the functions quickly and effortless, being possible to work with the complete functions. In this scenario, there is a high probability of having high dimensional data, in which the number of variables is greater than the number of sampling individuals. This fact makes that traditional statistical methods could not be appropriate. Depending on the final purpose, in this thesis these data are tackled from two different and complementary statistical perspectives: Functional Data Analysis (FDA) or Reliability Analysis (RA) based on Phase-type (PH) probability distributions.

FDA arose facing the need of building robust tools to model and predict functional data, whose observations are normally curves depending on time or any other continuous argument. In the last two decades, FDA has been subject of intensive research in which most multivariate techniques have been generalized, specially dimension reduction, regression and classification methods. Functional Principal Component Analysis (FPCA) stands out because reduces the dimension and explains the variability structure in terms of a small number of uncorrelated variables.

In the reliability field, one of the main objectives is to study the behaviour of complex systems, whose operation is conditioned by several uncontrollable variables. In this sense, RA attempts to identify the probability distribution of the data to shed light about the variability behind the systems operation. A suitable solution is to contemplate the Markovian processes and the PH distributions. This class is known to be able to approximate any non-negative distribution as much as desired thanks to its versatility and to model complex problems with well-structured results.

The methodological contributions of this thesis are elaborated in based to data-driven problems of great interest related to Resistive Random Access Memories (RRAMs) and COVID-19 pandemic. RRAMs awaken much expectation because

are one important source of incomes in the industry, whereas for mitigating the spread of the virus, it is crucial developing suitable models to make correct decisions

A new statistical approach based on PH distributions is developed to analyze the RRAM variability, which is one of the key issues to solve. A wide comparison with experimental data shows that the fitted PH distributions works better than the classic probability distributions and helps to know the RRAM internal performance.

A new stochastic process is built by considering the internal performance of macro-states in which the sojourn time is PH distributed. It is showed as the internal behaviour of the process is Markovian but both the homogeneity and Markovianity is lost for the new macro-state model. Other associated measures are also obtained. The new methodology allows the modeling of complex systems in an algorithmic way, in particular, the noise produced inside the RRAMs.

FPCA based on Karhunen-Loève expansion enables to characterize the stochastic evolution of RRAMs. Nevertheless, it is essential to identify the distribution of the principal components (pc's) to describe the entire process. In this sense, a new class of distributions, Linear PH (LPH) distributions, are introduced. Specifically, it was proved that if the principal components are LPH distributed then the process follows a LPH distribution at each point.

In relation to pc's, sometimes their interpretation is not immediate and a rotation is needed to facilitate it. We develop two new functional Varimax rotation approaches based on the equivalence between FPCA and PCA. One method consists of rotating the eigenvectors, and the other one, rotates the loadings of the standardized pc's scores. They are applied to interpret the variability of the positive cases curves of COVID-19 in the Spanish autonomous communities.

Additionally, two different parametric and non-parametric functional homogeneity testing approaches are proposed by assuming a basis expansion of sample curves. They consists of testing multivariate homogeneity on a vector of basis coefficients and on a vector of pc's scores, respectively. This fact will be useful to check the influence of the material and thickness in the RRAM behaviour.

For the case of more than one functional response variable, the previous methodology for testing homogeneity based on multivariate FPCA is extended. It is used to test if there are differences between the levels of several pollutants in terms of the location of measuring stations in the Region Abruzzo, Italy. Also, an approach for repeated measures is considered to study if the level of each pollutant decreased during the lockdown established by the Italian Government for COVID-19 pandemic.

Finally, a multiple function-on-function regression model in terms of pc's is proposed for the imputation of missing data for the functional response, by assuming that the multiple functional predictors are completely observed. This approach will enable to impute missing data related to COVID-19.

The content of this thesis are presented as a compendium of seven publications. The full version of the papers is included in the Appendices.

Quality indicators of the thesis

Appendix A1

The following reference is given in Appendix A1.

- C. Acal, J.E. Ruiz-Castro, A.M. Aguilera, F. Jimenez-Molinos and J.B. Roldán (2019). Phase-type distributions for studying variability in resistve memories. *Journal of Computational and Applied Mathematics*, vol. 345, pp. 23-32. DOI: `https://doi.org/10.1016/j.cam.2018.06.010`

Mathematics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	2.037	43/261	Q1

This work has contributed to the publication of the following papers not included in this thesis.

- E. Perez, D. Maldonado, C. Acal, J.E. Ruiz-Castro, F.J. Alonso, A.M. Aguilera, F. Jimenez-Molinos, Ch. Wenger, J.B. Roldan (2019). Analysis of the statistics of device-to-device and cycle-to-cycle variability in TiN/Ti/Al: HfO2/TiN RRAMs. *Microelectronic Engineering*, vol. 214, pp. 104-109. DOI: `https://doi.org/10.1016/j.mee.2019.05.004`

Optics			
JCR Year	Impact Factor	Rank	Quartile
2019	2.305	42/97	Q2

- C. Acal, J.E. Ruiz-Castro, A.M. Aguilera (2019). Distribuciones Tipo Fase en un estudio de fiabilidad. *TEMat*, vol. 3, pp. 63-74. DOI: `https://temat.es/articulo/2019-p63/`

- E. Perez, D. Maldonado, C. Acal, J.E. Ruiz Castro, A.M. Aguilera, F. Jimenez-Molinos, J.B. Roldan, Ch. Wenger (2021). Advanced temperature dependent statistical analysis of forming voltage distributions for three different HfO2-basesd RRAM technologies. *Solid-State Electronics*, vol. 176, pp. 107961. DOI: `https://doi.org/10.1016/j.sse.2021.107961`

Physics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	1.437	110/155	Q3

- J. E. Ruiz-Castro, C. Acal, A.M. Aguilera (2021). Phase-Type distributions computation aspects and applications in electronics. *Boletín de Estadística e Investigación Operativa*, vol. 37, num. 1, pp. 3-18.

Part of this study was presented in the conferences:

- 10th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2017).
  Title: A new statistical approach for modelling reset/set voltages in resistive memories.
  Mode of participation: Poster.
  Authors: C. Acal, J.E. Ruiz-Castro, A.M. Aguilera, F. Jiménez-Molinos, J.B. Roldán.
  Organizer: CFE-CMSTATISTICS.
  Celebration: London (UK), 16-18 December 2017.

- 23rd International Conference on Computational Statistics (COMPSTAT 2018).
  Title: Reliability analysis of switching parameters in resistive random access memories.
  Mode of participation: Poster.
  Authors: C. Acal, J.E. Ruiz-Castro, A.M. Aguilera, F. Jiménez-Molinos, J.B. Roldán.
  Organizer: International Association for Statistical Computing (IASC).
  Celebration: Iasi (Romania), 28-31 August 2018.

- X Jornadas de Usuarios de R.
  Title: Distribuciones tipo fase en memorias RRAM con R.
  Mode of participation: Poster.

Authors: C. Acal, J.E. Ruiz-Castro, A.M. Aguilera.
Organizer: University of Murcia and Association of R of Spain.
Celebration: Murcia (Spain), 22-23 November 2018.

- IX Jornadas de Enseñanza y Aprendizaje de la Estadística y la Investigación Operativa (GENAEIO).
  Title: Herramientas computacionales para el aprendizaje de las distribuciones Tipo Fase: Aplicación con datos reales de memorias resistivas.
  Mode of participation: Poster.
  Authors: C. Acal, J.E. Ruiz-Castro, A.M. Aguilera.
  Organizer: University of Granada and the Spanish Society of Statistics and Operations Research (SEIO).
  Celebration: Granada (Spain), 4-5 April 2019.

Appendix A2

The following reference is given in Appendix A2.

- J.E. Ruiz-Castro, C. Acal, A.M. Aguilera, J.B. Roldan (2021). A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories. *Mathematics*, vol. 9, num. 4, pp. 390. DOI: `https://doi.org/10.3390/math9040390`

Mathematics			
JCR Year	Impact Factor	Rank	Quartile
2019	1.747	28/235	Q1

Part of this study was presented in the following conference:

- 13th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2020).
  Title: Phase-type distributions in a vector Markov process to analyze random telegraph noise in resistive memories.
  Mode of participation: Invited talk.
  Authors: C. Acal, J.E. Ruiz-Castro, A.M. Aguilera, J.B. Roldán.
  Organizer: CFE-CMSTATISTICS.
  Celebration: London (UK), 19-21 December 2020.

Appendix A3

The following reference is given in Appendix A3.

- J.E. Ruiz-Castro, C. Acal, A.M. Aguilera, M.C. Aguilera-Morillo, J.B. Roldán (2021). Linear Phase Type probability modelling of functional PCA with applications to resistive memories. *Mathematics and Computers in Simulation*, vol. 186, pp. 71-79. DOI: `https://doi.org/10.1016/j.matcom.2020.07.006`

Mathematics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	1.620	68/261	Q2

Part of this study was presented in the following conferences:

- Mathematical and Computational Modelling, Approximation and Simulation, MACMAS 2019.
  Title: Phase-type probability modelling of functional PCA with applications to resistive memories.
  Mode of participation: Talk.
  Authors: C. Acal, J. E. Ruiz-Castro, A. M. Aguilera, M. C. Aguilera-Morillo and J. B. Roldán.
  Celebration: Granada (Spain), 9-11 September 2019.

- V Congreso de Jóvenes Investigadores de la RSME.
  Title: Un nuevo enfoque a la modelización tipo fase en Análisis de Datos Funcionales con aplicaciones en electrónica.
  Mode of participation: Invited talk.
  Authors: C. Acal, J. E. Ruiz-Castro, A. M. Aguilera.
  Organizer: The Royal Spanish Mathematical Society.
  Celebration: Castellón (Spain), 27-31 January 2020.

- IV Jornadas de Estadística como Herramienta Científica.
  Title: Distribuciones Tipo Fase Lineales para la Modelización Probabilística de Datos Funcionales en Electrónica.
  Mode of participation: Talk.
  Authors: C. Acal, J. E. Ruiz-Castro, A. M. Aguilera.
  Organizer: University of Jaen.
  Celebration: Jaen (Spain), 24-26 March 2021.

Appendix A4

The following reference is given in Appendix A4.

- C. Acal, A.M. Aguilera, M. Escabias (2020). New Modeling Approaches Based on Varimax Rotation of Functional Principal Components. *Mathematics*, vol. 8, num. 11, pp. 2085. DOI: `https://doi.org/10.3390/math8112085`

Mathematics			
JCR Year	Impact Factor	Rank	Quartile
2019	1.747	28/235	Q1

Appendix A5

The following reference is given in Appendix A5.

- A.M. Aguilera, C. Acal, M.C. Aguilera-Morillo, F. Jiménez-Molinos, J.B. Roldán (2021). Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Mathematics and Computers in Simulation*, vol. 186, pp. 41-51. DOI: `https://doi.org/10.1016/j.matcom.2020.05.018`

Mathematics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	1.620	68/261	Q2

Part of this study was presented in the following conferences:

- III International Workshop on Advances in Functional Data Analysis.
  Title: One-Way ANOVA modelling for RRAM reset curves.
  Mode of participation: Invited talk.
  Authors: C. Acal, M.C. Aguilera-Morillo, A.M. Aguilera, F. Jiménez-Molinos, J.B. Roldán.
  Organizer: The Spanish Society of Statistics and Operations Research (SEIO).
  Celebration: Castro Urdiales (Spain), 23-24 May 2019.

- Mathematical and Computational Modelling, Approximation and Simulation, MACMAS 2019.
  Title: Homogeneity problem for basis expansion of functional data with applications to resistive memories.
  Mode of participation: Talk.
  Authors: A. M. Aguilera, C. Acal, M. C. Aguilera-Morillo, F. J. Jiménez-Molinos and J. B. Roldán.
  Celebration: Granada (Spain), 9-11 September 2019.

Appendix A6

The following reference is given in Appendix A6. This work has been the result of a stay of three months at the University D'Annunzio of Chieti-Pescara (Italy).

- C. Acal, A.M. Aguilera, A. Sarra, A. Evangelista, T. Di-Battista, S. Palermi (2021). Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis. *Stochastic Environmental Research and Risk Assessment*, under revision.

Statistics & Probability			
JCR Year	Impact Factor	Rank	Quartile
2019	2.351	21/124	Q1

Appendix A7

The following reference is given in Appendix A7.

- C. Acal, M. Escabias, A.M. Aguilera, M. Valderrama (2021). COVID-19 data imputation by multiple function on function principal component regression. *Mathematics*, in press.

Mathematics			
JCR Year	Impact Factor	Rank	Quartile
2019	1.747	28/235	Q1

Part of this study was presented in the following conferences:

- I Jornada de Investigación Matemática en tiempos de Covid.
  Title: Aplicación de modelos de predicción en componentes principales para la imputación estadística de datos del Covid19.
  Mode of participation: Invited talk.
  Authors: M. Escabias, C. Acal, A.M. Aguilera, M. Valderrama.
  Organizer: The Math Institute of the University of Granada.
  Celebration: Granada (Spain), 10 June 2020.

- 19th Applied Stochastic Models and Data Analysis International Conference (ASMDA2021).
  Title: Covid-19 data imputation by Principal Component Prediction (PCP) models.
  Mode of participation: Invited talk.

Authors: M. Escabias, C. Acal, A.M. Aguilera, M. Valderrama.
Organizer: Applied Stochastic Models and Data Analysis International Society.
Celebration: Athens (Greece), 1-4 June 2021.

Contents

Chapter 1

Introduction

Nowadays, it is quite common to have high dimensional data associated with a great number of correlated variables, in which the sample sizes tend to be smaller than the number of variables. The classical statistical regression, classification and prediction methods are not usually efficient for these data on account of problems related to the sample size and overfitting. Data providing information about curves or more general functions that evolve over time, space or other continuous argument, are a particular case of high dimensional data, whose stochastic modeling belongs to the probabilistic theoretical framework of stochastic processes. This thesis aims to address two different analysis perspectives for this kind of processes, which might be complementary in many applications to study the variability associated with the analyzed random processes. On the one hand, we develop models for functional data analysis (Ramsay and Silverman, 2005), which will be estimated from discrete observations of the sample curves. On the other hand, we tackle the reliability analysis of complex systems based on the probability Phase-type distributions (Neuts, 1975; 1981) together with the Markovian Arrival Processes (Neuts, 1979).

Functional Data Analysis (FDA) comprehends a wide variety of statistical methods in which data can be described through functions, e.g., curves, surfaces or images. Historically, the treatment of this type of data has been carried out by means of multivariate approaches, since it is technically impossible to register complete curves in practice. In fact, the curves are discretely observed. FDA acquires body of doctrine at the end of the last century thanks to the publication of the first edition of the book entitled 'Functional Data Analysis' (Ramsay and Silverman, 1997). Ever since, the number of contributions to this field has gone through the roof from any area of knowledge. The main reason of this growth is due to the great computational advances that computers have suffered in last years. The increasing power of the computers enables the monitoring and analysis of large dataset coming from stochastic processes without too much effort. At this point, the advice is to analyze the complete behaviour of these trajectories, instead of working with the vector of discrete observations at different time points. Otherwise, essential information such as the continuity or smoothness of the curves could be lost. A broad revision about the most important FDA aspects from methodological and computational viewpoint, as well as, many examples of applications can be seen in Ramsay and Silverman (2002; 2005; 2009), Ferraty and Vieu (2006) and Horvath and Kokoszka (2012). On this matter, many authors have extended the classical multivariate statistical techniques to the field of FDA. Several of these techniques are canonical correlation analysis (Krzysko and Waszak, 2013; Keser and Kocakoç, 2015), cluster analysis (Tokushige et al., 2007; Jacques and Preda, 2014; Fortuna et al., 2018; Fortuna and Maturo, 2019; Caruso et al., 2021), discriminant analysis (Araki et al., 2009; Gorecki et al., 2014; Aguilera-Morillo and Aguilera, 2020), linear regression models (Aguilera et al., 1999), generalized linear models (James, 2002; Escabias et al., 2004; Muller and Stadtmuller, 2005; Escabias et al., 2014; Guo et al., 2015), ANOVA problem (Cuevas et al., 2004; Cuesta-Albertos and Febrero-Blande, 2010; Gorecki and Smaga, 2015, Zhang et al., 2019; Aguilera et al., 2020), variable selection (Gregorutti et al., 2015), classification (Galeano et al., 2015) or confidence intervals (Lian, 2012; Di-Battista and Fortuna, 2017). Even, the multidimensional layout (more than one functional variable) has been also considered in FDA (Benhenni et al., 2007; Tokushige et al., 2007; Gorecki and Smaga, 2017). Likewise, FDA is also strongly connected with longitudinal data analysis, when the information is measured on the same element in different periods of time or conditions (Davidian et al., 2004; Zhao et al., 2004; Martínez-Camblor and Corral, 2011). The main tool in this thesis is Functional Principal Component Analysis that is considered for many reasons the most predominant technique in the area of FDA.

FPCA can be seen as the natural extension of the multivariate Principal Component Analysis (PCA) for the case of a continuous time stochastic process (Deville, 1973; 1974). Among other advantages, it reduces the dimension of the problem and explores the main modes of variation in terms of a small set of uncorrelated variables, called principal components (pc's). In particular, FPCA based on Karhunen-Loève (K-L expansion) provides an orthogonal representation of a stochastic process, which can be approximated in terms of the most explicative components by truncating the K-L expansion. This finite-dimension representation is a key tool for using FPCA in the estimation of functional regression models. Thanks to its outstanding properties, FPCA has always been object of research. Dauxois et al. (1982) developed asymptotic theory and statistical inference on FPCA, meanwhile Ocaña-Lara et al. (1999) focused their attention on the study of FPCA when the metric is changed in the Hilbert space where the sample functions belong to. On the other hand, James et al. (2000) and Yao et al. (2005) introduced nonparametric models to perform FPCA when there is a small number of irregularly space observations for each sample curve. Hall and Hosseini-Nasab (2006) discuss how the properties of functional principal component analysis can be elucidated through stochastic expansions and related results. Additionally, different Bayesian approaches to FPCA were considered in Van der Linde (2008) and in Suarez and Ghosal (2017). Besides, FPCA for multivariate functional data has already been suggested and developed by several authors (Ramsay and Silverman, 2005; Berrendero et al., 2011; Jacques and Preda, 2014). An interesting and very useful result was given in Ocaña et al. (2007). If the basis expansion is taken into account to approximate the real form of curves, FPCA is reduced to a multivariate PCA of a transformation of the basis coefficients matrix. The basis expansion approach consists of assuming that curves belong to a finite-dimension space spanned by a basis. Normally, Fourier and B-spline bases (Kano et al., 2005; Aguilera and Aguilera-Morillo, 2013) are selected for periodic data and non-periodic data, respectively. In the meantime, Wavelet bases (Johnstone and Silverman, 1997; Chui, 2016; Liu et al., 2020) are used when derivatives are not required and curves have a strong local behaviour. Penalized estimation approaches were also developed to improve the estimation and interpretation of FPCA results (Silverman, 1996; Aguilera and Aguilera-Morillo, 2013). There are other possibilities in the literature, but they are not common in practice.

Functional regression models are also object of intensive research in recent years, and a very important part of the contributions of this thesis are related to them. The estimation of these models is an ill-posed problem that is usually solved by least squares penalized approaches and basis expansion of functional parameters and/or sample curves, reducing in many situation the functional model to a multivariate linear model in terms of the matrices of the basis coefficients of the response and predictors variables. Unfortunately, the interpretation is usually difficult because this multivariate model presents a high multicollinearity. A suitable solution is to turn the current problem into a linear regression on uncorrelated predictor variables. For that purpose, approaches based on FPCA (Chiou et al., 2004; Escabias et al., 2004; Muller and Stadtmuller, 2005; Aguilera-Morillo et al., 2013) or functional Partial Least Squares (Preda and Saporta, 2005; Escabias et al., 2007; Preda et al., 2007; Aguilera et al., 2010; Aguilera et al., 2016; Delaigle and Hall, 2017) have already studied for different functional regression models.

As we said at the beginning, in addition to FDA whose antecedents have already described, the reliability field also play a fundamental role in the current thesis. The main objective of the reliability (or survival) analysis is the modeling and optimization of systems to find certain efficiency and optimal cost. Specially, facing the presence of non-repairable failures that might produce big costs and irreparable damages. This statistical perspective has a wide scope of application, such as the medicine or electronic, computational and industrial engineering. In general terms, reliability analysis is in charge of studying the behaviour of systems, whose operation is conditioned by several uncontrollable variables. These variables provoke that systems are subjected themselves to a continuous deterioration. Another notable aspect is that the lifetime (or analogously, the failure time) is random among distinct experimental units. This means that the systems will not survive the same time, even though they are manufactured and run under the same conditions. Consequently, the branch of Statistics, and principally, the Probability Theory, play a fundamental role in the modeling of systems, given that the life or failure times will be able to be fitted by some probability distribution. For convenience, we are making reference to the time, but the reliability analysis can explore other variables that do not represent time, although it is true that they can be highly correlated with it. For instance, as it is displayed later, the operation of the resistive memories is based on the formation and rupture of a conductive filament, whose process depends on the supplied voltage. Here, the variable of interest is the voltage, but the process is operating concurrently a certain time.

The first probability distribution employed in the sector of reliability analysis was the exponential distribution (Epstein and Sobel, 1953). During many years, it has been considered as the distribution of reference thanks to its outstanding properties, simplicity and applicability. But as time went by, the exponential distribution became obsolete because it only models the behaviour of units that fail at constant rate, independently of the cumulated time. This situation is not very credible in practice. Since then, other distributions started to be applied such as Erlang, Weibull, Gamma and Log-Normal distribution, among others (McPherson, 2013). In the majority of reliability studies is not habitual to further beyond and these distributions are considered suitable to address any problem. Nevertheless, the development of new systems with internal structures more and more sophisticated causes that these distributions do not always achieve an accurate fitting. Therefore, it is easy to commit misinterpretation of the reality through the results. At this point, it makes sense to contemplate another approach that improves the quality of the study. Multiple complex models have been developed by introducing different aspects of interest. Some examples are multi-state systems, preventive maintenance, loss of units and multiple repairers. Traditional binary models have been generalized by multi-state models in order to better describe the evolution of systems that undergo different operation phases. The multi-state models were proposed at the middle of the 1970s and ever since, the methodological and applied results have been increasing (Andersen et al., 1993; Lisnianski and Levitin, 2003; Meira-Machado et al., 2009). The principal disadvantage of these models is the complexity of the expressions that are obtained in the modeling, which complicates the interpretation and applicability of results. In contrast, Markov models provide the opportunity to develop methodologically the modeling of complex systems in an algorithmic-matrix form, being the interpretation and the application of results much more simple. Semi Markov models have been contemplated for multi-state systems (Barbu et al., 2016; 2017), meanwhile recent advances in multi-state systems reliability can be consulted in Lisnianski et al. (2018). Likewise, Phase-type (PH) distributions come into play when absorbent Markov processes are considered.

PH distributions enable to model complex problems with well-structured results, thanks to its matrix-algebraic form. PH distributions, which were introduced by Neuts (1975; 1981), are defined as the distribution of the lifetime up to the absorption in an absorbing Markov process with finite state space. They constitute a class of non-negative distributions and fulfil the closure properties (maximum, minimum and addition). Besides, this class generalises a large number of known distributions such as exponential, Erlang or Coxian distribution, among others. That is, these distributions can be expressed with PH structure. However, one of the most important result related to PH distributions is given by Asmussen's theorem (Asmussen, 2000). The PH class is dense in the set of probability distributions on the non-negative half-line, so any non-negative probability distribution can be approximated as much as desired by a PH distribution. In relation to the fitting of their parameters by maximum likelihood, we resort to a recurring method called EM algorithm that alternates two steps for the estimation: expectation and maximization. This algorithm was developed by Asmussen et al. (1996) and assumed by Buchholz et al. (2014). The main aspects of PH distributions are discussed in depth by He (2014). Then, taking into account the great power of PH class, the PH distributions, together with the processes of Markovian arrivals, are considered in many areas of knowledge to make easier the study of complex systems (Ausin et al., 2004). For example, in queueing theory (Artalejo and Chakravarthy, 2006; Ramirez-Cobo et al., 2010) and in risk theory (Asmussen and Bladt, 1996) for continuous-time. Besides, the performance of discrete unitary multi-state systems with different types of maintenance (Ruiz-Castro, 2014; 2016), as well as redundant complex systems that evolve in discrete time (Ruiz-Castro and Quan-Lin, 2011; Ruiz-Castro, 2015) and the case of loss of units (Ruiz-Castro et al., 2018; Ruiz-Castro, 2021) have been analyzed by means of the PH class accompanied by the Markovian arrival processes. On the other hand, Ruiz-Castro (2020) and Ruiz-Castro and Dawabsha (2020) make use of PH distributions in order to model reliability complex systems, whereas Ruiz-Castro and Zenga (2020) analyze the breast cancer through them. Furthermore, Coxian distribution has been studied to uncover suitable fits in general problems from the survival perspective (Marshall and Zenga, 2012). In particular, this distribution has been considered for patient survival (Marshall and Zenga, 2009), to forecast elderly patient length of stay in hospital (Gordon et al., 2018) and to model students' length of stay at University (Marshall et al., 2013).

The challenge of this dissertation lies in developing models that will be used in order to tackle problems of great social impact based on the analysis and decision making (data-driven) from high dimensional data, through the methodological fields described above. Problems related to electronics and COVID-19 illness are principally addressed.

In the field of electronics, the current technology of non-volatile memories (those that do not need energy to run) is one of the most important sources of incomes throughout the world in the semiconductor industry. According to important consultants of this sector, the profits are constantly increasing year by year. For example, in 2013 the benefits overcame the 13 trillion of dollars in all the world. The main reason of this success is the massive sale of portable devices (tablets, smartphone, etc.) and the rise of Solid-State Drive as supports in the domestic computers. The attractive of RRAMs resides in the small size of the cells that form these memories. However, the reduction of their cells cannot be undefined and then, the industry is looking for other possibilities to improve the quality and performance of the gadgets. Among others, an option is the development of new devices that give a solution in short term and do not alter too much the current manufacturing processes. In this context, the Resistive Random Access Memories (RRAMs) have been stood as the best candidate to substitute the non-volatile memories employed to date. RRAMs have shown a fantastic potential (low power of operation, good scalability, fast speed, etc.) in the present CMOS technology (Waser y Aono, 2007; Waser, 2012; Lanza, 2014; Ielmini y Waser, 2015). The physical and internal properties have been studied by means of exhaustive experimental analysis (Tsuruoka et al., 2010; Long et al., 2013; Pan et al., 2014) and through simulation models (Villena et al., 2016; Aldana et al., 2017; Roldan et al., 2018). Recently, a new mathematical approach has been introduced in order to describe and simulate complex and thermochemical processes that occur in these devices (Mauri et al., 2015). Nevertheless, prior to the process of industrialization and commercialization of these memories, there exists a great needed to study the variability and noise behind the RRAMs' operation.

Regarding the variability, RRAMs operation is based on the stochastic nature of resistive switching processes that, in most cases, create (set process) and rupture (reset process) a conductive filament that changes drastically the device resistance (Pan et al., 2014; Lanza, 2014; Villena et al., 2017). Figure 1.1 shows the functioning principle of a RRAM: after an initial formation process of the conductive filament, the supply of voltage provokes the rupture of the filament to reach the High Resistance State, or rebuild again the filament to return to the Low Resistance State. These changes of resistance give raise to a sample of current-voltage curves corresponding to the reset-set cycles. The aforementioned variability is translated to different voltages and currents associated with the reset and set processes, as well as the behaviour of these curves up to the formation/rupture of the conductive filament. Several reset-set curves are displayed in Figure 1.2 (left panel). It is observable as the reset points are characterized by the sudden drop of the current (rupture of the conductive filament), whereas the set points are determined by the instant spike of the current (formation of the filament). These points lead to different reset-set voltages and currents, which are usually modeled by means of Weibull distribution (Long et al., 2012; Luo et al., 2012; Pan et al., 2014) in order to shed

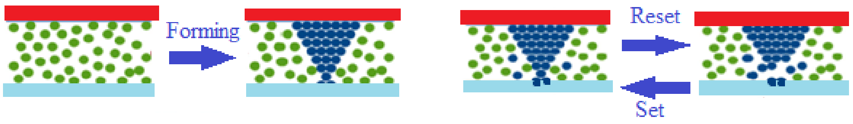


Figure 1.1: Schematic representation about the RRAM operation process.

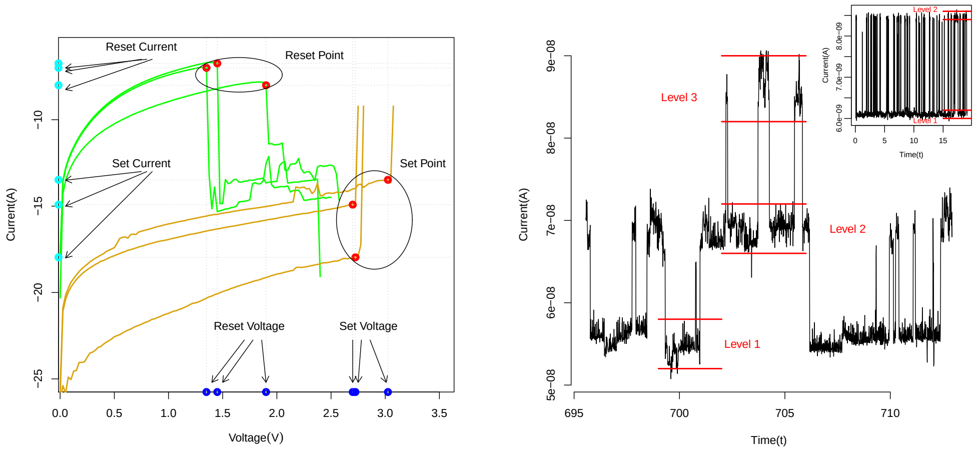


Figure 1.2: **Left panel)** Experimental current versus applied voltage for several reset (green lines) and set (orange lines) curves, including their corresponding reset-set points. **Right panel)** Current versus time trace for two RTN signal with different number of levels.

light about the statistical properties of the experimental data.

However, Weibull distribution does not work correctly for some resistive memories and then, other statistical approach is needed. Faced with this scenario, our motivation is to propose a new methodology based on PH distributions in order to improve the quality of the fitting. This new approach will be compared by means of experimental results with the classic probability distributions.

In relation to the noise, Random Telegraph Noise (RTN) is another significant issue to bear in mind before the massive industrialization of RRAM memories (Simoen and Claeys, 2017). RTN are defined as disturbances provoked by several traps that cause current fluctuations (Puglisi et al., 2015; Gonzalez-Cordero et al., 2019; 2020). Two different RTN signals can be seen in Figure 1.2 (right panel). This noise may affect the correct device operating in some applications related to neuromorphic hardware (Puglisi et al., 2018; Grasser, 2020; Alonso et al., 2021), but it can be also useful for other fields, for example, in cryptography (Chen et al., 2016). While it is true that most of the approaches to RRAMs consider the variability case; the RTN problem, despite its evident interest, is more rarely considered in the literature. For this reason, it is crucial to develop new methodologies that allow to control this kind of noise. From the mathematical viewpoint, the purpose is to describe the number of levels in the signal (see the marked levels in red in the right panel of Figure 1.2) and the sojourn time in each of these levels. Therefore, taking into account the stochastic nature of these traces (processes that evolves over time) and that the current fluctuations (levels) change randomly, a theoretical framework based on Markov chains can be contemplated as a suitable option in order to model and to study the evolution of the signals. In fact, some authors have already addressed the RTN problem through this approach (Puglisi, 2020). The problem in many applications is that the spent times in each level are not exponentially distributed, which is pretty fatal for Markov models.

In this work, the goal is to build a new model by assuming that the sojourn time is Phase-type distributed and that each level is formed by multiple internal states (no observables) whose behaviour is Markovian. Besides, the stationary distribution will be calculated through matrix-algorithmic methods, meanwhile the distribution of the number of visits to a determined macro-state will be given by considering a Laplace transform.

On the other hand, developing models capable of simulating the internal behaviour of RRAM memories is a challenging task both for the industry and for the academia in the area of semiconductor. The processes of simulation are fundamental for the introduction of a technology with new integrated electronic circuits. In this regard, it is very important to have compact methods to check if the devices are useful and work without failure prior to manufacture them. So far, several authors have proposed different complex options for RRAMs (Biolek et al., 2009; Shin et al, 2010; Jimenez-Molinos et al., 2015; Picos et al., 2015). Here, simulated data are obtained by using compact models that describe the electric current in terms of voltage through analytical equations. With the purpose of substituting the current complex methods, Aguilera-Morillo et al. (2019) conducted FPCA based on K-L expansion to model, explain and simulate the internal behaviour of RRAMs. This approach simplifies a lot the aforementioned complex models, given that the device current can be described only with few random parameters (principal components). Therefore, if the distribution of the components is known, the variability can be analyzed in an intuitive way by interpreting the parameters of the distribution. Moreover, curves like those of 1.2 can be simulated as well. In previous studies, some transformations have been considered to adjust different distributions successfully. However, to find the appropriate transformation and its probability distribution is not an easy task.

Our objective is to develop a new methodology based on PH distribution that could be accurately fitted for any transformation. At this regard, we will introduce a new class of distributions, the Linear Phase-type (LPH) to model the principal components. This class will be studied in detail to prove, through the K-L expansion, that certain linear transformations of the process at each point are PH distributed and then, the one dimensional distributions of the process will be modeled by the LPH distributions.

Regardless, the internal performance of these devices, and hence, their corresponding variability, could be influenced both for the type of material and for the conductive filament thickness used in the fabrication processes. From statistical viewpoint, the purpose is to decide if there are significant differences in the probability distribution that generates the set and reset processes associated with RRAMs fabricated making use of different materials and thicknesses. In other words, testing if several independent samples of curves come from the same population. Faced with the stochastic nature of the data measured on these memories, our aim is to introduce new functional parametric and non-parametric approaches to solve this homogeneity problem. Assuming the basis expansion of the curves, we will propose to carry out multivariate homogeneity tests on a vector of basis coefficients and, on the other hand, after applying FPCA, on a vector of principal component scores.

The theoretical framework of stochastic processes also have an important weight in the applications of data related to medicine. As you would expect, this dissertation is also focusing on the modeling of the evolution of the virus SARS-Cov-2 and its impact in other areas such as environment. At the end of January 2020, the World Health Organization declared worldwide public health emergency due to the rapid propagation of the virus. Ever since, the impact of the pandemic has been really devastating, where the number of deceases continues toward the rise in the whole world (Dong et al., 2020). Even though sanitary and economic crisis is the main worry of institutions, other areas of society have been also compromised: education (Torres-Martín et al., 2021), politics (Greer et al., 2020) and environment (Zambrano-Monserrate et al., 2020), among others. Thus, mitigating the virus incidence is the principal challenge of the governments. For that purpose, the Committee of Experts of each country analyzes the daily data of the pandemic in order to establish the best measures that help to recuperate as soon as possible the people's normal life.

These decisions are substantiated on the results produced by mathematical (statistical) models. The better the predictive power of the models, the better they will be able to know the response of the virus in the future. For this reason, the scientific community is dedicating all its time to the development of tools capable of modeling and predicting properly the behaviour of the pandemic. Normally, the number of confirmed cases, deaths, recovered, hospitalized and in intensive care unit people are usually considered for these models. An interesting comparison between Spain and Italy before and after their respective national lockdowns by means of quasi-Poisson regression is carried out in Tobias (2020). Berihuete et al. (2020) introduce a new Bayesian indicator to predict the start of a new outbreak. Additionally, Mora et al. (2020) propose semi-empirical models based on the logistic map with the aim of forecasting the evolution of these variables in distinct stages of the pandemic in Spain. Likewise, SIR models are applied in Agarwal and Jhajharia (2021) to analyze the trend of the illness over the world and more specifically, in India. Besides, these variables can be also addressed from the field of artificial intelligence through deep learning methods (Zeroual et al., 2020). On the other hand, Qi et al. (2020) conducted a generalized additive model in order to check if the number of cases in 30 Chinese provinces is connected with the daily average temperature and relative humidity. In this sense, the conclusion about if there are relationship between the spreading of the virus and certain environmental conditions may be influenced by the selection of the spatiotemporal model (Briz-Redon, 2021). In view of the functional nature of the variables, they have been also tackled from the FDA viewpoint. FPCA is performed in Tang et al. (2020) to analyze the COVID-19 data in the United State, meanwhile in Carroll et al. (2020) different functional tools are used in order to model cumulative curves of COVID-19 positive cases across countries. A multivariate FDA approach can be seen in Torres-Signes et al. (2021) to forecast the number of deaths in Spain.

In the face of the effort that Spanish Mathematics Committee requested for the development of mathematical techniques for the fight against COVID-19, we propose to use FPCA in order to explain the different modes of variability in the evolution of number of cases in each Spanish autonomous community. However, since it is common that the first principal component explains a high percentage of the total variability because all communities had a growing behaviour, the objective is to develop new ways of Varimax rotation to make easier the interpretation of the results. In particular, we will focus on the equivalence between FPCA and PCA of a transformation of the matrix of basis coefficients. As a natural extension of the multivariate case, our proposal is to make the rotation on the eigenvectors or on the loadings of the standardized principal component scores. The first one will preserve the orthogonality between the eigenfunctions but rotated principal component scores will be correlated each other, whereas the opposite scenario will take place for the second method (uncorrelatedness and non-orthogonality). This two new approaches are complementary to the ways of functional Varimax rotation proposed by Ramsay and Silverman (2005), where the rotation is made on the weigh function coefficients and on the weight function values. In both cases, the rotated component scores are correlated but the rotated eigenfunctions are still orthonormal.

On the other hand, decisions to restrict the propagation of the virus have had important effects on the air quality. In this sense, several authors revealed that the air pollution level has been decreased across the world thanks to lockdown measures adopted by the governments (Agarwal et al., 2020; Mahato et al., 2020; Berman and Ebisu, 2020). In this dissertation, we will analyse through functional methods the impact of quarantine policies on air quality in the district of Pescara-Chieti (Italy). The hourly average evolution of concentrations for four air pollutants in two different periods (pre and during lockdown) for different monitoring stations are available in this study. Some recent studies for environmental data from the FDA perspective can be seen in Park et al. (2013), Escabias et al. (2013), Hormann et al. (2015), Aguilera-Morillo et al. (2017) and Gautam and Trivedi (2020). Firstly, we will extend functional ANOVA techniques for repeated measures to evaluate if the level of each pollutant is different between the considered periods. Secondly, the purpose is to build a theoretical development for multivariate functional ANOVA for independent measures to compare jointly the level of all pollutants according to the localization of the monitoring stations. This methodology will be based on the multivariate FPCA and will consists of testing multivariate homogeneity on the vectors of the most explicative principal component scores.

Anyway, all these models require all input data must be validated in order to avoid bias in the estimations. In other words, data must be complete and with the sufficient quality to reach rigorous predictions. Nevertheless, these assumptions are barely fulfilled during a pandemic. Specially, it is common to have incomplete data due to changes in the way of registering data or because governments do not supply information some days, for instance, at weekends. Little and Rubin (2019) and Graham (2012) study in depth the imputation problem for multivariate data. For the functional case, He et al. (2011) propose a novel approach for multiple imputation in a longitudinal data context. Rao and Reimherr (2021) analyze distinct imputation methods for sparse and irregular functional data settings. When the response variable is not functional but the predictors have functional character, scalar-on-function regression can be applied (Ferraty et al., 2013; Ling et al., 2015; Ling et al., 2016; Crambes and Henchiri, 2019; Febrero-Bande et al., 2019). In our case, we will consider the situation where the response variable is functional and there are multiple functional predictors (function-on-function models). The idea is to use forecast models based on principal components regression for the statistical imputation of COVID-19 data in order to avoid problems associated with the multicollinearity.

Summarizing, the main objective of this thesis is to provide important advances for modeling high dimensional data with applications in areas of high social impact such as engineering, environment or medicine. In particular, in Appendix A1, a new approach based on Phase-type distributions is proposed in order to model the reset/set voltages and currents associated with RRAM processes when the Weibull distribution does not achieve an accurate fitting. In Appendix A2, we introduce a novel macro-state stochastic process whose sojourn time in each macro-state is Phase-type distributed. This model is used to study the performance of RTN signals when the internal behaviour of levels is not observable. In Appendix A3, a new class of distributions called Linear Phase-type distributions is introduced and its properties deeply studied. After considering the K-L expansion to describe the stochastic evolution of curves, the purpose is to identify the distribution of the principal components as well as to identify the distribution of the own process. This methodology enables to simulate as many reset/set curves of RRAMs as we desire. Two new Functional Varimax rotation approaches are presented in Appendix A4 to make easier the interpretation of principal components when almost all variability falls on one or two principal components. These criterions have been applied in order to understand the evolution of infections by COVID-19 in the Spanish autonomous communities during the first wave of the pandemic. In Appendix A5, two different parametric and non-parametric homogeneity testing approaches are proposed by assuming a basis expansion of the sample curves. The motivation is to check if the kind of material and thickness employed for the fabrication of RRAMs influence in their performance. Likewise, this methodology is extended for multivariate data case in Appendix A6 to study the differences between the temporal evolution of four pollutants in terms of the location of the monitoring stations (traffic or background stations) situated in Abruzzo Region (Italy). Furthermore, the statistics available in the literature for repeated measures have also been extended considering the basis expansion of the curves to study whether the level of each of the pollutants decreased during the lockdown period after Italian Government declared the home confinement at the middle of March because of COVID-19 pandemic. In Appendix A7, the extension of the function-on-function linear regression model to the case of multiple functional predictors is proposed for estimating the curves of hospitalized and intensive care unit people in terms of confirmed, deceased and recovered curves, during the first outbreak of COVID-19 in Spain.

Chapter 2

Objectives

2.1 Phase-type distributions for studying variability in resistive memories

Before the manufacturing process and commercialization of the *Resistive Random Access Memories* (RRAM), it is crucial to analyze the variability behind the RRAM operation. A great amount of experimental analysis have been carried out in order to study the physical and internal properties of these devices. Furthermore, mathematical models capable of describing and simulating their behaviour have been also developed. A crucial aspect is related with the analysis of voltages and currents (also resistances) related to the processes of formation (set process) and rupture (reset process) of a conductive filament (Pan et al., 2014; Lanza, 2014; Villena et al., 2017). The common statistical analysis performed on experimental data is through Weibull distribution (Luo et al., 2012; Pan et al., 2014; González-Cordero et al., 2016). The interpretation of its parameters allows us to understand the performance of these memories better.

However, there are many situations where the Weibull distribution does not work adequately. A new approach based on Phase-type (PH) distributions (Neuts, 1981) to improve the quality of the modeling is proposed in the current manuscript. In particular, we fit the reset voltages through PH distributions, which will help us to study the intermediate states of degradation in the process of destruction of the conductive filament. An exhaustive comparison will be also conducted to test if the introduced methodology reaches a better fitting than the classic procedure based on Weibull distribution.

2.2 A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories

In many applications, such as electronics and computing engineering, the target is to study the random temporal evolution of complex devices with several performance levels (macro-states), being possible the existence of internal phases in each one. Despite the macro-states are visible, the internal states are not. In this point, it is of great interest to analyze the internal behaviour among levels in order to better understand the transitions structure and unfolding. For this purpose, it is common to contemplate Markov processes, but in many situations the sojourn time in each level does not follow the exponential distribution, which is a crucial issue in this field. Faced with this scenario, the main objective of this work is to construct a new stochastic process by considering the internal perfomance of macro-states for which the sojourn time is Phase-type distributed and only macro-states can be observed. In addition, measures associated with this new stochastic process (e.g., stationary distribution and number of visits to a certain macro-state) will be determined by means of Markovian Arrival Processes (He, 2014).

This work is motivated through the need to explore the variability patterns of different Random Telegraph Noise signals associated with RRAM memories (Puglisi et al., 2015; Gonzalez-Cordero et al., 2019; 2020). From the mathematical viewpoint, these signals can be seen as stochastic processes where the level of electric current (current fluctuations) changes randomly. So far, several attempts based on classical Markov chains have been carried out (Puglisi, 2020). However, the sojourn time in each state is not exponentially distributed when the RTN signal is large enough. On this matter, the developed methodology is applied here. This novel perspective will shed more information about the internal performance of these gadgets. Besides, Hidden Markov Models (Rabiner, 1989) are proposed for substituting the graphical techniques used in the sector to compute the number latent levels hidden into the process. Afterwards, multiple previous studies are conducted to determine the number of phases in each macro-state.

2.3 Linear Phase-Type probability modelling of functional PCA with applications to resistive memories

A stochastic process can be represented by Functional Principal Component Analysis in terms of the Karhunen-Loève expansion. This approach enables to reduce the dimension of the problem and to describe the main stochastic characteristics related to multiple systems using a small set of uncorrelated random variables called principal components. Nevertheless, the process will not be characterized entirely until the probability distribution of the principal components is not identified. Unfortunately, finding a suitable distribution is not an easy task in practice because the classic probability distributions do not always achieve a rigorous fitting. In order to solve this handicap, a new approach based on Phase-type (PH) distributions is introduced in the current work. Thanks to the good properties of PH distributions and that any non-negative distribution can be approximated as needed through a PH distribution, it is expected to be able to identify the distribution of the principal components and thus, characterize the whole process for any given situation. Taking into account that principal components could take negative values, the main goal of this work is to introduce a new class of distributions, called Linear Phase-type distributions, to characterize the distribution of the stochastic process through the LPH distribution of the principal components. This class of distributions will be studied in detail to demonstrate that certain linear transformations of the process at each any time point are Phase-type distributed.

The motivation of this work is to provide a novel solution to an existent problem in the RRAM context. Aguilera-Morillo et al. (2019) modeled the reset curves by means of the K-L expansion. This procedure enabled to describe satisfactorily the main internal characteristic of these devices and awakened a great interest from the circuit simulation viewpoint thanks to its simplicity. Regrettably, it is essential to know the distribution of the principal components to simulate the associated stochastic process. Aguilera-Morillo et al. (2019) made a first attempt to fit the distribution of the principal components, but without too much success. Under this outlook, LPH distributions will be fitted in order to try to achieve a more accurate probability modelling of the principal components of the set/reset curves.

2.4 New Modeling Approaches Based on Varimax Rotation of Functional Principal Components

Functional Principal Component Analysis is for many reasons a key technique in the functional data framework. Among other good properties, FPCA reduces the dimension of the problem and explores the main features characterizing a functional variable in terms of a small set of uncorrelated variables. Regarding the description of the dependence structure, there are occasions where the principal components can not always be interpreted straightforward. In some occasions the problem lies in the lack of smoothness, which can be solved by means of penalizing the roughness of the weight functions (Silverman, 1996; Cardot, 2000; Aguilera and Aguilera-Morillo, 2013). In other applications, the difficulty of the interpretation is due to one or two components explain a very high percentage of variability. A usual way of solving this problem consists of rotating the weights functions in order to make easier the interpretation. In this regard, the main goal of this work is the proposal of two new functional rotation approaches.

Although there are different options to carry out the rotation in multivariate analysis, the most used method in practice is the orthogonal rotation, and in particular, the Varimax criterion. This technique has its origin in Factor Analysis. Jolliffe (2002) makes a complete review of Varimax criterion in PCA. From a functional viewpoint, Ramsay and Silverman (2005) proposed to apply Varimax rotation in two different ways: the first one is based on Varimax rotation of the matrix of values of the weight functions in a grid of equally spaced time points, meanwhile the second one consists of applying the Varimax method on the matrix of basis coefficients of weigh functions. Both procedures provide that the rotated principal component scores are no longer uncorrelated anymore.

In this work, we propose two new techniques for rotation of FPCA based on the equivalence between FPCA and PCA (Ocaña et al., 2007). The FPCA is equivalent to PCA of a certain transformation of the matrix of basis coefficients. The first method consists of the rotation of the eigenvectors. Then, the eigenfunctions remain orthogonal but the rotated component scores are not uncorrelated. The second approach involves a rotation of the loadings of the standardized principal component scores. This approach guarantees that rotated scores are uncorrelated but the orthogonality among eigenfunctions is lost. We carried out a simulation study to analyze the performance of these approaches by comparing the outcomes with Ramsay and Silverman's methods. Besides, an application with data related to the number of infected by COVID-19 during the first wave in Spain is conducted in order to analyze the evolution of the pandemic in the country.

2.5 Homogeneity problem for basis expansion of functional data with applications to resistive memories

The functional homogeneity problem consists of deciding if several independent samples of curves have been generated by the same stochastic process. If all independent samples fulfil the normality assumption, the homogeneity problem is known as one-way analysis of variance for functional data (FANOVA). This popular technique has as objective to check the hypothesis of the equality of several mean functions coming from independent groups. Cuevas et al. (2004), Shen and Faraway (2004), Ramsay and Silverman (2005), Delicado (2007) or Zhang (2014) tackle this problem by a broad variety of methods. An interesting comparison of tests for the one-way functional ANOVA problem can be seen in Gorecki and Smaga (2015). Flores et al. (2018) propose two new sample tests for homogeneity based on the concept of functional depth measures.

In this article, we focus on the fact that the FANOVA model is equivalent to multivariate ANOVA (MANOVA) when the basis expansion of the curves is considered. This means that the analysis is reduced to apply the MANOVA tests on a vector of basis coefficients of the sample curves. However, two problems appear for this theoretical framework. On the one hand, there are many situations where the samples curves are not generated by a Gaussian process and therefore, MANOVA tests can not be conducted. On the other hand, it is well-known that the multivariate tests do not work well when the dimension of the problem is high. For the first issue, some solutions lie in making use of bootstrap/permutation tests, but for the second matter there are not answers yet. Thus, the main objective of this work is to provide new solutions to these problems based on basis expansion of the sample curves. In order to solve the lack of normality, we propose multivariate non-parametric homogeneity tests (Oja, 2010) on the matrix of basis coefficients. To reduce the dimension of the problem, a novel methodology based on Functional Principal Component Analysis is introduced. This procedure consists of testing homogeneity on the vector of the most explicative principal component scores by means of parametric or non-parametric tests depending on the nature of the sample curves.

This methodology is motivated with the purpose of detecting if there are significant physical differences between RRAM technologies considering different materials and thicknesses. That is, if the type of material or thickness employed in the fabrication of these memories affects to the internal switching operation, whose behaviour is modeled by a functional variable. The two proposed approaches are carried out to achieve this goal. The performance of these procedures will be also tested through an extensive simulation study.

2.6 Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis

This work concerns the functional ANOVA when more than one functional response variable is available in the analysis. Despite its great interest, the multivariate functional perspective is rarely considered in the literature. As far as we are aware, only Gorecki and Smaga (2017) deal the multivariate ANOVA for functional data. They developed permutation tests based on a basis function representation and tests based on random projections. Here, we propose to extend the parametric and nonparametric approaches introduced in the previous section, by considering the multivariate FPCA.

Additionally, the one-way functional ANOVA problem for the case of repeated measures (the information is collected for the same subjects in different conditions or periods of time) is also tackled in the current paper. In particular, a basis expansion approach for the statistics proposed by Martinez-Camblor and Corral (2011) and Smaga (2020) to test the equality of two mean functions is considered.

Both methodologies will be applied to study the impact of quarantine policies on air quality in the district of Pescara-Chieti (Italy). For that purpose, hourly average measurements collected by the Regional Agency for the Environmental Protection of Abruzzo about four air pollutants concentrations have been considered in the analysis. These data were measured in two different periods of time, pre-lockdown and during lockdown, for different monitoring stations, which are the tools established to measure and manage the compliance with national ambient air quality standards. Then, the objective is to ascertain whether the level of each pollutant has changed during the lockdown period and to assessing the differences between the temporal evolution of all pollutants in terms of the location of measuring stations (traffic and background stations).

2.7 COVID-19 data imputation by multiple function on function principal component regression

Faced with the need to find models capable of modeling and predicting the evolution of the worldwide pandemic provoked by COVID-19 illness, both governments and institutions are investing enormous amount of money in order to provide the best possible equipment and tools to the scientific community. Although there are many factors of interest to gauge the situation of the pandemic in a country, the main researches are focused on the treatment of variables such as number of positive, recovered and deceased cases, as well as the hospital occupancy rate measured by the number of hospitalized people and in intensive care units. As the observed data are curves, different FDA aproaches have been developed (Tang et al., 2020; Torres-Signes et al., 2021). The inherent problem is that during an epoch of pandemic as in which we live, it is not common to have complete and high quality data. This is an essential requirement for the models to be able to provide accurate results. Then, a mechanism based on FDA for the imputation missing data is proposed in the current work.

The inspiration of this work is the imputation of missing values after a modification in the way of registering data in hospitalized and intensive care curves by some Spanish autonomous communities during the first wave of COVID-19. For that purpose, we propose to apply functional linear regression for the imputation of the missing functional responses (curves of hospitalized and intensive care people) in order to have complete data and to use the predictive models with guarantees.

In particular, the extension of the function-on-function linear regression models (Valderrama et al., 2010; Aguilera et al., 2015; Qi and Luo, 2018) is proposed in this paper for the case of multiple functional predictors (curves of positives, deaths and recovered people). The functional parameters of this model are estimated in terms of principal components regression with the completely observed data. Finally, once all curves are properly homogenized, recorded and smoothed so that they can be comparable, the missing data are imputed and the relationship between the hospital occupancy rate and the illness response variables is analyzed through a canonical correlation analysis in terms of principal components.

Chapter 3

Methodology

Before proceeding with the specific methodology of each work, we are going to introduce some basic theoretical aspects considered along this thesis.

3.1 Theoretical framework

3.1.1 Stochastic processes

In the real life it is common to find many systems that evolve over time. The queue of customers at a service station, spread of a pandemic or movement of a gas molecule are some examples of phenomena whose behaviour is varying over time. Depending on whether the evolution of these events is considered random or certain beforehand, stochastic or deterministic models can be used, respectively, in order to make predictions about their state in the future. Systems are not inherently stochastic or deterministic, rather to model an event as stochastic or deterministic may be subject to the choice of the observer. One of the most important differences between both approaches is that deterministic models predict an outcome with absolute certainty, whereas stochastic models provide only the probability of an outcome (Allen, 2010). In the case of stochastic processes, the underlying mathematical methodology is based on probability theory. On the other hand, the nature of the time variable can be discrete or continuous. For the discrete case, the processes are observed at a discrete set (numerable) of instants, meanwhile the continuous processes are observed constantly over time. In this last situation, the evolution of the variable is described by a continuous-time stochastic process. This thesis is developed in the theoretical framework of the continuous stochastic processes (Todorovic, 1992; Taylor and Karlin, 1994; Ross et al., 1996).

Formally, it is well-known that given $T \subset \mathbb{R}$ an interval of the real line, a continuous stochastic process is described as a family of non-numerable random variables $\{X(t) : t \in T\}$ defined on the same probabilistic space $(\Omega, \mathcal{A}, P)$. We focus on the stochastic processes in which the random variables $X(t)$ are real.

3.1.1.1 Basic hypotheses

A stochastic process can be seen as a random variable with values in a functional space denoted by H. We assume that this space has structure of Hilbert space (Young, 1988; Berberian, 1999), given a separable Hilbert space $(H, \langle, \rangle_H)$, a random variable on H is defined as a measurable function

$$\begin{aligned} X:\ & \Omega \to H \\ & \omega \to X(\omega), \end{aligned}$$

such that $X^{-1}(\mathcal{B}) \in \mathcal{A}$, being $\mathcal{B}$ a Borel set of the Borel σ-algebra generated by the space H.

In this thesis we will focus on observations coming from variables of a continuous stochastic process whose trajectories belong to the functional space $L^2[T]$ of integrable square functions on T defined by

$$L^2(T) = \left\{ f : T \to \mathbb{R} : \int_T f^2(t)dt < \infty \right\},$$

with the usual scalar product

$$\langle f, g \rangle = \int_T f(t)g(t)dt, \ \ \forall f, g \in L^2(t).$$

Moreover, let us consider that $L^2(\Omega)$ is the space of real random variables X on Ω with finite second order moments. Then, a stochastic process (random function) X is second order if satisfies

$$E[||X||^2] = \int_\Omega ||X(\omega)||^2 dP(\omega) < \infty, \ \ \forall X \in L^2(\Omega),$$

with $|| \bullet ||$ being the norm associated to the Hilbert space in which we consider the random variable. Associated with a second order stochastic process, the following elements which play a fundamental role throughout this dissertation are defined:

- Mean function

$$\begin{aligned} \mu :\ & T \to \mathbb{R} \\ & t \to \mu(t) = E[X(t)] = \int_\Omega X(t, \omega) dP(\omega). \end{aligned}$$

- Covariance function

$$
\begin{aligned}
C:\ & T \times T \to \mathbb{R} \\
& (t,s) \to C(t,s) = E[(X(t)-\mu(t))(X(s)-\mu(s))] \\
& \qquad = \int_{\Omega} [(X(t,\omega)-\mu(t))(X(s,\omega)-\mu(s))]dP(\omega).
\end{aligned}
$$

- Covariance operator

$$
\begin{aligned}
\mathcal{C}:\ & L^2(T) \to L^2(T) \\
& f \to \mathcal{C}(f)(t) = \int_T C(t,s)f(s)ds.
\end{aligned}
$$

Another interesting definition is the continuity in quadratic mean. A stochastic process is continuous in quadratic mean if

$$\lim_{h \to 0} E[(X(t+h)-X(t))^2] = 0,\ \forall t \in T.$$

The property of continuity in quadratic mean of a process guarantees the continuity of its covariance function (Todorovic, 1992). This fact is crucial because many of the employed functional techniques require the continuity of covariance function in $T \times T$. For instance, it allows to obtain the spectral decomposition of the covariance operator which is key in Functional Principal Component Analysis.

Hereinafter, we assume that $\{X(t) : t \in T\}$ is defined on a probabilistic space $(\Omega, \mathcal{A}, P)$ and that the following hypotheses are true:

H_1: The proces is second order.

H_2: The process is continuous in quadratic mean.

H_3: The sample trajectories belong to the Hilbert space $L^2[T]$ of squared integrable functions with the usual inner product.

3.1.1.2 Continuous-time Markov processes

Let us assume that we have a stochastic process $\{X(t) : t \geq 0\}$ with state space denoted by $\mathcal{S}$. Then, $\{X(t) : t \geq 0\}$ is a continuous-time Markov process if it is verified that

$$P[X(t_{m+1}) = x_{m+1}|X(t_1) = x_1, \ldots, X(t_m) = x_m] = P[X(t_{m+1}) = x_{m+1}|X(t_m) = x_m],$$

for any $0 \leq t_1 < \ldots < t_m < t_{m+1}$, possible states $x_1, x_2, \ldots, x_{m+1} \in \mathcal{S}$ and for any $m \geq 0$ (Kijima, 2013; Kulkarni, 2016). This means that given the current state, the rest of the past is irrelevant to forecast the future. Besides, the process is homogeneous if for $0 \leq s < t$ and $i, j \in \mathcal{S}$

$$p_{ij}(t) = P[X(s+t) = j|X(s) = i] = P[X(t) = j|X(0) = i],\ \forall s, t.$$

By considering these values, we can construct the transition probability matrix $\mathbf{P}(t) = (p_{ij}(t))$. This matrix verifies the following properties:

1. $p_{ij}(t) \geq 0\ \forall i, j$ and any time t.
2. $\mathbf{P}(t)\mathbf{e} = \mathbf{e}$, being $\mathbf{e}$ a column vector of ones with an appropriate order.
3. $\mathbf{P}(s+t) = \mathbf{P}(s)\mathbf{P}(t)$, Chapman-Kolmogorov equation.
4. $\lim_{t\to 0} \mathbf{P}(t) = \mathbf{I}$, being $\mathbf{I}$ the identity matrix with an appropriate order.

There are important elements related to the chain: the initial distribution and the transient distribution. The initial distribution represents the probability of being in the state i at the start of the process and it is denoted by $p_i(0) = P[X(0) = i],\ \forall i \in \mathcal{S}$. The second term makes reference to the probability of occupying a state at time t. It is denoted by $p_i(t) = P[X(t) = i],\ \forall i \in \mathcal{S}$ with $\sum_i p_i(t) = 1$. Likewise, the rates of jumps between states are calculated through the derivative at the origin of the transition probabilities:

- For $i \neq j$

  $$p'_{ij}(0) = \lim_{h\to 0} \frac{p_{ij}(h) - p_{ij}(0)}{h} = \lim_{h\to 0} \frac{p_{ij}(h)}{h} = q_{ij},$$

  being q_{ij} the transition rate from the state i to the state j.

- For $i = j$

  $$p'_{ij}(0) = \lim_{h\to 0} \frac{p_{ij}(h) - p_{ij}(0)}{h} = \lim_{h\to 0} \frac{p_{ij}(h) - 1}{h} = q_{ii},$$

  with $q_{ii} < 0$ and being $q_i = -q_{ii}$ the exit rate from the state i.

Besides, the sojourn time in each state is distributed as an exponential distribution with parameter q_i for state i. These results imply that the probability of jumping to state j after staying a determined exponential time at the state i is $p_{ij} = q_{ij}/q_i$, if $q_i \neq 0$. Otherwise, the state would be absorbent. Finally, we define the following matrix of order $m \times m$ in order to sum up the performance of the chain

$$\mathbf{Q} = \begin{pmatrix} -q_1 & q_{12} & \cdots & q_{1m} \\ q_{21} & -q_2 & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & -q_m \end{pmatrix}.$$

The matrix $\mathbf{Q}$ is called infinitesimal generator matrix (Q-matrix) which verifies that its non-diagonal elements are non-negative, its diagonal elements are negatives or zero and the sum of elements of each row is equal to zero (conservative matrix). The Q-matrix and the initial distribution identify the Markov process.

Absorbing Markov processes

A Markov chain is called an absorbing Markov process when it is composed by a series of transient states and, at least, one absorbing state (Buchholz, 2014). Note that a state $i \in \mathcal{S}$ is a transient state if the probability of returning to i is lower than 1. Also, an state $i \in \mathcal{S}$ is an absorbing state if the chain does not change the state once the state i is reached, i.e., the transition probability from i to j is 0 with $i \neq j$, being $q_i = 0$ in this case.

Let $\{X(t) : t \geq 0\}$ be a continuous-time Markov process with finite transient state space $E = \{1, \ldots, m\}$ and one absorbing state $m + 1$. This process will be absorbed by the state $m + 1$ with probability equal to one and its infinitesimal generator can be expressed by matrix blocks as

$$\mathbf{Q} = \left(\begin{array}{c|c} \mathbf{T} & \mathbf{T}^0 \\ \hline \mathbf{0} & 0 \end{array} \right),$$

being $\mathbf{T}_{m\times m}$ the matrix that contains the intensities between transient states and $\mathbf{T}^0_{m\times 1}$ the column vector that represents the exit transition rate from each transient state to the absorbent state. Additionally, the row vector $\mathbf{0}$ is due to the fact that it is impossible to leave the absorbing state and 0 is the transition rate out off $m+1$ state.

3.1.2 Phase-type distributions

A Phase-type probability distribution (PH) with representation $(\boldsymbol{\alpha}, \mathbf{T})$ is defined as the distribution of the time up to the absorption of a continuous-time Markov process with m transient states and one absorbing state $m+1$ (Neuts, 1975; 1981). The 2-tuple $(\boldsymbol{\alpha}, \mathbf{T})$ represents the initial distribution and the transition intensities matrix among transient states, respectively, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ and α_i being the probability of finding itself initially in phase i. Here, it is crucial to emphasise that $\boldsymbol{\alpha}$ is a substochastic vector of order m and $\mathbf{T}$ is a subgenerator of order m. Besides, we assume that α_{m+1} is equal to zero (therefore, we assume that alpha is a probability distribution). Thus, a non-negative random variable T is phase-type distributed if its cumulative distribution function is given by the following expression

$$F(t) = 1 - \boldsymbol{\alpha} e^{\mathbf{T}t}\mathbf{e} = 1 - \boldsymbol{\alpha} \left(\sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{T}^n \right) \mathbf{e}, \;\; t \geq 0.$$

From the cumulative distribution function the density, the density function is given by

$$f(t) = \boldsymbol{\alpha} e^{\mathbf{T}t}\mathbf{T}^0, \;\; t \geq 0,$$

where $\mathbf{T}^0 = -\mathbf{T}\mathbf{e}$. On the other hand, the reliability function of T is

$$R(t) = 1 - F(t) = \boldsymbol{\alpha} e^{\mathbf{T}t}\mathbf{e}, \;\; t \geq 0,$$

so that the cumulative hazard rate is determined by

$$H(t) = -\ln(R(t)) = -\ln(\boldsymbol{\alpha} e^{\mathbf{T}t}\mathbf{e}), \;\; t \geq 0.$$

Therefore, the hazard rate is

$$h(t) = \frac{f(t)}{R(t)} = \frac{\boldsymbol{\alpha} e^{\mathbf{T}t}\mathbf{T}^0}{\boldsymbol{\alpha} e^{\mathbf{T}t}\mathbf{e}}, \;\; t \geq 0.$$

Obviously, there exists other functions that are usually considered in reliability and survival studies or in the area of queueing theory, e.g., availability function, MTTF (Mean Time Total Failure), etc. But here, we describe the functions that are more interesting for the applications carried out in this memory.

PH distributions have extraordinary features that make them very attractive from the modeling viewpoint. The main characteristics of this class of distributions can be seen in detail in He (2014). However, it is worth highlighting that PH distributions enable the application and interpretation of the results in a simple way as well as to express the main associated measures in algorithmic form. Likewise, the PH class is closed under a series of operations such as maximum, minimum and addition for independent variables. Furthermore, as we said in the introduction of this thesis, several classical probability distributions are special cases of PH distribution, that is, they can be dealt with PH structure. Below, we show the PH representation of several probability distributions commonly used in practice. They can be obtained directly by using the aforementioned definitions.

1. Exponential distribution

$$F(t) = 1 - e^{-\lambda t}, \;\; t \geq 0: \;\; \boldsymbol{\alpha} = 1, \;\; \mathbf{T} = -\lambda \text{ and } m = 1.$$

2. Erlang distribution $F(t) = 1 - \sum_{j=0}^{m-1} e^{-\lambda t}(\lambda t)^j / j!$ for $t \geq 0$, $m \geq 1$ and $\lambda > 0$,

$$\boldsymbol{\alpha} = (1, \ldots, 0), \;\; \mathbf{T} = \begin{pmatrix} -\lambda & \lambda & & \\ & -\lambda & \ddots & \\ & & \ddots & \lambda \\ & & & -\lambda \end{pmatrix}_{m \times m}.$$

3. Hypo-exponential distribution $F(v) = 1-\sum_{X=0}^{v}\sum_{i=1}^{m}\lambda_i e^{-\lambda_i X}\left(\prod_{\substack{j=1\\ j\neq i}}^{m}\frac{\lambda_j}{\lambda_j-\lambda_i}\right)$ for $v \geq 0$ and $\lambda_i \neq \lambda_j$ for $i \neq j$ with $\lambda_i, \lambda_j > 0$,

$$\boldsymbol{\alpha} = (1, \dots, 0), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & & \\ & -\lambda_2 & \ddots & \\ & & \ddots & \lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

4. Hyper-exponential distribution $F(v) = 1 - \sum_{i=1}^{m}\alpha_i(1-e^{-\lambda_i v})$ for $v \geq 0$ and $\lambda_i > 0$,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

5. Coxian distribution for $\lambda_i > 0$ with $i = 1, \dots, m$ and $0 < g_j \leq 1$ with $j = 1, \dots, m-1$,

$$\boldsymbol{\alpha} = (1, \dots, 0), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & g_1\lambda_1 & & \\ & -\lambda_2 & g_2\lambda_2 & \\ & & \ddots & g_{m-1}\lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

6. Generalized Coxian distribution for $\lambda_i > 0$ with $i = 1, \dots, m$ and $0 < g_j \leq 1$ with $j = 1, \dots, m-1$,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & g_1\lambda_1 & & \\ & -\lambda_2 & g_2\lambda_2 & \\ & & \ddots & g_{m-1}\lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}_{m\times m}.$$

In particular, the modeling studies carried out with microelectronic experimental data in this thesis converge to the Erlang distribution. In this regard, the mean time in each state can be computed by $1/\lambda$, and the mean time from the beginning up to the absorption as m/λ. Besides, the associated variance would be m/λ^2.

Nevertheless, one of the main results in this area is the theorem proposed by Asmussen (2000). He demonstrated that the PH class is dense in the set of probability distributions defined on the non-negative half-line. Taking this theorem into account, it is possible to approximate as much as desired whatever non-negative distribution by means of a Phase-type distribution.

3.1.3 Basic tools for Functional Data Analysis

The field of Functional Data Analysis aims to analyze sample of functions instead of vectors in which multivariate analysis is based. In this thesis, we will focus on the case in which data are curves obtained as trajectories of a stochastic process (functional variable) verifiying the hypotheses H_1, H_2 and H_3 aforementioned. In particular, let us assume that $x_1(t), \ldots, x_n(t)$ are realizations of i.i.d. stochastic processes $X_1(t), \ldots, X_n(t)$ with the same distribution that $X(t)$.

From sample curves, the following functions are defined:

- Sample mean function

$$\bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t), \ \forall t \in T.$$

- Sample variance function

$$\sigma_x^2(t) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i(t) - \bar{x}(t))^2, \ \forall t \in T.$$

- Sample covariance function

$$\hat{C}(s,t) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t)), \ \forall s, t \in T.$$

These functions are unbiased and consistent estimators that converge almost surely to the corresponding population moments (Deville, 1974).

3.1.3.1 Basis expansion approach

One of the biggest problems that we find when working with functional data is the fact that it is not usual to have the explicit expression of the sample paths. Instead of that, we count on a series of curves observed at discrete time on a finite set of time instants $\{t_{i0}, t_{i1}, \ldots, t_{im_i} \in T\}$ $\forall i = 1, \ldots, n$, that could be different for each sample curve. Because of this fact, the first step in FDA is to reconstruct the functional form of the curves. The main approaches are based on non-parametric techniques (Ferraty and Vieu, 2006) or basis expansions for the sample curves (Ramsay and Silverman, 2002; 2005). The last perspective is considered here, which consists of assuming that sample curves belong to a finite-dimension space spanned by a basis $\{\phi_1(t), \ldots, \phi_p(t)\}$. Hence, the sample processes can be expressed as follows

$$X_i(t) = \sum_{k=1}^{p} a_{ij}\phi_j(t) = \boldsymbol{a}_i^T \boldsymbol{\phi}(t), \;\; i = 1, ..., n, \tag{3.1}$$

where a_{ij} are the basis coefficients of the reconstruction with $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{ip})^T$ and $\boldsymbol{\phi} = (\phi_1(t), \ldots, \phi_p(t))^T$. Note that p must be sufficiently large to ensure a rigorous precision. Another concern topic is the suitable choice of the basis depending on the nature of the curves. Ramsay and Silverman (2002, 2005) make a wide review about different kinds of bases and ways to proceed, whereas Ramsay et al. (2009) detail how to do the computation with the software R and Matlab. The most useful basis systems are Fourier functions for periodic data, B-Spline basis when non-periodic paths are smooth enough and wavelets basis for curves with a strong local behaviour. Due to the fact that B-Spline bases will be considered throughout this dissertation, we briefly summarized them now. More information about the study of spline functions and their implementation by means of computational algorithms with B-Splines bases can be seen in Green and Silverman (1994), De Boor (2001) and Kano et al. (2005, 2011).

Splines of order $q+1$ consists of piecewise polynomials of degree q which are softly joint at a set of knots. Besides, their derivatives are continuous up to degree $q-1$ on the knots. B-Splines of certain degree generate the splines of the same degree and have no the problem called *boundary effect* (curve rapidly spreads toward zero away of the domain). This fact is common in many types of smoothers such as Kernel smoothers. Theoretically speaking, let us consider that $\tau_0 < \ldots < \tau_r$ is a partition of knots of the observation interval T. In order to give a formal definition, we can add more knots to the B-Splines basis, being now the partition $\tau_{-q} < \ldots < \tau_{-2} < \tau_{-1} < \tau_0 < \ldots < \tau_r < \tau_{r+1} < \tau_{r+2} < \ldots \tau_{r+q}$. At this point, they can be computed through the following iterative method:

$$B_{j,1}(t) = \begin{cases} 1 & \tau_{j-2} \leq t < \tau_{j-1} \\ 0 & \text{otherwise} \end{cases} \;\;, \;\; j = -1, 0, 1, \ldots, r+4$$

$$B_{j,q+1}(t) = \frac{t - \tau_{j-2}}{\tau_{j+q-2} - \tau_{j-2}} B_{j,q}(t) + \frac{\tau_{j+q-1} - t}{\tau_{j+q-1} - \tau_{j-1}} B_{j+1,q}(t)$$

$$q = 1, 2, \ldots ; \; j = -1, 0, 1, \ldots, r - q + 4.$$

From here on out, cubic B-splines will be used in this document ($q = 3$).

Depending on whether the sample curves are measured with or without error, the basis coefficients can be computed by smoothing or interpolation methods, respectively (Aguilera et al., 1995; Aguilera et al., 1996; Ramsay and Silverman, 1997; Valderrama et al., 2000). In particular, if curves are observed with error and B-Splines are chosen as suitable basis, there are different approaches to estimate the basis coefficients: regression splines, smoothing splines and penalized splines (P-Splines). In the former case, the coefficients are computed by Ordinary Least Squares without penalization, meanwhile in the last two methods, the coefficients are obtained by Penalized Least Squares. The continuous penalty for smoothing splines measures the roughness of a function by means of the integrated squared second derivative, whereas in P-Splines, the penalty is discrete based on the differences of certain order between adjacent coefficients. A comparative study of regression splines and smoothing splines can be consulted in Durban (2009). Durban (2013) analyzes how to deal with P-Splines in different models. Aguilera and Aguilera-Morillo (2013) carried out a complete comparison of different types of penalized smoothing with B-Splines basis. Techniques based on regression splines and penalized splines have been taking into account in the current document thanks to their remarkable perfomance with the curves that we have.

Regression splines

Let us suppose that the sample curves are expressed as in Equation (3.1) and a basis of B-splines is considered. This method consists of obtaining the basis coefficients by minimizing the least squares error

$$MSE(\boldsymbol{a}_i|x_i) = (x_i - \boldsymbol{\Phi}_i\boldsymbol{a}_i)^T(x_i - \boldsymbol{\Phi}_i\boldsymbol{a}_i),$$

with $\boldsymbol{\Phi}_i = (\phi_j(t_{ik}))_{m_i \times p}$. Then, the basis coefficients are given by

$$\hat{\boldsymbol{a}}_i = (\boldsymbol{\Phi}_i^T\boldsymbol{\Phi}_i)^{-1}\boldsymbol{\Phi}_i^T x_i.$$

Therefore, fitted curves can be expressed as follows

$$\hat{x}_i(t) = \hat{\boldsymbol{a}}_i^T\boldsymbol{\phi}(t), \ \ i = 1, ..., n.$$

Penalized splines

In a similar manner to regression splines and under the same assumptions, we should minimize the following penalized least squares criterion in order to calculate the basis coefficients

$$PMSE_d(\boldsymbol{a}_i|x_i) = (x_i - \boldsymbol{\Phi}_i\boldsymbol{a}_i)^T(x_i - \boldsymbol{\Phi}_i\boldsymbol{a}_i) + \lambda\boldsymbol{a}_i^T\boldsymbol{P}_d\boldsymbol{a}_i,$$

where λ is the smoothing penalty parameter and $\boldsymbol{P}_d = (\Delta^d)^T \Delta^d$ with Δ^d being the matrix representation of the d-order difference operator. Then, the basis coefficients for each curve are estimated by

$$\hat{\boldsymbol{a}}_i = (\boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i + \lambda \boldsymbol{P}_d)^{-1} \boldsymbol{\Phi}_i^T x_i.$$

An important concern about P-Splines is to determine which is the best decision for the order of the penalty and the smoothing parameter. The usual advice is to apply a quadratic penalty and to use cross-validation for computing the smoothing parameter. Regarding the number of knots, a good choice would be to take approximately one knot for every four observations till around 40 knots as maximum (Ruppert, 2002).

3.1.3.2 Functional Principal Component Analysis

Functional Principal Component Analysis is the natural generalization of multivariate PCA when the available sample information is a set of sample curves coming from a continuous-time stochastic process (Deville, 1974). Its main goal is to reduce the dimension of the problem and to explain the main modes of variation in terms of a small set of uncorrelated variables called functional principal components.

Let us consider that we have a sample of functions from $\{X(t) : t \in T\}$ denoted by $X_1(t), \ldots, X_n(t)$. We will assume that the process is centered without loss of generality. The pc's are zero-mean variables computed as uncorrelated generalized linear combinations of the process with maximum variance. Hence, the j-th principal component score is defined as

$$\xi_{ij} = \int_T X_i(t) f_j(t) dt, \;\; i = 1, \ldots, n,$$

where $f_j(t)$ are the weight functions or loadings. These functions are obtained by maximizing the following objective function

$$\begin{cases} \max_f \; \text{var}[\int_t X_i(t) f(t)] dt \\ \text{r.t. } ||f||^2 = 1 \text{ and } \int f_l(t) f(t) dt = 0, \; l = 1, \ldots, j-1. \end{cases}$$

Then, the loadings are the solution to the eigenequation

$$\hat{\mathcal{C}}(f_j)(t) = \int \hat{C}(t,s) f_j(s) ds = \lambda_j f_j(t),$$

with $\hat{\mathcal{C}}(f_j)(t)$ being the sample covariance operator, $\hat{C}(t,s)$ the sample covariance function and $\{\lambda_j\}$ a decreasing sequence of non null eigenvalues such that $\lambda_j =$

var$[\xi_j]$. The sample covariance function can be expressed as follows by assuming the hypotheses H_1, H_2 and H_3:

$$\hat{C}(s,t) = \sum_{j=1}^{n-1} \lambda_j f_j(s) f_j(t),$$

which provides the following orthogonal representation (Karhunen-Loève expansion) of sample curves:

$$X_i(t) = \sum_{j=1}^{n-1} \xi_{ij} f_j(t);\ i = 1, \dots, n.$$

This representation is optimal because it is the best approximation of the sample curves in the least squares sense. Besides, this principal component decomposition can be approximated by truncanting in terms of the first q principal components as follows:

$$X_i^q(t) = \sum_{j=1}^{q} \xi_{ij} f_j(t),$$

whose explained variance is given by $\sum_{j=1}^{q} \lambda_j$.

A noteworthy result was developed in Ocaña et al. (2007). Keeping in mind the basis expansion of the sample curves, they proved that the FPCA is equivalent to multivariate PCA of matrix $A\Psi^{1/2}$, being $A = (a_{ij})_{n\times p}$ the matrix of basis coefficients and $\Psi = (\Psi_{ij})_{p\times p} = \int_T \phi_i(t)\phi_j(t)dt$ the matrix of inner product between basis functions. In this framework, it is possible to express the principal component weight function f_j in terms of the basis expansion as well, i.e.

$$f_j = \sum_{k=1}^{p} b_{jk}\phi_k(t) = \boldsymbol{b}_j^T \boldsymbol{\phi}(t),$$

where $\boldsymbol{b}_j = \Psi^{-1/2}\boldsymbol{v}_j$, with $\boldsymbol{v}_j$ being the solutions to the eigenvalue problem $n^{-1}\Psi^{1/2}A^T A\Psi^{1/2}\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j$. Let us observe that $n^{-1}\Psi^{1/2}A^T A\Psi^{1/2}$ is the sample covariance matrix and the functional pc's scores are the multivariate principal components scores of matrix $A\Psi^{1/2}$.

Multivariate FPCA

The PCA of a functional variable can be extended to the case of a vector of functional variables defined on the same probabilistic space. Let us consider a set of curves $X_{ih}(t)$ with $i = 1, ..., n$; $h = 1, ..., H$ obtained as observations of a multivariate functional variable $(X_1, X_2, \dots, X_H)$. Then, the information for each subject is recorded in a vector denoted by $\boldsymbol{X}_i(t) = (X_{i1}(t), ..., X_{iH}(t))^T$. Besides, we assume that $\boldsymbol{X}_i(t)$ are i.i.d. multivariate functional variables with mean vector $\boldsymbol{\mu} = (\mu_1(t), ..., \mu_H(t))^T$ and sample matrix covariance function $\hat{\mathbf{C}}$ such that $\hat{\mathbf{C}}(t,s) = (\hat{C}_{h,h'}(t,s))$, $t,s \in \mathcal{T}$ and $h,h' = 1,...,H$. Note that if $h = h'$, then $\hat{C}_{h,h}$ is the covariance function and otherwise, that is $h \neq h'$, $\hat{C}_{h,h'}$ represents the cross-covariance function.

Ramsay and Silverman (2002) discussed in detail the bivariate FPCA. When there are more than two response variables, the j-th principal component scores are determined by

$$\xi_{ij} = \int_{\mathcal{T}} (\boldsymbol{X}_i(t) - \boldsymbol{\mu}(t))^T \boldsymbol{f}_j(t)dt = \sum_{h=1}^{H} \int_{\mathcal{T}} (X_{ih}(t) - \mu_h(t)) f_{jh}(t)dt,$$

where $\boldsymbol{f}_j(t) = (f_{j1}(t), ..., f_{jH}(t))^T$ are the vector of weight functions that maximizes the variance restricted to $\sum_{h=1}^{H} \int_{\mathcal{T}} f_{jh}(t) f_{j'h}(t)dt = 1$ if $j = j'$ and 0 otherwise. These functions are obtained as the solutions to the eigenequation system

$$\hat{\mathcal{C}} \boldsymbol{f}_j = \lambda_j \boldsymbol{f}_j,$$

with $\hat{\mathcal{C}}$ being the covariance operator and the sequence $\{\lambda_j\}_{j \geq 1}$ of positive real eigenvalues decreasing to zero indicating the amount of variance attributable to each component.

Hereinafter, we assume that $\boldsymbol{\mu}(t) = \mathbf{0}$. Then, the process can be expressed in terms of the K-L expansion

$$\boldsymbol{X}_i(t) = \sum_{j=1}^{n-1} \xi_{ij} \boldsymbol{f}_j(t),$$

which can be truncated by considering the first q principal components as

$$\boldsymbol{X}_i^q(t) = \sum_{j=1}^{q} \xi_{ij} \boldsymbol{f}_j(t).$$

Multivariate FPCA can be estimated through the basis expansion of the curves (Jacques and Preda, 2014; Schmutz et al., (2020)). Briefly, if the basis expansion is considered, the curves can be expressed as

$$\boldsymbol{X}_i(t) = \boldsymbol{\Phi}(t)\mathbf{a}_i^T,$$

where the basis coefficients are gathered as $\mathbf{a}_i = (a_{i11},\ ..., a_{i1p_1}, ..., a_{iH1}, ..., a_{iHp_H})$

with p_h being the number of basis functions for the h-th response variable and

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

In general $\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{\Phi}(t)^T$, where $\boldsymbol{A}$ is the resultant matrix after joining by row all $\mathbf{a}_i$. The spectral decomposition of the covariance operator C becomes

$$\boldsymbol{\Phi}(s)\Sigma_A \boldsymbol{W}\boldsymbol{b}_j^T = \lambda_j \boldsymbol{\Phi}(s)\boldsymbol{b}_j^T,$$

with Σ_A being the covariance matrix of $\boldsymbol{A}$, $\boldsymbol{b}_j$ being a row-vector that contains the basis coefficients of $\boldsymbol{f}_j(t) = \boldsymbol{\Phi}(t)\boldsymbol{b}_j^T$ and $\boldsymbol{W}$ being the matrix of inner products between basis functions with dimension $\sum_{h=1}^{H} p_h \times \sum_{h=1}^{H} p_h$. Since the presented spectral decomposition is true for all s, the expression can be reduced as $\Sigma_A \boldsymbol{W}\boldsymbol{b}_j^T = \lambda_j \boldsymbol{b}_j^T$. Now, by considering $\boldsymbol{v}_j = \boldsymbol{b}_j \boldsymbol{W}^{1/2}$, the multivariate FPCA is equivalent to the multivariate PCA of the matrix $\boldsymbol{A}\boldsymbol{W}^{1/2}$, whose covariance matrix can be diagonalized as ${\boldsymbol{W}^{1/2}}^T \Sigma_A \boldsymbol{W}^{1/2}\boldsymbol{v}_j^T = \lambda_j \boldsymbol{v}_j^T$.

3.2 Phase-type distributions for studying variability in resistive memories

In the field of microelectronics, the academia usually goes to graphical methods in order to estimate the Weibull distribution parameters (Luo et al., 2012; Long et al., 2013; Pan et al., 2014). Graphical method is a parametric technique based on the principle of least squares (Lawless, 2003). It is highly employed thanks to its simplicity and because it enables us to have a first graphical idea about the quality of the fitting. Briefly, this technique consists of constructing a point cloud from observed experimental data and fitting a straight line according to the least squares criterion. The final form of the cloud will depend on the considered probability distribution. Specifically, Deaves and Lines (1997) presented the graphical method to find the Weibull distribution parameters. A revision of this technique for others distributions, as well as an exhaustive study of simulation to prove the power of PH distributions against the classic ones can be seen in Acal et al. (2019).

Graphical techniques are not enough to estimate the parameters when PH distributions are considered. In fact, the process of estimation in this class of distributions is not a simple topic because they are highly redundant in general. Here, we must use the recursive method called EM algorithm to find the maximum likelihood estimate of the parameters of an underlying distribution from given trace data

(Asmussen et al., 1996; Buchholz et al., 2014). EM algorithm alternates two steps: Expectation and Maximization. The first step (E-step) determines the expectation of likelihood function through the inclusion of latent variables as if they were observables. The second one (M-step) computes the maximum likelihood estimators of the parameters by maximising the expected likelihood function obtained in E-step. The found parameters in M-step are used to start the following E-step and so on. The computational aspects of the estimation in this field, through statistical programmes such as R or Matlab, are revised in Ruiz-Castro et al. (2021).

In our application we assume that the variable of interest is the reset voltage (voltage where the conductive filament is broken) instead of time. Obviously, in a more general context of reliability analysis, the voltages are usually replaced by the failure times. Let $v_1, \dots, v_n$ be a sequence of reset voltages. From theoretical viewpoint, these points can be considered as the voltages up to the absorption associated with an absorbing Markov process. Besides, we suppose that the set $\{v_1, \dots, v_n\}$ are the values of n independent replications of a random variable distributed by a Phase-type distribution with representation $(\boldsymbol{\alpha}, \mathbf{T})$. EM algorithm optimises the following likelihood function

$$L(\boldsymbol{\alpha}, \mathbf{T}) = \prod_{i=1}^{m} \boldsymbol{\alpha}_i^{N_i} \prod_{i=1}^{m} e^{x_i \mathbf{T}_{ii}} \prod_{i=1}^{m} \prod_{\substack{j=1 \\ j \neq i}}^{m+1} \mathbf{T}_{ij}^{N_{ij}},$$

being x_i the total time spent in state i, N_i the number of times that the process began in phase i and N_{ij} the number of jumps between both states. If the current estimation of PH is $(\boldsymbol{\alpha}, \mathbf{T})$, the conditional expectations of x_i, N_i and N_{ij} (step-E) adopt the following expressions

$$\begin{aligned}
E_{(\boldsymbol{\alpha},\mathbf{T})}[x_i] &= \frac{1}{n} \sum_{k=1}^{n} \frac{\left[\int_0^{v_k} \left(\boldsymbol{\alpha} e^{\mathbf{T}(v_k-u)}\right)^T \left(e^{\mathbf{T}u}\mathbf{T}^0\right)^T du\right]_{ii}}{\boldsymbol{\alpha} e^{\mathbf{T}v_k}\mathbf{T}^0}, \\
E_{(\boldsymbol{\alpha},\mathbf{T})}[N_i] &= \frac{1}{n} \sum_{k=1}^{n} \frac{\boldsymbol{\alpha}_i (e^{\mathbf{T}v_k}\mathbf{T}^0)_i}{\boldsymbol{\alpha} e^{\mathbf{T}v_k}\mathbf{T}^0}, \\
E_{(\boldsymbol{\alpha},\mathbf{T})}[N_{ij}] &= \frac{1}{n} \sum_{k=1}^{n} \frac{\left[\int_0^{v_k} \left(\boldsymbol{\alpha} e^{\mathbf{T}(v_k-u)}\right)^T \left(e^{\mathbf{T}u}\mathbf{T}^0\right)^T du\right]_{ii} \mathbf{T}_{ij}}{\boldsymbol{\alpha} e^{\mathbf{T}v_k}\mathbf{T}^0},
\end{aligned}$$

$$E_{(\boldsymbol{\alpha},\mathbf{T})}[N_{i,m+1}] = \frac{1}{n} \sum_{k=1}^{n} \frac{\left(\boldsymbol{\alpha} e^{\mathbf{T}v_k}\right)_i \mathbf{T}_i^0}{\boldsymbol{\alpha} e^{\mathbf{T}v_k}\mathbf{T}^0}.$$

Therefore, the M-step results in the estimation of new parameters

$$\hat{\boldsymbol{\alpha}}_i = E_{(\boldsymbol{\alpha},\mathbf{T})}[N_i] \quad ; \quad \hat{\mathbf{T}}_{ij} = \frac{E_{(\boldsymbol{\alpha},\mathbf{T})}[N_{ij}]}{E_{(\boldsymbol{\alpha},\mathbf{T})}[x_i]},\ i \neq j;$$

$$\hat{\mathbf{T}}_i^0 = \frac{E_{(\boldsymbol{\alpha},\mathbf{T})}[N_{i,m+1}]}{E_{(\boldsymbol{\alpha},\mathbf{T})}[x_i]} \quad ; \quad \hat{\mathbf{T}}_{ii} = -\left(\hat{\mathbf{T}}_i^0 + \sum_{\substack{j=1\\ j\neq i}}^{m} \hat{\mathbf{T}}_{ij}\right).$$

3.3 A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories

A new macro-state stochastic process is developed in this work in order to model complex systems formed by different levels (macro-states) whose sojourn time in these levels does not follow the exponential distribution. The new model is built in transient and stationary regimes.

The Model

- We assume that the stochastic process $\{X(t) : t \geq 0\}$ is composed of r macro-states. In turn, each macro-state is composed of multiple states, that is, the macro-state k is formed by n_k internal states. We assume that there is an embedded internal process denoted by $\{J(t) : t \geq 0\}$ which is a Markov process with the following generator matrix expressed by blocks

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1k} & \cdots & \mathbf{Q}_{1r} \\ & \ddots & \vdots & \ddots & \\ \vdots & \cdots & \mathbf{Q}_{kk} & \cdots & \vdots \\ & \ddots & \vdots & & \ddots \\ \mathbf{Q}_{r1} & \cdots & \mathbf{Q}_{rk} & \cdots & \mathbf{Q}_{rr} \end{pmatrix}.$$

- We consider the following two matrices: $\mathbf{Q}_k$ represents the transition rate to the macro-state k; $\mathbf{Q}_{-k}$ contains the output velocity to any macro-state, except to macro-state k. Clearly, $\mathbf{Q} = \mathbf{Q}_k + \mathbf{Q}_{-k}$.

- Due to the internal process verifies the properties of Markovianity and homogeneity, it is direct to compute the transition probability matrix and the transient distribution at time t for the process $\{J(t) : t \geq 0\}$:

  - Transition probability matrix. $\Rightarrow \mathbf{P}(t) = \exp\{\mathbf{Q}t\}$.
  - Transient distribution at time t. $\Rightarrow \mathbf{a}(t) = \boldsymbol{\theta}\mathbf{P}(t)$, being $\boldsymbol{\theta}$ the initial distribution.

- As it is shown in Appendix A2.2.1, the transition probabilities for the macro-state process $\{X(t) : t \geq 0\}$ is obtained through the Markovian process $\{J(t) : t \geq 0\}$. Knowing the probability of being in the macro-state $\boldsymbol{i}$ at time s, the transition between macro-states $\boldsymbol{i} \to \boldsymbol{j}$ is expressed in an algorithmic-matrix form.

Stationary Distribution

- The objective is to resolve the system given by the balance equation and the normalization equation that are verified by the stationary distribution in a matrix and algorithmic form by using matrix-analytic methods. All steps for the resolution of the system and an algorithm to calculate the stationary distribution can be seen in Appendix A2.2.2.

- We achieve the stationary distribution for the macro-state process by means of the internal matrices blocks from the stationary distribution for the embedded process.

Sojourn Time Phase-Type Distribution

One of the most important aspect related to the new process is to know the probability distribution for the sojourn time in each macro-state. Although for the Markov process $J(t)$, the sojourn time in each internal state is exponentially distributed, this fact is not corroborated for $X(t)$.

- In Appendix A2.3.1, it is proved that the probability distribution of the random sojourn time in each macro-state is Phase-type distributed depending on the initial observed time.

- Additionally, we show that if the macro-state process has reached the stationary regime and the process is in the macro-state $\boldsymbol{i}$, then, the probability function of the sojourn time is also Phase-type distributed.

- The first step time for the process $X(t)$ from one macro-state to another macro-state follows a Phase-type distribution.

Number of Visits to a Macro-State

Another interesting topic is the number of visits to a determine macro-state.

- For its calculus, we have consider a matrix $\mathbf{p}_k(n,s,t)$, whose element (i,j) represents the probability that the embedded process is in the internal state j at time t and has visited n times the macro-state k from s up to time t, knowing that it was in the internal state i at time s.
- This matrix verifies the differential equations described in Appendix A2.4 and it is computed by means of the inverse of Laplace transform.
- Once this matrix is obtained, we have determined the number of visits to a certain macro-state. It can be seen in A2.4.

Expected Number of Visits to a Determined Macro-State

Regarding the expected number of visits to a determined macro-state, by considering the definition of the expected value and keeping in mind the differential equations, there are two different ways to compute the mean number of visits to the macro-state k depending on whether the initial state is considered or not:

- For the case where the initial state is not counted
$$E[N_k(t)] = \boldsymbol{\theta} \cdot \int_0^t \mathbf{P}(u)du \left(\mathbf{Q}_k - \tilde{\boldsymbol{Q}}_{kk}\right) \cdot \mathbf{e}.$$
- For the case where the initial state is considered
$$E[N_k(t)] = \boldsymbol{\theta} \cdot \mathbf{A}_k \cdot \mathbf{e} + \boldsymbol{\theta} \cdot \int_0^t \mathbf{P}(u)du \left(\mathbf{Q}_k - \tilde{\boldsymbol{Q}}_{kk}\right) \cdot \mathbf{e},$$

All the progress up to reach these expressions can be consulted in Appendix A2.4.1.

Parameter Estimation for the stochastic process $X(t)$

When m independent devices are considered, the purpose is to maximise the following likelihood function to estimate the parameters

$$L = \prod_{l=1}^{m} \boldsymbol{\alpha}_{x_0^l} \left[\prod_{a=0}^{m_l-1} e^{\mathbf{Q}_{x_a^l x_a^l}\left(t_{a+1}^l - t_a^l\right)} \left(\mathbf{Q}_{x_a^l x_{a+1}^l}\right)^{\tau_l} \right] \mathbf{e},$$

where t_a^l corresponds to the transition from macro-state x_a^l to macro-state x_{a+1}^l, τ_l is zero if the last time is a censoring time and one otherwise, $\mathbf{Q}_{aa}$ is a square sub-stochastic matrix whose main diagonal is negative and the rest positive values, and $\mathbf{Q}_{ab}$ a matrix with positive elements with $\sum_{b=1}^r \mathbf{Q}_{ab}\mathbf{e} = 0$ for any a and b.
Other log-likelihood function is given from the transition probabilities of the process

$$\log\, L = \sum_{l=1}^{m} \sum_{a=0}^{m_l-1} \log \left(h_{x_a^l x_{a+1}^l} \left(t_a^l, t_{a+1}^l\right)\right).$$

3.4 Linear Phase-Type probability modelling of functional PCA with applications to resistive memories

In this work we describe a new probability distribution stemmed from Phase-type (PH) distributions, defined as Linear Phase-type distribution. The formal definition of a LPH distribution is the following.

- A random variable X follows a Linear Phase-type distribution with representation $(a, b, \boldsymbol{\beta}, \mathbf{S})$ if $Y = a + bX$ for a, b $(b \neq 0) \in \mathbb{R}$ is Phase-type distributed with representation $(\boldsymbol{\alpha}, \mathbf{T})$. In this case $\boldsymbol{\beta} = \boldsymbol{\alpha} e^{\mathbf{T}a}$ and $\mathbf{S} = b\mathbf{T}$.

From this definition, we have obtained several measures related to a LPH distribution. For instance, the reliability function of this class of distributions is

$$R_X(x) = \left\{ \begin{array}{lll} \boldsymbol{\beta} e^{\mathbf{S}x}\mathbf{e} & ; & \text{for} \quad x > \frac{-a}{b},\ b > 0 \\ 1 - \boldsymbol{\beta} e^{\mathbf{S}x}\mathbf{e} & ; & \text{for} \quad x < \frac{-a}{b},\ b < 0 \end{array} \right. .$$

Besides, it is possible to derive all the moments by means of the function

$$M_X(t) = -\boldsymbol{\beta}(\mathbf{S} + \mathbf{I}t)^{-1} e^{-(\mathbf{S}+\mathbf{I}t)a/b}\mathbf{S}^0,$$

so that the n-th moment is computed as $E[X^n] = \frac{\partial^n M_X(t)}{\partial t^n}|_{t=0}$.

Additionally, LPH class has been studied in detail, thereby achieving important results. Taking into account the properties of the PH distributions, these outcomes can be summarized as follows (all the details and demonstrations of them can be seen in the Appendix A3.2.2).

1. The finite summation of independent LPH distributions with PH distributions associated follows a LPH distribution.

2. A positive homothecy of a LPH distribution is also LPH distributed.

3. The set of LPH distributions is dense in the set of probability distributions defined on any half-line of real numbers.

The developed methodology together with these last results makes possible the modeling of a random variable defined on any semi-line real. Particularly, we display in next Chapter as the LPH distribution of the principal components is inherited by the stochastic process through the Karhunen-Loève expansion.

3.5 New Modeling Approaches Based on Varimax Rotation of Functional Principal Components

The Varimax rotation can be extended to the field of multivariate Principal Component Analysis (PCA). Taking into account that PCA can be performed by considering Single Value Decomposition (SVD), there are available two different ways to conduct the Varimax criterion (Jolliffe, 2002). Note that both procedures provide different results whose properties must be kept in mind when for the interpretation.

- The first possibility consists of rotating the weight matrix (eigenvectors of the covariance matrix) and scores of the principal components. This approach guarantees that the orthogonality among axes is maintained but scores will not be uncorrelated anymore. This fact is a mishap for PCA because it is not how rotations are usually understood and applied.
- The second technique is based on rotating the loadings (eigenvectors scaled by the corresponding singular value) and scores of the standardized principal components. Now, the rotated axes are not orthogonal which provokes that the projections of the data do not make sense, but the rotated scores remain uncorrelated.

Anyway, the explained variance by the first q components are still the same after applying Varimax rotation in both approaches. However, variances redistributed among rotated components are not arranged in descending order, as it happens with PCA performed without rotation.

Functional Varimax Rotation

The objective is to develop new modeling approaches based on Varimax rotation in order to better understand the variability features after applying Functional Principal Component Analysis. Taking into account the equivalence between FPCA of a functional variable X and multivariate PCA of $A\Psi^{1/2}$. We propose two new functional approaches based on Varimax rotation of PCA matrix $A\Psi^{1/2}$, being A the basis coefficients matrix and Ψ the matrix of inner product between basis functions.

Let us assume that V is the matrix whose columns are the eigenvectors associated with the covariance matrix of $A\Psi^{1/2}$. Let us denote the j-th column by $\boldsymbol{v}_j$. Then, $Z = (\xi_{ij})_{n\times p} = (A\Psi^{1/2})V$ is the matrix whose columns are the principal component scores. Besides, the basis expansion of eigenfunctions can be expressed in matrix form as $f = B^T\boldsymbol{\phi}$ with $f = (f_1, \dots, f_p)^T$ and $B = (b_{jk})_{p\times p} = \Psi^{-1/2}V$ being the basis coefficients matrix of eigenfunctions.

FPCA rotation would comprise rotating the first q weight functions as $f_q^{R^T} = f_q^T R$. Thus, the sample functions can be approximated by considering the first q

principal components as follows

$$X^q = Z_q f_q = (Z_q R)(R^T f_q) = Z_q^R f_q^R,$$

where the vector of rotated eigenfunctions is expressed as

$$f_q^{R^T} = \boldsymbol{\phi}^T B_q R = \boldsymbol{\phi}^T (\Psi^{-1/2} V_q) R,$$

with B_q being the matrix of basic coefficients associated with the first q eigenfunctions and V_q the matrix whose columns are the first q eigenvectors. Hence, there are four different possibilities to apply the Varimax criterion in FPCA:

R1 Applying VARIMAX rotation criterion to weight function values.

The objective is to calculate a matrix R that maximizes the variance of the squares of the elements of the matrix

$$F_q^{R^T} = F_q^T R,$$

where F_q is the $q \times m$ matrix whose elements are the values of the first q eigenfunctions evaluated at a grid of time points $t_1, ..., t_m$, given by $F_q^T = \Gamma^T \Psi^{-1/2} V_q$, with Γ being the $p \times m$ matrix that contains as rows the values of each basis function at the time points.

R2 Applying VARIMAX rotation criterion to weight function coefficients.

The method consists of determining a matrix R that maximizes the variability of the squares elements of $B^R = BR = \Psi^{-1/2} VR$. Then, the rotated principal factors are given by

$$f_q^{R^T} = \boldsymbol{\phi}^T B^R.$$

R3 Applying VARIMAX rotation criterion to PCs by rotating the matrix of eigenvectors

The aim is to find a matrix R that maximizes the variability of the squares elements of the rotated matrix of eigenvectors $V_q^R = V_q R$. Then, the rotated principal factors are given by

$$f_q^{R^T} = \boldsymbol{\phi}^T (\Psi^{-1/2} V^R).$$

R4 Applying VARIMAX rotation criterion to the standardized PCs by rotating the matrix of loadings

The goal is to compute a matrix R that maximizes the variance of the squares elements of the matrix $\Delta_q^R = \Delta_q R = V_q \Lambda_q^{1/2} R$. Then, the rotated principal factors are given by

$$f_q^{R^T} = \boldsymbol{\phi}^T \left(\Psi^{-1/2} \Delta_q^R \Lambda_q^{-1/2} \right).$$

Methods **R3** and **R4** would be the main contribution of this work, whereas the methods **R1** and **R2** are the options proposed by Ramsay and Silverman (2005). **R3** and **R4** are inspired in the possibility of rotation the eigenfunctions instead of the basic coefficients. A depth discussion about the main features of these rotations is made in next Chapter.

3.6 Homogeneity problem for basis expansion of functional data with applications to resistive memories

Homogeneity problem aims to test if more than two independent samples of functional data have been generated from the same stochastic process. Let us consider that we have m independent samples of i.i.d. stochastic processes (functional variables) denoted by $\{X_{ij}(t) : i = 1, \dots, m;\ j = 1, \dots, n_i;\ t \in T\}$ with distribution $SP(\mu_i(t), \gamma(s,t))$, $\forall i = 1, \dots, m$, with $\mu_i(t)$ and $\gamma(s,t)$ being the mean function and the common covariance function associated with each of the m stochastic processes, respectively. Besides, we assume that $\{x_{ij}(t) : i = 1, \dots, m;\ j = 1, \dots, n_i;\ t \in T\}$ are m independent samples of curves which can be seen as realizations of the stochastic processes aforementioned. All sample curves verify the hypothesis H_1, H_2 and H_3.

In the context of functional analysis of variance, the objective is to check the equality of the mean functions among the different obtained samples. This means to test the following hypothesis

$$H_0 : \mu_1(t) = \cdots = \mu_m(t),\ \forall t \in T,$$

against the alternative that its negation holds.

We propose two new methods based on basis expansion of functional data to tackle the FANOVA problem distinguishing if the data are generated by a Gaussian process or another process. The first one consists of testing multivariate homogeneity on the matrix of basis coefficients, and the second approach resides in conducting the multivariate analysis of variance on the principal component scores given by FPCA.

Homogeneity testing on basis coefficients

When the one-way FANOVA problem is considered, functional data can be expressed in terms of next linear model:

$$X_{ij}(t) = \mu(t) + \alpha_i(t) + \epsilon_{ij}(t),\ i = 1, \dots, m;\ j = 1, \dots, n_i,$$

where $\mu(t)$ is the overall mean function, $\alpha_i(t)$ is the main-effect function for the group i and $\epsilon_{ij}(t)$ are the error functions i.i.d. $SP(0, \gamma(s,t))$.

The main-effect functions are not identifiable so that in order to be estimated some constraint must be imposed. The most used constraint is $\sum_{i=1}^{m} \alpha_i(t) = 0$. Under this constraint, there has been proved (see Appendix A5.2.1) that FANOVA testing problem is equivalent to the usual multivariate ANOVA on the matrix of basis coefficients $A = (a_{(ij)k})_{n \times p}$ with $n = \sum_{i=1}^{m} n_i$. At this point, there are two possibilities:

- From parametric viewpoint, the MANOVA problem can be applied through one of the following tests: the Wilks's lambda, the Lawley-Hotelling's trace, the Pillai's trace, and the Roy's maximum root. The four chances do not have exact null distribution but they can be approximated by F-test statistics (Rencher and Christensen, 2012). Nevertheless, the following assumptions are required: (1) Normality for the basis coefficients; (2) observations are randomly obtained; (3) the dimension of the sample in each group must be larger than the variables space; (4) homogeneity of variance-covariance matrices in the m groups; (5) no multicollinearity.

- Unfortunately, the problem inherent is that the processes are seldom Gaussian and therefore, the basis coefficients are not Gaussian either. Hence, other approaches should be considered facing the impossibility of using the parametric tests described earlier. One option would be to perform the bootstrap and permutation versions of these tests (Gorecki and Smaga, 2015). On the contrary, we propose to carry out nonparametric multivariate tests such as the extension of the univariate Kruskal Wallis's test and Mood's test (Oja, 2010; Ellis et al., 2017). These procedures are based on the use of the median instead of mean and moreover, provide a solution when the sample size is small by means of permutation techniques. The principal difference between Kruskal Wallis's test and Mood's test is that the first one is more powerful when data are generated from some distribution but at the same time it is more sensitive with the presence of outliers.

Homogeneity testing on functional principal components

We introduce a new approach based on FPCA in order to solve the FANOVA problem. Theoretically, now the basis is represented by the eigenfunctions of the covariance operator, meanwhile the coefficients of matrix A are substituted by the most explicative principal components scores (the first q principal components scores). The choice of q will be made ensuring that the proportion of variance explained by the first q principal components is as close as possible to one.

Under this conceptual framework, we propose again two new procedures (parametric and nonparametric methods) to solve the homogeneity problem on the vector of the first q principal components in the m groups.

- When multivariate normality is satisfied, univariate ANOVA on each principal component score should be performed, since the principal components are uncorrelated. It is well-known that uncorrelatedness implies independence for the multivariate normality case. Then, if the response variables are independent, the multivariate tests do not make sense because they are less powerful. Besides, Bonferroni inequality must be applied to control the Type I error produced by these multiple ANOVA tests. This correction consists of dividing the overall level by the number of tests, whose result will be the alpha level for each ANOVA test.
- Nonparametric multivariate tests will be conducted when normality is not verified.

This new approach is really interesting because reduces notably the great dimension problem when a high number of basis functions is selected to achieve an accurate representation of the curves.

3.7 Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis

In the current work, we deal with the functional ANOVA problem for repeated measures and the multivariate functional ANOVA problem for independent groups and a vector functional variables.

Functional ANOVA for repeated measures

The objective is to compare two or more mean functions from paired design, in which we have the repeated functional data for the same subjects submitted to R conditions or time periods. In this context, it is assumed that the sample functions can be represented as $X_{jr}(t)$ with $t \in \mathcal{T} = [a, b]$, $j = 1, ..., n$ and $r = 1, ..., R$, such that $E[X_{jr}(t)] = \mu_r(t)$. Only two different conditions or periods of time are evaluated in the current work ($R = 2$). The goal is to test the hypothesis

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

For that purpose, Martinez-Camblor and Corral (2011) proposed the following statistics

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

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

However, $\mathcal{C}_n$ does not take the within group variability into account. To solve this aspect, Smaga (2020) proposed the following two statistics:

$$\begin{aligned} \mathcal{D}_n &= n \int_T \frac{\left(\bar{X}_1(t) - \bar{X}_2(t)\right)^2}{\hat{K}(t,t)} dt, \\ \mathcal{E}_n &= sup_{t\in[a,b]} \left\{ \frac{n\left(\bar{X}_1(t) - \bar{X}_2(t)\right)^2}{\hat{K}(t,t)} \right\}, \end{aligned}$$

with $\hat{K}(t,t) = \frac{\sum_{j=1}^n \left[(X_{j1}(t) - \bar{X}_1(t)) - (X_{j2}(t) - \bar{X}_2(t))\right]^2}{n-1}$.

In this paper, $\mathcal{C}_n$, $\mathcal{D}_n$ and $\mathcal{E}_n$ are computed by considering the basis expansion approach. Now, if we generalise the expression (3.1) for this design, the curves are expressed as $X_{jr}(t) = \mathbf{a}_{jr}^T \boldsymbol{\phi}(t)$, with $j = 1, ..., n$ and $r = 1, 2$. Then,

$$\begin{aligned} \left(\bar{X}_1(t) - \bar{X}_2(t)\right)^2 &= \left(\bar{\mathbf{a}}_1^T \boldsymbol{\phi}(t) - \bar{\mathbf{a}}_2^T \boldsymbol{\phi}(t)\right)^2 \\ &= \left(\boldsymbol{\phi}(t)^T \bar{\mathbf{d}}\right)^2 = \boldsymbol{\phi}(t)^T \bar{\mathbf{d}} \bar{\mathbf{d}}^T \boldsymbol{\phi}(t), \end{aligned}$$

and

$$\begin{aligned} \hat{K}(t,t) &= Var(X_1(t)) - 2Cov(X_1(t), X_2(t)) + Var(X_2(t)) \\ &= \hat{C}_1(t,t) - 2\hat{C}_{12}(t,t) + \hat{C}_2(t,t) \\ &= \boldsymbol{\phi}(t)^T (\hat{\Sigma}_1 - 2\hat{\Sigma}_{12} + \hat{\Sigma}_2) \boldsymbol{\phi}(t), \end{aligned}$$

with $\bar{\mathbf{d}} = (\bar{d}_1, ..., \bar{d}_p)^T = \bar{\mathbf{a}}_1 - \bar{\mathbf{a}}_2 = (\bar{a}_{11}, ..., \bar{a}_{1p})^T - (\bar{a}_{21}, ..., \bar{a}_{2p})^T$, where $\bar{a}_{rk} = n^{-1} \sum_{j=1}^n a_{jrk}$ $r = 1, 2$; $k = 1, ..., p$. In addition, $\hat{\Sigma}_r$ is the sample covariance matriz of the matrix A_r of basis coefficients in the group r, whose elements are $A_r = (a_{jrk})$, and $\hat{\Sigma}_{12}$ is the sample cross-covariance matrix between A_1 and A_2. Note for major clarity that $\bar{X}_r = n^{-1} \sum_{j=1}^n \mathbf{a}_{jr}^T \boldsymbol{\phi}(t) = \bar{\mathbf{a}}_r^T \boldsymbol{\phi}(t)$.

In order to approximate the null distribution of these statistics, different approaches can be consulted in Martinez-Camblor and Corral (2011), Smaga (2019; 2020). A brief summary can be checked in Appendix A6.2.1.

Multivariate FANOVA for independent measures

The goal is to test the equality of the mean functions coming from independent groups when more than one functional response variable are considered in the analysis. The developed methodology for this theoretical framework is the extension of the parametric and nonparametric methods proposed in the previous section. In particular, we focus on the approach based on principal components. We leverage that, although the eigenfunctions are vectors of functions, the components are still scalar. Therefore, the multivariate FANOVA is again reduced to a MANOVA problem on the vector of the most explicative pc's scores. We assume the initial multivariate scenario described in Section 3.1 *Multivariate FPCA*, but now the multivariate functional variables are denoted by $\boldsymbol{X}_{ij}(t) = (X_{ij1}(t), \ldots, X_{ijH}(t))^T$ with $i = 1, \ldots, g$; $j = 1, \ldots, n_i$; $h = 1, \ldots, H$ and mean vector $\boldsymbol{\mu}_i$. Once the principal component scores are obtained by considering the basis expansion, two different methodologies are proposed in this paper in order to solve the multivariate testing problem

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

against the alternative that its negation holds. Both approaches are based on testing homogeneity on the vector of the first q principal components scores in the g groups.

- The first one lies in performing univariate ANOVA on each principal component by correcting the level of significance when the normality is satisfied.

- The second method consists of applying non-parametric multivariate tests such as the extensions of the univariate Kruskal Wallis's test and Moods's test when the sample curves are not generated from a Gaussian process.

3.8 COVID-19 data imputation by multiple function on function principal component regression

In this article, an extension of the function-on-function regression model for the case of several functional predictors is introduced for the imputation of missing values of the functional response. In particular, the objective is to complete the information of COVID-19 hospitalized and intensive care people curves (functional response variables) through multiple functional predictors (curves of positive, recovered and deceased cases).

Multiple function-on-function linear model

The multiple function-on-function linear regression (MFFLR) model enables to estimate a functional response Y from a vector of more than one functional predictor variable, i.e., $X = (X_1, \ldots, X_J)^T$. By considering that $\{(x_i, y_i) : i = 1, \ldots, n\}$ with $x_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})^T$, is a random sample from (X, Y) where each functional variable takes take values on the Hilbert space $L^2(T)$ defined above, the model written in matrix form adopts the following expression

$$\begin{aligned} y_i(t) &= \alpha(t) + \sum_{j=1}^{J} \int_T x_{ij}(s)\beta_j(s,t)ds + \varepsilon_i(t) \\ &= \alpha(t) + \int_T x_i(s)^T \beta(s,t)ds + \varepsilon_i(t),\ i = 1, \ldots, n, \end{aligned}$$

where $\alpha(t)$ is the intercept function, $\beta(s,t) = (\beta_1(s,t), \beta_2(s,t), \ldots, \beta_J(s,t))^T$ are the J coefficient functions, $x_i(s) = (x_{i1}(s), x_{i2}(s), \ldots, x_{iJ}(s))^T$ and $\epsilon_i(t)$ are independent functional errors. Although all variables are defined in the same interval T, the model can be easily extended to the case of variables with different domains.

There are several approaches in order to estimate the model. The most common techniques are based on least squares penalized approaches and basis expansion of sample curves and/or functional parameters. Although this methodology is very attractive because the problem turns into a multivariate linear model for the matrix of response basis coefficients, the resultant multivariate model presents a high multicollinearity. One solution would be to represent the response and the predictor functional variables in terms of the principal components (uncorrelated variables). Again, we would obtain a multivariate linear model but without multicollinearity.

Functional principal component regression

If we consider the principal component decompositions for the response and the predictor functional variables, the MFFLR model is reduced to a linear regression model for each principal component of the functional response Y on all principal components of the functional predictors. On this subject, by truncating each principal component decomposition, the principal component MFFLR model can be expressed as follows:

$$\hat{y}_i(s) = \bar{y}(s) + \sum_{k=1}^{K} \hat{\xi}^y_{ik} f^y_k(s) = \bar{y}(s) + \sum_{k=1}^{K} \left(\sum_{j=1}^{J} \sum_{l \in L_{kj}} \hat{b}^{x_j}_{kl} \xi^{x_j}_{il} \right) f^y_k(s), \tag{3.2}$$

with $\hat{b}^{x_j}_{kl}$ being the linear least squared estimation of the regression coefficients b_{kl} when $\beta_j(s,t) = \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} b^{x_j}_{kl} f^{x_j}_k(s) f^y_l(t)$ is expressed in terms of basis expansion; f^y_k are the eigenfunctions of the sample covariance operator of $y_i(t)$; $\xi^{x_j}_{il}$ and ξ^y_{ik} are the principal component scores of predictor and response functional variables, respectively. For more information about the theoretical development, see Appendix A7.3.2.

Additionally, another important aspect is the selection of principal components of each predictor variable. On this point, the most explicative components might be independent or slightly correlated with the response variable. A good option would be to adapt common selection models procedures based on stepwise and best subset regression combined with cross-validation for this functional framework.

Imputation of missing response curves

We assume that the information about the predictor variables is totally known beforehand and only certain values are missing for the response variable. This means that we have n curves whose evolution is completely observed and m curves with incomplete observations for the response. Then, the imputation of the missing response curves has the following steps:

- We estimate the parameters b_{kl} with the complete n curves.
- The principal component scores of predictors $\{\xi_{il}^{x_j} : i = n+1, \ldots, n+m, l = 1, \ldots, n-1\}$ are computed through the expression given in its definition.
- These scores are substituted in (3.2) in order to estimate the missing response curves $\{y_i^{miss}(s) : i = n+1, \ldots, n+m\}$.

Then, the estimated model can be used to predict new values of the response Y on a test sample and to provide accurate interpretation of the relationship between the predictor and the response variables. If objective is to predict the response variable in a future interval of amplitude k denoted by $[T, T+k]$, (3.2) could be estimated in terms of the predictor variables in the past interval of time $[0, T]$.

Chapter 4

Results

We summarise here the main results achieved throughout the research articles which constitute the current thesis, detailing both theoretical and numerical results.

4.1 Phase-type distributions for studying variability in resistive memories

Our intention is to study the inner performance of RRAM memories by estimating the internal structure of the associated Phase-type distribution. For that purpose, we make use of EM algorithm given in Section 3.2. Here, we focus on the modeling of reset voltages (voltage at which the conductive filament breaks), but this approach can be extended to other mesures (resistances and currents) or to the case of established process data. In order to fit a distribution to these observed values, we consider multiple general Phase-type distributions. Also, we assume any internal structure for the transition intensities matrix $\mathbf{T}$. After doing the analysis, we observed that all cases converged to the same result for a fixed number of phases

$$\boldsymbol{\alpha} = (1, \dots, 0), \quad \text{and} \quad \mathbf{T} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\lambda & \lambda \\ 0 & \cdots & \cdots & 0 & -\lambda \end{pmatrix},$$

so that, the internal structure of PH representation depends on one parameter. This structure corresponds to an Erlang distribution with parameters $E(m, \lambda)$, with m being the number of phases. Then, we can conclude that the voltage until the conductive filament is broken can be modeled by an Erlang distribution. The interpretation of its parameters is very intuitive: the process until the failure of the conductive filament begins in phase 1 (once the filament has been completely formed) and it undergoes a degradation evolution of m well differenced states, with $1/\lambda$ being the reset voltage mean for each stage. The optimum value was reached for 15 phases with $\hat{\lambda} = 9.279325$ via the EM algorithm.

Finally, we graphically compare the fit of Weibull and Erlang distribution through the experimental cumulative risk rate, the reliability function and the hazard rate function. The results achieved with the Erlang distribution are much better than with the Weibull distribution. As a matter of fact, the accuracy of the fit by means of Erlang distribution is very outstanding. All computational aspects were made with the EMpht program and the software R (R Core Team, 2020).

4.2 A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories

A new stochastic process with macro-states is proposed in order to model systems whose sojourn time in each macro-state is not exponentially distributed and only macro-states can be observed. It is assumed that the macro-state where the process is located is known at each instant of time, but its internal behaviour is not observable. Besides, the model is built in transient and stationary regimes and multiple measures are achieved. Finally, the proposed process to model the RTN signals that occur within RRAM memories will be considered.

The main theoretical results obtained throughout the work are summarized below. More information on mathematical developments can be found in Appendix A2.

- The generator of the macro-state stochastic process $\{X(t) : t \geq 0\}$ is built by means of matrix blocks. In turn, each macro-state is made up of multiple internal states.

- Inside the macro-state process, there is an embedded Markov process, $\{J(t) : t \geq 0\}$, that represents the internal behaviour of the system.

- The transition probability matrix for the general process $X(t)$ is worked out in a matrix-algorithmic form by considering the transition probabilities of the Markov process.

- The process $X(t)$ is non-homogeneous and not Markovian. The process $J(t)$ verifies both properties.

- The stationary distribution of $X(t)$ is computed through the internal matrix blocks in order to reduce the computational cost.

- The sojourn time in any state for the Markov process is exponentially distributed, meanwhile the sojourn time in each macro-state is Phase-type distributed.

- If $X(t)$ has reached the stationary regime and the process is in macro-state $\boldsymbol{i}$, the probability distribution of the sojourn time is also Phase-type distributed.

- The distribution of the number of visits to a given macro-state between two different times is obtained from several differential equations and the inverse Laplace transform.

- The expression of the mean number of visits is developed explicitly depending on times and if the initial state is considered.

- Two different likelihood functions have been derived in order to estimate the parameters of the built model.

Regarding the application, we analyze four RTN signals (called RTN25, RTN26, RTN27 and long-RTN) coming from the same device. The discrepancies between them lie in the measurement time and the applied voltages that produces the variations of electric current. The measurement time was approximately about nineteen minutes for RTN25, RTN26 and RTN27 signals with a supplying of 0.34 volts, 0.35 volts and 0.36 volts respectively, meanwhile, the long-RTN trace was subjected more than three hours of record with 0.5 volts. In addition, we determine the number of latent levels into the signals by means of the hidden Markov models. We study the evolution of these series separately.

- **Series RTN25-26-27**. After an exhaustive study, it is shown that these signals have a Markovian behaviour. Thus, the classical methodology on Markov chains is carried out on them (see Appendix A2.6.1 for more information).

- **Serie long-RTN**. Once the hidden Markov model is applied, we estimate the proportion of times that the signal is in each latent state and the stationary distribution for the continuous Markov process associated with the latent states. Based on these results, we assume that the best arrangement is with 4 latent states, since insignificant proportions would be obtained if more than 4 levels were considered. Given that the spent time in each level does not follow the exponential distribution, the new approach is carried out. In particular, Phase-type distributions with a Coxian/Erlang structure have been considered. The best fit is achieve for the case in which the macro-states are composed of 2, 2, 4 and 3 internal states, respectively. Likewise, we perform the Anderson-Darling test as an indicator of the goodness of fit, being its p-values higher than the usual significance level $\alpha = 0.05$ for the four cases. Therefore, we assume that the sojourn time in the macro-state $\boldsymbol{i}$ follows a Phase-type distribution with representation $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$. In conclusion, the perfomance of the device is regulated by a non-homogeneous process $\{X(t) : t \geq 0\}$ with an embedded Markov process $\{J(t) : t \geq 0\}$. The macro-states space is $\{1, 2, 3, 4\}$ and for the embedded process is $\{1 = i_1^1, 2 = i_2^1, 3 = i_1^2, 4 = i_2^2, 5 = i_1^3, 6 = i_2^3, 7 = i_3^3, 8 = i_4^3, 9 = i_1^4, 10 = i_2^4, 11 = i_3^4\}$.

4.3 Linear Phase-Type probability modelling of functional PCA with applications to resistive memories

Let us consider that we have a stochastic process, $\{X(t) : t \in T\}$, that fulfills the hypotheses H_1, H_2 and H_3. The main objective of this work is to identify the distribution of the principal components in order to characterize the whole process through the Karhunen-Loève expansion.

A new class of distributions called Linear Phase-type distributions has been introduced in which the following results are satisfied:

1. The finite addition of independent LPH distributions with associated PH distributions follows a LPH distribution.
2. A positive homothecy of a LPH distribution is also LPH distributed.
3. The set of LPH distributions is dense in the set of probability distributions defined on any half-line of real numbers.

Considering these results, we have proved that

- if the principal components are LPH distributed with the same scale parameter, then the process expressed in terms of K-L expansion follows a LPH distribution at each instant of time. The corresponding representation of the LPH distribution is provided for the process (see Appendix A3.3).

This result has been considered in order to model the stochastic behaviour of RRAM memories. In particular, the goal is to fit the current probability distribution at each voltage in the reset process through the K-L expansion.

- A set of 232 reset curves have been considered in the application.
- Previously, curves were registered in the interval [0,1] and reconstructed by means of P-Spline. They are denoted as $\{I_i(u) : u \in [0, 1],\ i = 1, \ldots, 232\}$.

- After applying FPCA, we observed that more than 99% of the total variability of the process is explained only with the first principal component. Here, the curves can be approximated with great precision by truncating the Karhunen-Loève expansion in the first term.

- Then, the objective is to fit a probability distribution to the scores of the first principal component. Due to these scores are negative and positive values, some transformations are necessary to apply PH distributions. We considered the linear transformation $1 + 1000 \times \xi_1$ with ad hoc selection of the slope and constant parameters. Before this transformation all the scores values belonged to the interval [-1,1].

- The parameters of a PH distribution with m transient states and any internal structure for matrix $\mathbf{T}$ were estimated by using the EM algorithm. The optimum value was achieved for 21 phases.

- Besides,an in-depth comparison was made with other classic probability distributions such as Weibull, Normal and Cauchy. However, neither of them overcame the fitting reached by the PH distribution. In fact, only the PH distribution can be accepted according to the p-values provided by the Anderson-Darling test.

- Therefore, we can conclude that
  - $1 + 1000 \times \xi_1$ can be modeled by a PH distribution.
  - ξ_1 follows a LPH distribution with representation $(1, 1000, \boldsymbol{\beta}, \mathbf{S})$.
  - The process is also LPH distributed with representation

$$\left(|f_1(u)| - 1000\bar{I}(u) sgn\left(f_1(u)\right), 1000 sgn\left(f_1(u)\right), \boldsymbol{\alpha} e^{\mathbf{T}\left(1-\frac{1000}{f_1(u)}\bar{I}(u)\right)}, \frac{1000}{f_1(u)}\mathbf{T} \right).$$

4.4 New Modeling Approaches Based on Varimax Rotation of Functional Principal Components

In this paper, two new methods based on the equivalence between FPCA of basis expansion of the sample curves and multivariate PCA of a transformation of the matrix of basis coefficients are introduced to perform the Varimax criterion in FPCA. The first one consists of rotating the original scores and eigenvectors, whereas the second methods lies in rotating the standardized scores and loadings. These methods will be noted as **R3** and **R4**, respectively. Moreover, these techniques are compared with the options inspired by Ramsay and Silverman (2005) for the same purpose. They proposed to apply the Varimax criterion on the matrix of values after evaluating the weight functions in a grid of equally spaced time knots (**R1**) or on the matrix of basis coefficients of the eigenfunctions (**R2**). The most important theoretical outcomes are described in what follows:

- **R1**, **R2** and **R3** provide orthonormal factors but correlated PC scores, meanwhile **R4** supplies quite the opposite.
- When orthonormal basis functions are considered to reconstruct the real form of the curve, **R2** and **R3** match each other.
- The proportion of redistributed variance among the rotated standardized component scores (**R4**) is the same because of the properties of the SVD decomposition.

A simulation study was carried out in order to analyze the performance of the approaches proposed in this work. Also, a comparison with the rest of methods was done according to the integrated mean squares error (MSE) of each rotated eigenfunction in relation to the original rotation. The data were simulated from the approximation of the Wiener process given by its truncated Karhunen–Loève expansion (see the Appendix A4.3 for more information about this Gaussian process and how MSE is computed). Besides, 500 replications of 150 sample curves of the process were simulated for several scenarios: different sample sizes and distinct number of equally spaced knots in the observed domain [0,1] were contemplated (in particular, $t_k = k/m,\ k = 0, 1, \ldots, m;\ m = 25, 50, 100$). Curves were smoothed by means of a basis of cubic B-splines of dimension 8. Then, as a result of this study we can conclude that:

- When the sample size is sufficiently large, the outcomes are very similar.
- The method **R3** gets the most accurate results in relation to MSE.
- The method **R3** is more robust with respect to the number of observation nodes of the sample curves.
- It does not make sense to compare **R4** with the remainder options due to the fact that the estimated weight functions are not orthogonal for this method.

Regarding the application, we study the evolution of the number of infected cases by COVID-19 per 10000 inhabitants in the Spanish autonomous communities (AC) for the first wave of the pandemic.

- Pre-processing steps are necessary before applying FPCA in order to evaluate the behaviour of the illness from 20/02/2020 to 27/04/2020. Firstly, all curves are registered in the interval [0,1]. Besides, since curves are smooth, a cubic B-spline basis of dimension 10 with equally spaced knots in the interval [0,1] is chosen to approximate the sample curves. Finally, FPCA is carried out.

- The first principal component explains more than 99% of the total variability, so that the interpretation is not an easy task. In fact, only the first component could be interpreted. On this matter, the first eigenfunction is positive and strictly increasing during the entire observation period, being the weight almost two times bigger at the end of the period than at the beginning. This means that the number of cases shot up in the ACs over time.

- However, it is obvious to think that not all ACs suffered the same speed of spreading, but rather each AC had its own rhythm. In this sense, methods **R3** and **R4** for functional Varimax rotation were considered on the first four principal components.

- After rotation, we can see that there are three big groups of ACs according to evolution of the number of cases during the first wave. It is possible to interpret the temporal behaviour of them thanks to the rotated eigenfunctions given by the methods **R3** and **R4**.

- The results given by FPCA were compared with the ones obtained by applying a functional clustering analysis on the matrix of basis coefficients (Jacques and Preda, 2014b). We saw that effectively, both methodologies provide similar results and they are in accordance with other studies about the evolution of COVID-19 in Spain (Henriquez et al., 2020; Siqueira et al., 2020; Santamaria and Hortal, 2021).

- These results reveal the power and outstanding performance of the functional Varimax approaches introduced in this paper.

4.5 Homogeneity problem for basis expansion of functional data with applications to resistive memories

Developing new techniques to test the homogeneity of the distributions is the main motivation of this paper. We introduce two new procedures whose behaviour is evaluated through a rigorous simulation study. These methods are reduced to check the multivariate homogeneity on a vector of basis coefficients and on a vector of principal component scores. In both scenarios we should distinguish whether the data have been generated by Gaussian processes or not.

- For the parametric design (Gaussian processes), the objective is to checkt if the mean functions are equal among the different samples. In this case, classical multivariate ANOVA tests might be applied for both layouts.

- When Gaussianity is not satisfied, we propose to apply nonparametric multivariate homogeneity tests such as the extensions of the univariate Kruskal Wallis's tests and Mood's tests. Under this situation, the problem is equivalent to test the equality of the medians in all groups.

Regarding the simulation study, detailed information about the artificial aspects can be seen in Appendix A5.3. In broad terms, three groups are considered with three different models for the mean functions (M1, M2 and M3). M1 represents the situation where the equality of mean functions is true, meanwhile in M3 the discrepancies between group mean functions are smaller. For different sample sizes ($n_i = 15, n_i = 25, n_i = 35, \ i = 1, 2, 3$), the sample curves are generated at 51 equally spaced time points in the interval [0,1] according to the FANOVA linear model. Besides, Gaussian and non-Gaussian errors were considered by using five different values for the dispersion parameter ($\sigma = \{0.02, 0.05, 0.10, 0.20, 0.40\}$). A cubic B-spline basis of dimension 18 was employed to get the basis coefficients. On the other hand, Pillai's trace test (for basis coefficients approach) or multiple ANOVA tests using Bonferroni's correction (for principal components layout) and the extension of the univariate Kruskal Wallis's test (for both designs) were performed for the parametric and nonparametric cases, respectively. Finally, for the principal component procedure, different number of components were taken into account. After the simulation study, we deduce the following strengthens and weakness of the two new procedures:

- The parameter of error dispersion σ notably affects the tests quality. The power of the tests may be questioned for large values of σ, especially when $M3$ and $\sigma = 0.4$ are simultaneously considered. However, similar simulation studies do not use values as higher as $\sigma = 0.40$ (Gorecki and Smaga, 2015).

- Regarding sample sizes, as σ is becoming increasingly large, the tests convert into more conservative when n decreases for the cases where the null hypothesis is false.

- The choice of the number of principal components plays an important role as well. The outcomes display that the higher the explained variability for the components, the better the power of the test. The advice is to overcome the 99% of the total variability, primarily when σ is large.

- In general, the results are really outstanding, except when $M3$ and $\sigma = 0.4$ are given at the same time. The basis coefficients model is slightly better for the parametric scenario and the principal components approach for the nonparametric case. The power of the tests causes us concern in the more extreme situations.

- Parametric tests are more powerful and therefore, they should be considered as long as the assumptions are satisfied.

- Precaution when the multiple ANOVA tests are used for a great number of principal components. We think that the reason why the basis coefficients procedure has provided superior results for the parametric case is because we needed eight principal components to explain more than 99% of the total variability. Then, the significance level in the multiple ANOVA tests is lower due to Bonferroni's inequality and the tests are less powerful. Hence, this controversy would disappear when only two or three components were necessary, as the corrected significance level would not be so small and the acceptance proportion would increase.

Finally, an application with data from RRAM memories was carried out in order to test if the kind of material and thickness play a fundamental role in RRAM operation. We have the information about three kind of devices made of different materials and thicknesses (see Appendix A5.4 for more information about the features of the devices). The experimental data associated with these devices are current/voltage curves associated with the reset cycles. Before applying the methodology, all curves were registered in the interval [0,1] and reconstructed by P-Spline smoothing with B-Spline basis of dimension 20 and penalty parameter $\lambda = 0.5$. The purpose is to test if the sample mean functions of the three technologies are equal or not. Graphically, it seems that the material and thickness take part an important role in the RRAM behaviour. We check this suspicion by means of both considered approaches.

- **Basis coefficients approach**. We conduct a MANOVA on the matrix of basis coefficients. After applying the Kolmogorov-Smirnov's test for the univariate normality of each single basis coefficients and the Kullback's test for the homogeneity covariance matrix, neither multivariate normality nor homogeneity were accepted. At this point, and due to the presence of outliers, we consider the nonparametric procedure by means of the univariate Mood's test. Keeping in mind that the related p-values were practically null, it is possible to decide that there are significant physical differences according to the type of metal for the electrode and dielectric thickness employed in the RRAM devices.

- **Principal components procedure**. Once Functional Principal Component Analysis is applied, we observe that only the first component explains more than 99% of the total variability of the process, so that the reset process can be approximated with sufficient precision by considering the Karhunen-Loève expansion in terms of the first principal component. On this subject, the homogeneity problem can be reduced to one-way ANOVA for the first principal component. Nevertheless, the parametric layout assumptions are not verified again and then, nonparametric tests must be used. In particular, univariate Mood's median test, being the associated p-value smaller than 0.001. Additionally, Wilcoxon's rank sum test is also conducted for the pairwise comparisons through the Benjamini's method for adjusting p-values. Here, all p-values are also smaller than 0.001. Thus, we can conclude that the sample mean functions are different and therefore, both material and thickness have a high impact on RRAM perfomance.

4.6 Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis

The main objective of this work is to study the impact of quarantine policies on air quality in district of Pescara-Chieti, in the Abruzzo Region (Italy). For that purpose, models for FANOVA with repeated measures and a novel approach for multivariate FANOVA for independent measures based on multivariate FPCA have been proposed. On the one hand, we carry out a univariate analysis to evaluate the behaviour of each pollutant before (BL) and during lockdown (DL), and, on the other hand, we research if the mean function of all pollutants measured in the background stations is equal to that of the urban traffic ones. In particular, four pollutants are available (NO2, PM10, PM2.5 and benzene) and five monitoring stations are considered (two traffic and three background stations). The dataset was divided in two time frames of the same length (39 days) for BL and DL. Next, the phases of the study are summarized:

- A descriptive analysis of the data was applied in order to make visible the daily variation and weekday concentrations of pollutants BL and DL. It seems that NO2 decreased DL; PM10 and PM2.5 are independent from the measures adopted during the COVID-19 nation-wide lockdown because all stations undergo an increase DL; and benzene exhibits different behaviour in background stations compared to traffic ones.

- For the functional reconstruction of pollutant curves, each sample curve was approximated in terms of a basis of cubic B-splines of dimension 20. With 20 basis functions, we capture the trend and local behaviour of the curves.

- In order to statistically analyze the effect of the lockdown on the mean of each pollutant, functional ANOVA for repeated measures is considered. In particular, the statistics $\mathcal{D}_n$ and $\mathcal{E}_n$ are contemplated to control the between and within group variability. Besides, a permutation method is used to approximate their null distributions (this method is explained in detail in Appendix A6.4.2). The results showed that there are significant differences in the mean curves of each pollutant in both time periods.

- With the objective to evaluate if the behaviour of the background stations is different in relation to the traffic stations both for BL and for DL, the multivariate functional ANOVA is carried out. Additionally, functional ANOVA for each variable separately has been also applied. For both theoretical designs, the approach based on FPCA is considered. That is, the tests consist of testing homogeneity on the vectors of the most explicative principal component scores, being four components the optimum number of principal components (we guarantee more than 99% of the total variability with four components). Besides, the extension of the univariate Kruska-Wallis'test with the permutation version is conducted in order to solve the problem of small sample size.
  - For the BL period, we found differences between both groups of stations due to PM10, although there are evidences against benzene as well.
  - For the DL period, the multivariate test is not capable of detecting differences between the groups, but the univariate functional ANOVA does stablish that the level of benzene variates according to the groups. This fact is confirmed graphically.

4.7 COVID-19 data imputation by multiple function on function principal component regression

The main goal of this work is to complete the missing values related to the curves of hospitalized and intensive care people due to COVID-19 illness in several Spanish autonomous communities between February 2nd and April 27th, 2020. Taking into account the functional nature of these response variables, a functional linear regression model can be considered to make the imputation. The predictor variables of this model will be the curves of confirmed cases, deceased and recovered people, whose evolution is completely observed over time. Given that the predictors have also a functional behaviour, we propose to use a multiple function-on-function regression linear (MFFLR) model based on principal components in order to avoid the multicollinearity problem.

Briefly, we showed in last Chapter that, if the basis expansion of sample curves is considered and we truncate each principal component decomposition, the MFFLR model is equivalent to a multivariate linear regression model in terms of a reduced set of response and predictor principal components.

Previously to the process of imputation, several steps are necessary to make the curves comparable. This is fundamental because each AC followed a different criterion for recording the data and the size of the population varies a lot according to the AC. Therefore, all curves were properly homogenized, registered and smoothed. All the details of these phases can be seen in Appendix A7.2.

Subsequently, we proceed to the imputation of the missing data. Predictor variables are completely observed for all ACs and only the response variables have missing values in some ACs. In particular, we do not have the entire information of four ACs in the curves associated with the hospitalized and in intensive care units (ICU) people. Those ACs whose evolution is totally known will be considered as training sample in the model.

- After applying the Functional Principal Component Analysis for each variable, we observe that only with the first principal component we explain almost all variability of the variables.

- In addition, only the first principal component of each predictor variable was highly and significantly correlated with the first principal component of each response variable.

- Then, the principal component MFFLR model is reduced to the linear regression of first principal component of the response in terms of the first principal component of each of the predictors.

- The model provided good outcomes for the training sample. These statement is corroborated through the determination coefficient and the square root of the mean squared errors.

- Regarding the missing curves, the model yields a pointwise estimation of the response variables. That is, we obtain an estimation of the evolution of these curves if the ACs would not have modified the way of recording at the middle of the first wave. In fact, the model captures really well the trend of the curves up to the change.

Finally, we apply a canonical correlation analysis based on the first principal component scores. The objective is to study the association between the hospital occupancy rate (HOR) and the illness response (RI). HOR is formed by the response variables used in the process of imputation, meanwhile RI is composed by those variables dealt as predictors in the model. After the analysis (see Appendix A7.4.2 for more information about the results), we conclude that both groups of variables are highly correlated each other and moreover, each of the first canonical variables have an important predictive power over the opposite set of variables.

Chapter 5

Conclusions

5.1 Phase-type distributions for studying variability in resistive memories

One of the biggest problems when Phase-type distributions are considered in applications is the estimation of the parameters. The fact that Phase-type distributions do not have a unique representation increases more difficulty to the optimization problem. For that reason, once the likelihood function is built from a sample of a Phase-type population with general representation, a solution is to use the EM algorithm to estimate the parameters and to check the efficiency of the methodology. Thanks to the great power of Phase-type distributions, we apply this technique to a major current issue in microelectronic experimental data. In the semiconductor industry, an essential aspect to better understand the RAMM internal behaviour is the analysis of the variability associated with the RRAM operation. To date, Weibull distribution is commonly used to model the current, resistance and voltage related to the variability mentioned above. Weibull distribution has worked up relatively well but the development of more sophisticated memories has produced that it becomes obsolete in last years. In this sense, the accuracy of the fitting is not as desirable as we can expect, and therefore, another distribution must be considered.

This paper introduces a new perspective that replaces the current methodology and resolves the lack of precision problem. As we anticipated before, we propose a new approach based on Phase-type distributions. Estimation and selection of parameters of the Phase-type distribution via EM algorithm provide the Erlang distribution as the best fit, for any number of phases. This means that the internal system of the devices in terms of voltage is governed by a Phase-type distribution with Erlang structure instead of general structure. Then, we conclude that the voltage until the conductive filament breaks follows an Erlang distribution. From the physical viewpoint, after the process of forming of the conductive filament, (phase 1) it undergoes a sequential degradation through a series of well differentiated phases (first parameter of the Erlang distribution), being the mean reset voltage the inverse of the second parameter of the distribution. After an exhaustive analysis, we reach the conclusion that Phase-type distributions, and in particular, Erlang distribution, achieves better results than the Weibull distribution. Therefore, we look forward that the Phase-type distributions framework will play an important role in the analysis of data experimental measures in RRAM from now on.

5.2 A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories

Many real situations require the analysis of systems whose performance depends on several macro-states that evolve over time. Although these macro-states are observable, their inner phases are not recognized beforehand. Knowing the internal behaviour it is crucial to understand how these systems work in practice. Markov chains are normally conducted to model the stochastic process and then, the exponential distribution is considered as the most suitable option to adjust the spent time in each level. Unfortunately, sometimes this methodology does not achieve an accurate fit and therefore, another perspective should be contemplated. On this matter, a novel stochastic process is introduced by considering the internal performance of macro-states in which the sojourn time in each one is Phase-type distributed depending on the initial observed time. Thus, we assume that each macro-state is composed of internal states. This new model is built in transient and stationary regimes. Additionally, measures associated with this process are also derived making use of matrix analysis, Laplace transform technique and a series of algorithms given along the work. A really important result is that the homogeneity and Markovianity are lost for the new macro-states model, but the embedded internal process maintains both properties.

This methodology has been considered in the context of RRAM memories. An essential topic related to the RRAMs is the Random Telegraph Noise signal that is produced inside the processes of operation. From the electronics perspective, it is crucial to analyze the current fluctuation between levels and the sojourn time in each of these levels. On the one hand, Hidden Markov Models have been proposed in order to determine the number of levels and to substitute the employed graphical techniques in the sector until now. Regarding the modeling of the signal, Markov models do not supply rigorous fittings when the signal is sufficiently large. So far, researchers usually consider piecewise signals to analyze them separately. However, the dependence structure and the possible relationships between the different pieces are ignored by means of this methodology. At this point, we conducted the new methodology proving that the sojourn time in each level can be modeled through a Phase-type distribution. This result is of great interest because sheds light over the device's internal process. In particular, we showed that a latent state of long RTN signal coming from a resistive memory is composed of multiple states. Hence, we can conclude that this new approach is a good candidate to replace the common statistical analysis carried out to model long RTN signals. Naturally, this methodology can be extended to the rest of memories, not only to the RRAM context.

5.3 Linear Phase-Type probability modelling of functional PCA with applications to resistive memories

Motivated by the characterization and simulation of the stochastic processes associated with the RRAM memories, a new class of distributions called Linear Phase-type distributions have been introduced in the current work. These distributions haven been developed by studying important features such as the closure and density. Its algorithmic-matrix structure makes easier the subsequent computational implementation of the results. Under these properties, it has been proved that certain linear transformations of LPH distributions are in the same class. As a consequence, if the principal components are LPH distributed, then the K-L expansion provides that the distribution of the process at each time point are Phase-type distributed and therefore, the one-dimensional distribution of the process follows a LPH distribution.

These distributions enable to model the principal components in a matrix-algorithmic form, when the processes are represented accurately through the Functional Principal Component Analysis based on the K-L expansion. Adjusting the probability distribution of the principal components is really important in order to recognise the random evolution of the process.

The results have been carried out to fit the stochastic perfomance of RRAMs. In this case, one principal component is considered and the explicit representation of the LPH is given for the stochastic process at each point.

5.4 New Modeling Approaches Based on Varimax Rotation of Functional Principal Components

Functional Principal Component Analysis is an important technique to explain the main patterns of variation in functional data. The problem attached to many situations is that the interpretation of the principal components is not always a simple task. Specially, in those applications where the most part of the explained variance is accounted only for one or two components at much. Then, a solution would be to carry out some type of rotation in the weight functions in order to synthesize the factor structure and redistribute the variability among the components. From this perspective, the orthogonal Varimax criterion is without doubt the most famous rotation for its properties and simpleness. The main drawback of Varimax rotation is that it is not able to retain the two fundamental features of FPCA: orthogonality of the weight functions and uncorrelatedness of the components. So far, we know just two mechanisms are available to make the Varimax rotation in FPCA, but neither of them is directly applied on weight functions. On the one hand, one of the method consists of rotating the matrix that contains the values of the eigenfunctions at set of time knots (**R1**). On the other hand, the second method rotates the weight function coefficients after considering a basis expansion (**R2**). In both procedures, the orthogonality is held up but the rotated scores are correlated.

Here, we propose two new approaches based on the Varimax rotation by considering the equivalence between FPCA and PCA of a transformation of the basis coefficients matrix of the curves. One is based on rotating the matrix of eigenvectors, which provides the orthogonality of the axis but not the uncorrelatedness of the scores (**R3**). The second one lies in rotating the loadings matrix of the standardized principal components, where the rotated scores are still uncorrelated but the axis are not orthogonal anymore (**R4**). After a detailed simulation study, we conclude that (**R3**) guarantees a more accurate fitting versus the other rotations that share the same characteristics (**R1** and **R2**). In addition, it is also more robust with respect to the number of discrete time observation of the sample curves. Finally, by means of the combination of (**R3**) and (**R4**), we have been able to interpret the behaviour and evolution of the number of positive cases by COVID-19 in the Spanish autonomous communities during the first wave of the pandemic in the country. We look forward that these new procedures are welcomed in future investigations in any branch of the knowledge to study the variability structure of a functional data set.

5.5 Homogeneity problem for basis expansion of functional data with applications to resistive memories

The motivation of answering to the problem about if there are significant statistically differences in the probability distribution of more than two independent sample of curves leads the current work. In particular, we focus on the situation of a functional response variable and a categorical variable that forms the groups associated with the independent samples. This layout is clearly tackled through the functional analysis of variance. Although this technique is totally recognized in many areas of knowledge, the functional nature of the dependent variable (the samples are now curves instead of vectors) makes more complicated the study. In the literature some approaches are available to solve the issue when the sample curves are generated by a Gaussian process, in which the homogeneity problem is reduced to check the equality of the group mean functions. Likewise, bootstrap techniques have been proposed for the case of lack of normality.

In this regard, we develop two new procedures by assuming the basis expansion of the sample curves. The first one is based on conducting a multivariate analysis of variance on the matrix of basis coefficients, meanwhile the second one consists of reducing the dimension of the problem by means of Functional Principal Component Analysis and to apply the MANOVA test on the vector of the most explicative principal components scores. Parametric or nonparametric solutions are given depending on whether the MANOVA typical assumptions are verified. Likewise, an extensive simulation study has been performed to test the performance of these new methodologies. The study has revealed that the sample size and dispersion parameter of errors have a high influence in both approaches, but in general, they have provided excellent results for parametric and nonparametric designs. Additionally, an application has been carried out in order to shed light about the variability behind RRAM memories operation. There are suspicions that the type of material and thicknesses used in the processes of manufacturing play an important role on RRAM operation. Taking into account that the experimental data measured on these memories are curves, the two new proposals have been considered here.

Finally, we would like to highlight the great interest that the principal component approach may awaken in applications where the dimension of the basis is large and the sample size is small. In many occasions, the FANOVA problem could be reduced to a simple MANOVA for the first q principal components, with q being small.

5.6 Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis

The current work addresses the functional ANOVA problem for two different theoretical frameworks. The first one consists of having repeated functional data of a single variable for the same subjects submitted to different conditions or whose information is taken on distinct time periods. Specifically, we dealt with the problem where the goal is to test the equality of mean functions measured on two different conditions or instants. Faced with this scenario, we extend the statistics available in the literature by considering the basis expansion of the curves. The second concern is focused on the multivariate functional ANOVA problem for independent measures. This means that there are several independent groups in which more than one functional response variable is observed. Now, the objective is to check the equality of the multivariate dimensional group mean functions. A novel approach based on multivariate FPCA has been introduced, where the problem is reduced to test multivariate homogeneity on the vectors of the most explicative principal components.

These approaches are motivated to analyze the impact of quarantine policies on air quality in the Abruzzo Region (Italy). The available data represent the evolution of four air pollutants during two different periods of time (pre and during home confinement) coming from several monitoring stations. Then, the first goal is to detect if there are significant differences between the monitoring stations classified by their geographic location (traffic and background stations). Secondly, we want to study whether the level of each of the pollutants decreased during the lockdown. The proposed functional ANOVA has proven to be beneficial to monitoring the evolution of air quality before and during the lockdown tenure and to assessing the homogeneity of groups, individuated according to the location of measuring stations.

5.7 COVID-19 data imputation by multiple function on function principal component regression

The first notified case of SARS-CoV-2 was in December in Wuhan, in central China's Hubei province. Such has been the velocity of propagation of the virus that the countries were not able to confront the illness, which has been reflected in a high number of deceased people (Dong et al., 2020). Faced this catastrophic situation, all governments in collaboration with the scientific community are attempting to make correct decisions with the objective of mitigating the COVID-19 pandemic as soon as possible. The countries adopt measurements more or less restrictive on people's life according to the predictions made by the statistical models. However, the good performance of these models depends on the quality of the data, which is not usually really satisfactory during a pandemic. Specially, it is habitual to find situations where the data are incomplete. For instance, see the case that motivates this work: a modification in the way of registering data provoked missing values in hospitalized and intensive care curves during the first wave of COVID-19 in several Spanish autonomous communities. In order to solve this problem, we propose a principal components multiple function-on-function regression model for the imputation of missing data. This approach enables to forecast the functional responses (the curves of hospitalized and intensive care people) from multiple functional predictors (the curves of positive cases, deaths and recoveries). Note that to obtain the predictions of these ACs, it is necessary to estimate the model by means of a training sample (rest of ACs) whose information is entirely known beforehand.

This functional linear regression model has displayed a suitable behaviour for the training sample, since the similarity between observed and forecasted trajectories is well for both functional responses. With regard to the predictions for the ACs that changed the way of recording, the model captures the trend of the curves up to the change. Thus, the imputation represents the temporal evolution of these ACs if they would not have modified the mode of data registering. Likewise, a canonical correlation analysis based on the first principal component scores has been carried out in order to analyze how the hospital occupancy rate (number of hospitalized people and intensive care units admissions) is connected with the illness response (number of positive cases, deaths and recovered people). We have showed that both groups of variables are highly correlated each other, being the first canonical variable a good overall predictor of the opposite set of variables. In this sense, the number of positives, deaths and hospitalized exhibited a larger predictive power than the remainder one.

Chapter 6

Conclusiones

6.1 Phase-type distributions for studying variability in resistive memories

Uno de los principales problemas que presentan las distribuciones Tipo-fase es la estimación de sus parámetros. El hecho de que no tengan una representación única añade aún más dificultad al problema de la optimización. La solución más común en la práctica es utilizar el algoritmo iterativo EM y comprobar la eficiencia del ajuste. Esta técnica es aplicada para resolver un problema de gran interés dentro del ámbito de la electrónica. En particular, la industria de los semiconductores está centrada en el análisis de la variabilidad asociada al funcionamiento de las memorias RRAM con el fin de entender mejor su comportamiento interno. Esta variabilidad es traducida en diferentes corrientes, resistencias y voltajes que suelen ser modelizados a través de la distribución Weibull. Esta distribución ha mostrado unos buenos resultados, pero el desarrollo de memorias cada vez más sofisticadas ha provocado que en los últimos años no consiga un ajuste tan deseable como cabría esperar, y por tanto, otra distribución debe ser considerada.

En este artículo se introduce una nueva perspectiva que reemplaza la metodología actual y resuelve el problema de la falta de precisión. Como se ha dicho, se propone un nuevo enfoque basado en las distribuciones Tipo-fase. Después de los procesos de estimación y selección de parámetros a través del algoritmo EM, se ha llegado a la conclusión de que la distribución Erlang es la mejor opción para modelizar el voltaje de fallo (voltaje en el que se rompe el filamento). Esto significa que el sistema interno de las memorias está gobernado por una distribución PH con estructura Erlang en lugar de con una estructura general. Por tanto, se concluye que el voltaje hasta que se rompe el filamento conductor sigue una distribución Erlang. Desde el punto de vista físico, una vez el filamento ha sido formado (fase

1) sufre un proceso de degradación de una serie de fases bien diferenciadas (primer parámetro de la distribución Erlang), siendo el voltaje medio de fallo el inverso del segundo parámetro de la distribución. Después de un análisis exhaustivo se concluye que el enfoque propuesto logra mejores resultados que la distribución Weibull. Por tanto, se espera que este marco teórico juegue un papel importante en el análisis experimental de datos procedentes de las RRAM de aquí en adelante.

6.2 A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories

Muchas aplicaciones reales requieren el análisis de sistemas cuyo comportamiento depende de varios macro-estados que evolucionan en el tiempo. Aunque estos macro-estados son observables, las fases internas no son conocidas de antemano. Conocer el comportamiento interno es esencial para entender cómo funcionan estos sistemas en la práctica. Normalmente, las cadenas de Markov son aplicadas para modelar procesos estocásticos, asumiendo que la distribución exponencial es la mejor opción para ajustar el tiempo de permanencia en cada nivel. Desgraciadamente, a veces esta metodología no logra un ajuste preciso y por tanto, otra metodología debe ser considerada. A este respecto, en este trabajo se introduce un novedoso proceso estocástico considerando el comportamiento interno de los macro-estados, en los cuales, el tiempo de permanencia en cada uno es modelizado a través de una distribución Tipo-fase dependiendo el tiempo observado inicial. Asumimos que cada macro-estado está compuesto por múltiples estados internos. Este nuevo modelo es construido en régimen estacionario y transitorio, y además, se obtienen algunas medidas asociadas haciendo uso del análisis matricial, la transformada de Laplace y una serie de algorítmicos que son dados a lo largo del trabajo. Un resultado importante acerca de este modelo es que no se cumplen ni la propiedad de la homogeneidad ni la de la Markovianidad, aunque ambas sí son verificadas para el proceso interno incrustado.

Esta metodología ha sido aplicada en el contexto de las memorias RRAM para modelizar las señales de Ruido Telegráfico Aleatorio (RTN) que son producidas dentro de las mismas memorias cuando están funcionando. Desde el punto de vista electrónico, es importante analizar las fluctuaciones de corriente entre niveles y el tiempo de permanencia en cada uno de estos niveles. Por un lado, los modelos de Markov ocultos son propuestos para determinar el número de niveles y substituir las técnicas gráficas empleadas hasta ahora en el sector. En cuanto a la modelización de la señal, los modelos de Markov no son un buen candidato cuando la señal es lo suficientemente larga. Expertos de esta área suelen considerar trozos de señales y analizarlas por separado, pero de esta manera se está ignorando la estructura de dependencia y las posibles relaciones entre las distintas partes. En este sentido, se aplica la metodología propuesta en este trabajo, obteniéndose que el tiempo de permanencia en cada nivel sigue la distribución Tipo-fase. Este resultado es de gran interés porque arroja luz sobre el proceso interno de las memorias. En particular, se muestra que un espacio latente de señales RTN largas está compuesto por múltiples estados. Por tanto, se pone de manifiesto que este enfoque, el cual se puede utilizar para otras memorias, es un buen candidato para reemplazar los actuales análisis estadísticos llevados a cabo en este campo.

6.3 Linear Phase-Type probability modelling of functional PCA with applications to resistive memories

Ante la motivación de caracterizar y simular los procesos estocásticos asociados con las memorias RRAM, en este trabajo se introduce una nueva clase distribuciones llamada distribuciones Tipo Fase Lineal (LPH). Estas distribuciones han sido desarrolladas estudiándose importantes características como la densidad o las propiedades de clausura. Además, su estructura algebraico-matricial facilita la posterior implementación de los resultados. Bajo estas propiedades, se demuestra que ciertas transformaciones de las distribuciones LPH pertenecen a la misma clase. Como consecuencia, si las componentes principales son LPH distribuidas, entonces la expansión de Karhunen-Loève proporciona que la distribución del proceso en cada instante de tiempo sigue una distribución PH, y por tanto, la distribución unidimensional del proceso sigue una distribución LPH.

Estas distribuciones permiten modelar las componentes principales de forma algebraico-matricial cuando los procesos son representados de manera precisa a través del FPCA basado en la expansión de K-L. Ajustar la distribución de probabilidad de las componentes principales es crucial para conocer la evolución aleatoria del proceso.

Estos resultados han sido aplicados para ajustar el comportamiento de las RRAM. En este caso, solo se ha necesitado una componente principal y la representación explícita de la distribución LPH ha sido dada para el proceso estocástico en cada punto.

6.4 New Modeling Approaches Based on Varimax Rotation of Functional Principal Components

El Análisis de Componentes Principales Funcional es una técnica que permite explicar los principales patrones de variación en datos funcionales. El problema adjunto en muchas situaciones es que la interpretación de las componentes principales no es siempre sencilla, especialmente, en aquellas aplicaciones donde la mayor parte de la varianza explicada recae sobre una o dos componentes a lo sumo. Entonces, una solución sería aplicar algún tipo de rotación en las autofunciones con el fin de sintetizar la estructura factorial y redistribuir la variabilidad entre las componentes. Desde esta perspectiva, la rotación Varimax es sin lugar a dudas la más famosa por sus propiedades y simpleza. El principal inconveniente de la rotación Varimax es que no es capaz de retener las dos características fundamentales del FPCA: ortogonalidad de las autofunciones e incorrelación de las puntuaciones de las componentes. Hasta la fecha, solo existen dos mecanismos para realizar la rotación en FPCA, pero ninguna de ellas son aplicadas directamente a las autofunciones. Por un lado, uno de los métodos consiste en rotar la matriz que contiene los valores de las autofunciones en un conjunto de nodos (**R1**). Por otro lado, el segundo método rota los coeficientes de las autofunciones después de considerar la expansión básica de las curvas (**R2**). En ambos procedimientos, la ortogonalidad es mantenida, pero las puntuaciones rotadas son correladas.

En el presente artículo, proponemos dos nuevos enfoques basados en la rotación Varimax considerando la equivalencia entre el FPCA y el PCA tras una transformación de la matriz que contiene los coeficientes básicos de las curvas. El primero está basado en la rotación de la matriz de autovectores, el cual garantiza la ortogonalidad de los ejes pero no la incorrelación de las puntuaciones (**R3**). El segundo rota las cargas de la matriz de las componentes estandarizadas, donde las componentes rotadas siguen siendo incorreladas pero los ejes no son ortogonales (**R4**). Tras un estudio de simulación profundo, se concluye que (**R3**) proporciona un ajuste más preciso en comparación con los otras rotaciones con las que comparte las mismas características (**R1** y **R2**). Además, es más robusta con respecto al número de observaciones en tiempo discreto de las curvas muestrales. Finalmente, combinando (**R3**) y (**R4**), se ha interpretado el comportamiento y la evolución del número de casos positivos por COVID-19 en las comunidades autónomas españolas durante la primera ola de la pandemia en el país. Se esperan que estos nuevos procedimientos sean bienvenidos en futuras investigaciones en cualquier área del conocimiento para estudiar la estructura de variabilidad de un conjunto de datos funcional.

6.5 Homogeneity problem for basis expansion of functional data with applications to resistive memories

La motivación de este trabajo radica en dar respuesta al problema sobre si existen diferencias estadísticamente significativas en la distribución de probabilidad de más de dos muestras independientes de curvas. En particular, nos centramos en la situación en la que se dispone de una variable de respuesta funcional y una variable categórica que forma los grupos asociados con las muestras independientes. Este diseño es claramente abordado a través del análisis de la varianza funcional. Aunque esta técnica es totalmente reconocida y empleada en muchas áreas del conocimiento, la naturaleza funcional de la variable dependiente (las muestras son curvas en lugar de vectores) dificulta el estudio. En la literatura se encuentran disponibles algunos enfoques para resolver la cuestión planteada cuando las muestras son generadas por un proceso Gaussiano. En este caso, el problema de la homogeneidad es reducido a evaluar la igualdad de las funciones media de los grupos. Asimismo, técnicas bootstrap son también propuestas para cuando no se verifica el supuesto de normalidad.

A este respecto, se desarrollan dos nuevos procedimientos asumiendo la expansión básica de las curvas. El primer método se basa en aplicar un análisis de la varianza multivariante sobre la matriz de coeficientes básicos, mientras que el segundo consiste en reducir la dimensión del problema utilizando el FPCA y aplicar los test MANOVA sobre los vectores de las puntuaciones de las componentes más explicativas. Se proporcionan soluciones paramétricas y no paramétricas según si se verifican o no las condiciones clásicas del MANOVA. Asimismo, se ha llevado a cabo un amplio estudio de simulación para estudiar el buen funcionamiento de estas dos nuevas metodologías. El estudio ha revelado que tanto el tamaño muestral como el parámetro de dispersión tienen una alta influencia en el desempeño de los test, pero en general, ambos enfoques logran buenos resultados para los diseños paramétricos y no paramétricos. Además, estos enfoques han sido utilizados para analizar la variabilidad que hay detrás del funcionamiento de las memorias RRAM. En particular, se aborda el problema de si el tipo de material y grosor empleado en los procesos de fabricación de las memorias, juegan un papel importante en el funcionamiento de las RRAM, cuyo comportamiento es estocástico.

Finalmente, se hace hincapié sobre el gran interés que puede despertar el enfoque de las componentes principales en aplicaciones donde la dimensión de la base es grande y el tamaño de la muestra es pequeño. En muchas ocasiones, el problema FANOVA podría ser reducido a un simple MANOVA para las primeras q componentes principales, siendo q bastante pequeño.

6.6 Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis

Este trabajo aborda el problema del ANOVA funcional desde dos marcos teóricos diferentes. El primero se ocupa de la situación en la que se dispone de datos funcionales repetidos de una variable, tras someter a los sujetos a diferentes condiciones o tras medirles la información en distintos periodos de tiempo. Concretamente, se trata el problema donde el objetivo es comprobar la igualdad de funciones medias medidas en dos condiciones o instantes diferentes. Ante este escenario, se extiende los estadísticos ya disponibles en la literatura considerando la expansión básica de las curvas. La segunda cuestión que concierne al actual artículo está centrada en resolver el problema del ANOVA funcional multivariante para medidas independientes. En este diseño se trabaja con varios grupos independientes, definidos por una variable categórica, en los que se observan más de una variable de respuesta funcional. Entonces, el objetivo es evaluar la igualdad de los vectores de las funciones media de los grupos. Para tal propósito, se ha introducido un novedoso enfoque basado en el FPCA multivariante, donde el problema se reduce a aplicar los test de homogeneidad multivariante en los vectores de las componentes principales más explicativas.

Ambos enfoques vienen motivados con el fin de analizar el impacto que han tenido las políticas de confinamiento en la calidad del aire en la Región de Abruzzo (Italia). Los datos disponibles representan la evolución temporal de cuatro contaminantes durante dos períodos de tiempo (antes y durante el confinamiento domiciliario). Estos datos han sido recogidos por varias estaciones de monitoreo. Por tanto, el primer objetivo es detectar si existen diferencias significativas entre las estaciones de monitoreo clasificadas según su localización geográfica (estaciones situadas en zonas con mucho tráfico o en las afueras). En segundo lugar, se quiere estudiar si el nivel de cada uno de los contaminantes se redujo durante el periodo de confinamiento. Los métodos propuestos han resultado ser beneficiosos para monitorear la evolución de la calidad del aire antes y durante el confinamiento, así como para evaluar la homogeneidad de los grupos, individualizados según la ubicación de las estaciones de medición.

6.7 COVID-19 data imputation by multiple function on function principal component regression

El primer caso notificado de SARS-CoV-2 fue en Diciembre en Wuhan, en la provincia china de Hubei. Tal ha sido la velocidad de propagación del virus que los países no han sido capaces de hacer frente a la enfermedad, lo cual se ha visto reflejado en un alto número de fallecidos en todo el mundo (Dong et al., 2020). Ante esta catastrófica situación, todos los gobiernos en colaboración con la comunidad científica están intentando tomar decisiones correctas con el objetivo de frenar la pandemia lo antes posible. Los países adoptan medidas más o menos restrictivas en la vida de las personas según las predicciones que obtienen a través de los modelos estadísticos. Sin embargo, el buen desempeño de estos modelos depende de la calidad de los datos, que no suele ser muy satisfactoria en épocas de pandemia.

Especialmente, es habitual encontrar situaciones en los que los datos estén incompletos. Por ejemplo, véase el caso que motiva este trabajo: una modificación en el cambio de registro de los datos provocó valores faltantes en las curvas de personas hospitalizadas y en cuidados intensivos durante la primera ola de COVID-19 en varias comunidades autónomas españolas. Con el fin de resolver este problema, se propone un modelo de regresión función-sobre-función múltiple en componentes principales para la imputación de datos faltantes. Este enfoque permite predecir las respuestas funcionales (curvas de personas hospitalizadas y en cuidados intensivos) a partir de múltiple predictores funcionales (las curvas de casos positivos, fallecidos y recuperados). Téngase en cuenta que para obtener las predicciones de estas comunidades, es necesario estimar el modelo a través de una muestra de entrenamiento (el resto de comunidades) cuya información es totalmente conocida de antemano.

Este modelo de regresión lineal funcional ha mostrado un comportamiento adecuado para la muestra de entrenamiento, ya que las trayectorias observadas y predichas son bastantes similares para ambas respuestas funcionales. Con respecto a las predicciones para aquellas comunidades que cambiaron la forma de registro, el modelo captura la tendencia de las curvas hasta el cambio mencionado. Por tanto, la imputación representa la evolución temporal de estas comunidades si ellas no hubieran modificado el modo de registro. Asimismo, un análisis de correlaciones canónicas basado en las puntuaciones del primer componente principal ha sido llevado a cabo para analizar la relación entre la tasa de ocupación hospitalaria (número de personas hospitalizadas y en cuidados intensivos) y la respuesta de la enfermedad (número de positivos, fallecidos y recuperados). Se ha mostrado que ambos grupos de variables están altamente correlados entre sí, siendo la primera variable canónica un buen predictor para el conjunto de variables opuesto. En este sentido, el número de positivos, fallecidos y hospitalizados exhiben un poder predictivo más importante que el resto.

Open research lines

We briefly describe the main current research lines in which we are working in keeping with the results and conclusions obtained throughout this thesis.

- **Appendix A1**. Sometimes the fitting by the PH distributions presents certain weaknesses in distribution tails, or even, being suitable the adjustment, the number of parameters to be estimated is really high. In this sense, a solution is to develop a new methodology based on the mixture of PH distributions. A new distribution called the multiple cut-point phase type distribution will be introduced. This new distribution will enable to reduce the number of parameters in the estimate. For instance, if only one cut point is required, the number of parameters to estimate would be three in an Erlang distribution (the value of the cut point and the value for each Erlang distribution). Several measures such as the Laplace transform and therefore the moments will be studied. The EM-algorithm will be considered for the estimation. A parallel study will be conducted for the discrete case.

- **Appendix A2**. Neither homogeneity nor Markovianity are verified for the macro-state model developed in the current work. The objective is to develop new non-homogenous Markov processes and semi Markov processes with PH distributions. Afterwards, the models would be compared.

- **Appendix A3**. So far, the RRAM's set and reset processes have been tackled separately in the literature. However, there is a dependence structure between each other, since a set cycle takes place once the reset cycle has finished and vice versa. Then, the efforts will be focused on the joint modeling of reset-set cycles by using mixed multivariate ARIMA-FPCA models.

- **Appendix A4**. In this work, two new Varimax rotation for functional data have been introduced. These approaches are based on the equivalence between FPCA and PCA of a transformation of the matrix of basis coefficients. This result can be generalized for different types of rotation such as the oblique rotations.

- **Appendix A5-A6**. A novel approach based on FPCA is proposed in the currents articles to address the FANOVA problem for independent groups, both for the univariate and for multivariate design. These methodologies will be extended for repeated measures to solve current problems in biomechanics. In fact, the first results have been obtained for solving the two-way FANOVA problem with repeated measures in one of the categorical predictors. This methodology has been applied to test if there are significant statistical differences between gait curves of children going to school by using trolleys or backpacks with a considerable load (data from Sport and Health Institute of the UGR). This work is currently being written and will be submitted for publication soon.

- **Appendix A7**. The imputation of the missing data for each functional response variable has been carried out separately. This fact is ignoring possible relationships between the response variables. Therefore, one of the first subject to address in the near future is the generalization of the function-on-function models for a multivariate scenario.

Bibliography

[1] C. Acal, J.E. Ruiz-Castro, and A.M. Aguilera. Distribuciones tipo fase en un estudio de fiabilidad. *TEMat*, 3:63–74, 2019.

[2] A. Agarwal, A. Kaushik, S. Kumar, and R.K Mishra. Comparative study on air quality status in indian and chinese cities before and during the covid-19 lockdown period. *Air Quality, Atmosphere & Health*, 13:1167–11178, 2020.

[3] P. Agarwal and K. Jhajharia. Data analysis and modeling of covid-19. *Journal of Statistics and Management Systems*, 24(1):1–16, 2021.

[4] A. M. Aguilera and M. C. Aguilera-Morillo. Penalized PCA approaches for B-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219(14):7805–7819, 2013.

[5] A. M. Aguilera, M. Escabias, F. A. Ocaña, and M. J. Valderrama. Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress. *Methodology and Computing in Applied Probability*, 17(4):1015–1028, 2015.

[6] A. M. Aguilera, M. Escabias, C. Preda, and G. Saporta. Using basis expansions for estimating functional pls regression: Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305, 2010.

[7] A. M. Aguilera, F. Fortuna, M. Escabias, and T. Di Battista. Assessing social interest in burnout using google trends data. *Social Indicators Research, in press*, 2020.

[8] A. M. Aguilera, F. A. Ocaña, and M. J. Valderrama. Forecasting with unequally spaced data by a functional principal component approach. *Test*, 8(1):233–254, 1999.

[9] A.M. Aguilera, M.C. Aguilera-Morillo, and C. Preda. Penalized versions of functional pls regression. *Chemometrics and Intelligent Laboratory Systems*, 154:80–92, 2016.

[10] M. C. Aguilera-Morillo and A. M. Aguilera. Multi-class classification of biomechanical data: A functional lda approach based on multi-class penalized functional pls. *Statistical Modelling*, 20(6):592–616, 2020.

[11] M. C. Aguilera-Morillo, A. M. Aguilera, M. Escabias, and M. Valderrama. Penalized spline approaches for functional logit regression. *TEST*, 22(2):251–277, 2013.

[12] M. C. Aguilera-Morillo, A. M. Aguilera, F. Jiménez-Molinos, and J. B. Roldán. Stochastic modeling of random access memories reset transitions. *Mathematics and Computers in Simulation*, 159(1):197–209, 2019.

[13] M.C. Aguilera-Morillo, A.M. Aguilera, and M. Durban. Prediction of functional data with spatial dependence: a penalized approach. *Stochastic Environmental Research and Risk Assessment*, 31:7–22, 2017.

[14] S. Aldana, P. García-Fernández, A. Rodríguez-Fernández, R. Romero-Zaliz, M. B. González, F. Jiménez-Molinos, F. Campabadal, F. Gómez-Campos, and J. B. Roldán. A 3d kinetic monte carlo simulation study of resistive switching processes in Ni/HfO$_2$/Si-n^2-based rrams. *Journal of Physics D: Applied Physics*, 50(33):335103, 2017.

[15] L. J. Allen. *An introduction to stochastic processes with applications to biology*. CRC Press, 2010.

[16] F. J. Alonso, D. Maldonado, A. M. Aguilera, and J. B. Roldán. Memristor variability and stochastic physical properties modeling from a multivariate time series approach. *Chaos, Solitons and Fractals*, 143:110461, 2021.

[17] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[18] Y. Araki, S. Konishi, S. Kawano, and H. Matsui. Functional logistic discrimination via regularized basis expansions. *Communications in Statistics—Theory and Methods*, 38:2944–2957, 2009.

[19] J. R. Artalejo and S. R. Chakravarthy. Algorithmic analysis of themap/ph/1 retrial queue. *Top*, 14(2):293–332, 2006.

[20] S. Asmussen. *Ruin probabilities*. World Scientific, 2000.

[21] S. Asmussen and M. Bladt. Phase-type distributions and risk processes with state-dependent premiums. *Scandinavian Actuarial Journal*, 1996:19–36, 1996.

[22] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[23] M. C. Ausın, M. P. Wiper, and R. E. Lillo. Bayesian estimation for the m/g/1 queue using a phase-type approximation. *Journal of Statistical Planning and Inference*, 118(1):83–101, 2004.

[24] V. S. Barbu, A. Karagrigoriou, and A. Makrides. On semi-markov modeling and inference for multi-state systems. In *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)*, 2016.

[25] K. Benhenni, F. Ferraty, M. Rachdi, and P. Vieu. Local smoothing regression with functional data. *Computational Statistics*, 22(3):353–369, 2007.

[26] S. K. Berberian. *Introduction to Hilbert space (Vol. 287)*. American Mathematical Soc., 1999.

[27] A. Berihuete, M. Sanchez-Sanchez, and A. Suarez-Llorens. A bayesian model of covid-19 cases based on the gompertz curve. *Mathematics*, 9(3):228, 2021.

[28] J. D. Berman and K. Ebisu. Changes in u.s. air pollution during the covid-19 pandemic. *Science of The Total Environment*, 739:139864, 2020.

[29] J.R. Berrendero, J.R. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619–2634, 2011.

[30] Z. Biolek, D. Biolek, and V. Biolkova. Spice model of memristor with nonlinear dopant drift. *Radioengineering*, 18(2):210–214, 2009.

[31] A. Briz-Redon. The impact of modelling choices on modelling outcomes: a spatio-temporal study of the association between covid-19 spread and environmental conditions in catalonia (spain). *Stochastic Environmental Research and Risk Assessment*, 2021.

[32] P. Buchholz, J. Kriege, and I. Felko. *Input modeling with phase-type distributions and Markov models, Theory and Applications*. Springer, 2014.

[33] H. Cardot. Nonparametric estimation of the smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538, 2000.

[34] C. Carroll, S. Bhattacharjee, Y. Chen, P. Dubey, J. Fan, A. Gajardo, X. Zhou, H. G. Müller, and J. L. Wang. Time dynamics of covid-19. *Scientific reports*, 10:21040, 2020.

[35] G. Caruso, S. A. Gattone, F. Fortuna, and T. Di-Battista. Cluster analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences*, 73:100850, 2021.

[36] X. Chen, L. Wang, B. Li, Y. Wang, X. Li, Y. Liu, and H. Yang. Modeling random telegraph noise as a randomness source and its application in true random number generation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 35(9):1435–1448, 2016.

[37] J. M. Chiou, H. G. Müller, and J. L. Wang. Functional response models. *Statistica Sinica*, 14(3):659–677, 2004.

[38] C. K. Chui. *An introduction to wavelets*. Elsevier, 2016.

[39] C. Crambes and Y. Henchiri. Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, 201:103–119, 2019.

[40] J. A. Cuesta-Albertos and M. Febrero-Bande. A simple multiway anova for functional data. *Test*, 19(3):537–557, 2010.

[41] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational Statistics and Data Analysis*, 47(1):111–122, 2004.

[42] M. Davidian, X. Lin, and J. L. Wang. Introduction: emerging issues in longitudinal and functional data analysis. *Statistica Sinica*, 14(3):613–614, 2004.

[43] C. De Boor. *A practical guide to splines (revised edition)*. Springer, 2001.

[44] D. M. Deaves and I. G. Lines. On the fitting of low mean windspeed data to the weibull distribution. *Journal of Wind Engineering and Industrial Aerodynamics*, 66(3):169–178, 1997.

[45] A. Delaigle and P. Hall. Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, 40(1):322–352, 2012.

[46] P. Delicado. Functional k-sample problem when data are density functions. *Computational Statistics*, 22:391–410, 2007.

[47] J. C. Deville. Estimation of the eigenvalues and of the eigenvectors of a covariance operator. *Note interne de l'INSEE*, 1973.

[48] J. C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, 15:3–101, 1974.

[49] T. Di-Battista and F. Fortuna. Functional confidence bands for lichen biodiversity profiles: A case study in tuscany region (central italy). *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(1):21–28, 2017.

[50] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[51] M. Durban. *Splines con penalizaciones: Teoría y aplicaciones.* Universidad Pública de Navarra, 2007.

[52] M. Durban. An introduction to Smoothing with Penalties: P-splines. *Boletín de la sociedad Española de Estadística e Investigación Operativa (BEIO)*, 25(3):195–205, 2009.

[53] A. R. Ellis, W. W. Burchett, S. W. Harrar, and A. C. Bathke. Nonparametric inference for multivariate data: the r package npmv. *Journal of Statistical Software*, 76(4):1–18, 2017.

[54] B. Epstein and M. Sobel. Life testing. *Journal of the American Statistical Association*, 48(263):486–502, 1953.

[55] M. Escabias, A. M. Aguilera, and M. C. Aguilera-Morillo. Functional pca and base-line logit models. *Journal of Classification*, 31(3):296–324, 2014.

[56] M. Escabias, A. M. Aguilera, and M. J. Valderrama. Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4):365–384, 2004.

[57] M. Escabias, A.M. Aguilera, and M.J. Valderrama. Functional pls logit regression model. *Computational Statistics* & *Data Analysis*, 51(10):4891–4902, 2007.

[58] M. Escabias, M. J. Valderrama, A. M. Aguilera, M. E. Satofimia, and M. C. Aguilera-Morillo. Stepwise selection of functional covariates in forecasting peak levels of olive pollen. *Stochastic Environmental Research and Risk Assessment*, 27(2):367–376, 2013.

[59] M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. *Computational Statistics and Data Analysis*, 131:91–103, 2019.

[60] F. Ferraty, M. Sued, and P. Vieu. Mean estimation with data missing at random for functional covariables. *Statistics*, 47(4):688–706, 2013.

[61] F. Ferraty and P. Vieu. *Nonparametric functional data analysis. Theory and practice.* Springer-Verlag, 2006.

[62] R. Flores, R. E. Lillo, and J. Romo. Homogeneity test for functional data. *Journal of Applied Statistics*, 45(5):868–883, 2018.

[63] F. Fortuna and F. Maturo. K-means clustering of item characteristic curves and item information curves via functional principal component analysis. *Quality & Quantity*, 53(5):2291–2304, 2019.

[64] F. Fortuna, F. Maturo, and T. Di-Battista. Clustering functional data streams: Unsupervised classification of soccer top players based on google trends. *Quality and Reliability Engineering International*, 34(7):1448–1460, 2018.

[65] P. Galeano, E. Joseph, and R. E. Lillo. The mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291, 2015.

[66] S. Gautam and U. Trivedi. Global implications of bio aresol in pandemic. *Environment, Development and Sustainability*, 22:3861–3865, 2020.

[67] G. González-Cordero, M. B. González, A. Morell, F. Jiménez-Molinos, F. Campabadal, and J. B. Roldán. Neural network based analysis of random telegraph noise in resistive random access memories. *Semiconductor Science and Technology*, 35(2):025021, 2020.

[68] G. González-Cordero, J. B. Roldán, F. Jiménez-Molinos, J. Suñé, S. Long, and M. Liu. A new compact model for bipolar rrams based on truncated cone conductive filaments, a verilog-a approach. *Semiconductor Science and Technology,*, 31(11):115013, 2016.

[69] G. González-Cordero, M. González, F. Jiménez-Molinos, F Campabadal, and J. B. Roldán. New method to analyze random telegraph signals in resistive random access memories. *Journal of Vacuum Science & Technology B*, 37(1):012203, 2019.

[70] A. S. Gordon, A. H. Marshall, and M. Zenga. Predicting elderly patient length of stay in hospital and community care using a series of conditional coxian phase-type distributions, further conditioned on a survival tree. *Health care management science*, 21(2):269–280, 2018.

[71] T. Gorecki, M. Krzysko, and L. Waszak. Functional discriminant coordinates. *Communications in Statistics-Theory and Methods*, 43(5):1013–1025, 2014.

[72] T. Gorecki and L. Smaga. Comparison of tests for the one-way anova problem for functional data. *Computational Statistics*, 30(4):987–1010, 2015.

[73] T. Gorecki and L. Smaga. Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, 44(12):2172–2189, 2017.

[74] J. W. Graham. *Missing data: Analysis and design.* Springer Science & Business Media, 2012.

[75] T. Grasser. *Noise in Nanoscale Semiconductor Devices.* Springer Nature, 2020.

[76] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models.* Monographs on Statistics and applied probability. Chapman & Hall, 1994.

[77] S. L. Greer, E. J. King, E. M. da Fonseca, and A. Peralta-Santos. The comparative politics of covid-19: The need to understand government responses. *Global Public Health*, 15(9):1413–1416, 2020.

[78] B. Gregorutti, B. Michel, and P. Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35, 2015.

[79] M. Guo, L. Zhou, J. Z. Huang, and W. K. Hardle. Functional data analysis of generalized regression quantiles. *Statistics and Computing*, 25(2):189–202, 2015.

[80] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006.

[81] Q. M. He. *Fundamentals of matrix-analytic methods.* Springer, 2014.

[82] Y. He, R. Yucel, and T. E. Raghunathan. A functional multiple imputation approach to incomplete longitudinal data. *Statistics in medicine*, 30(10):1137–1156, 2011.

[83] J. Henriquez, E. Gonzalo-Almorox, M. Garcia-Goñi, and F. Paolucci. The first months of the covid-19 pandemic in spain. *Health Policy and Technology*, 9(4):560–574, 2020.

[84] S. Hörmann, L. Kidziǹski, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 77(2):319–348, 2015.

[85] L. Horvath and P. Kokoszka. *Inference for functional data with applications.* Springer-Verlag, 2012.

[86] D. Ielmini and R. Waser. *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications.* Wiley-VCH, 2015.

[87] J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71:92–106, 2014.

[88] J. Jacques and C. Preda. Functional data clustering: a survey. *Advanced Data Analysis and Classification*, 8:231–255, 2014b.

[89] G. M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society. Series B*, 64(3):411–432, 2002.

[90] G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000.

[91] F. Jiménez-Molinos, M. Villena, J. B. Roldán, A. M. Roldán, et al. A spice compact model for unipolar rram reset process analysis. *Electron Devices, IEEE Transactions on*, 62(3):955–962, 2015.

[92] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):319–351, 1997.

[93] I. T. Jolliffe. *Principal Component Analysis (Second Edition).* Springer, 2002.

[94] H. Kano, H. Fujioka, and C. F. Martin. Optimal smoothing and interpolating splines with constraints. *Applied Mathematics and Computation*, 218(5):1831–1844, 2011.

[95] H. Kano, H. Nakata, and C. F. Martin. Optimal curve fitting and smoothing using normalized uniform B-splines: a tool for studying complex systems. *Applied Mathematics and Computation*, 169(1):96–128, 2005.

[96] I. Keser and I. Kocakoç. Smoothed functional canonical correlation analysis of humidity and temperature data. *Journal of Applied Statistics*, 42(10):2126–2140, 2015.

[97] M. Kijima. *Markov processes for stochastic modeling.* Springer, 2013.

[98] M. Krzysko and L. Waszak. Canonical correlation analysis for functional data. *Biometrical Letters*, 50(2):95–105, 2013.

[99] V. G. Kulkarni. *Modeling and analysis of stochastic systems.* Crc Press, 2016.

[100] M. Lanza. A review on resistive switching in high-k dielectrics: A nanoscale point of view using conductive atomic force microscope. *Materials*, 7(3):2155–2182, 2014.

[101] J. F. Lawless. *Statistical models and methods for lifetime data (Second Edition)*. John Wiley & Sons, 2003.

[102] H. Lian. Empirical likelihood confidence intervals for nonparametric functional data analysis. *Journal of Statistical Planning and Inference*, 142(7):1669–1677, 2012.

[103] N. Ling, L. Liang, and P. Vieu. Nonparametric regression estimation for functional stationary ergodic data with missing at random. *Journal of Statistical Planning and Inference*, 162:75–87, 2015.

[104] N. Ling, Y. Liu, and P. Vieu. Conditional mode estimation for functional stationary ergodic data with responses missing at random. *Statistics*, 50:991–1013, 2016.

[105] A. Lisnianski, I. Frenkel, and A. Karagrigoriou. *Recent advances in multi-state systems reliability: Theory and applications*. Springer, 2018.

[106] A. Lisnianski and G. Levitin. *Multi-state system reliability: assessment, optimization and applications*. World scientific, 2003.

[107] R. J. Little and D. B Rubin. *Statistical analysis with missing data (Third Edition)*. John Wiley & Sons, 2019.

[108] J. Liu, J. Chen, and D. Wang. Wavelet functional principal component analysis for batch process monitoring. *Chemometrics and intelligence laboratory systems*, 196, 2020.

[109] S. Long, C. Cagli, D. Ielmini, M. Liu, and J. Sune. Analysis and modeling of resistive switching statistics. *Journal of Applied Physics*, 111(7):074508, 2012.

[110] S. Long, X. Lian, T. Ye, C. Cagli, L. Perniola, E. Miranda, M. Liu, and J. Suñé. Cycle-to-cycle intrinsic reset statistics in hfo$_2$-based unipolar rram devices. *IEEE Electron Device Letters*, 34(5):623–625, 2013.

[111] W. C. Luo, J. C. Liu, H. T. Feng, Y. C. Lin, J. J. Huang, K. L. Lin, and T. H. Hou. Rram set speed-disturb dilemma and rapid statistical prediction methodology. In *2012 International Electron Devices Meeting*, pages 9.5.1–9.5.4, 2012.

[112] S. Mahato, S. Pal, and K.G. Ghosh. Effect of lockdown amid covid-19 pandemic on air quality of the megacity delhi, india. *Science of the Total Environment*, 730:139086, 2020.

[113] A. H. Marshall and M. Zenga. Simulating coxian phase-type distributions for patient survival. *International Transactions in Operational Research*, 16(2):213–226, 2009.

[114] A. H. Marshall and M. Zenga. Experimenting with the coxian phase-type distribution to uncover suitable fits. *Methodology and computing in applied probability*, 14(1):71–86, 2012.

[115] A. H. Marshall, M. Zenga, and S. Giordano. Modelling students' length of stay at university using coxian phase-type distributions. *International Journal of Statistics and Probability*, 2(1):73, 2013.

[116] P. Martínez-Camblor and N. Corral. Repeated measures analysis for functional data. *Computational Statistics and Data Analysis*, 55:3244–3256, 2011.

[117] A. Mauri, R. Sacco, and M. Verri. Electro-thermo-chemical computational models for 3d heterogeneous semiconductor device simulation. *Applied Mathematical Modelling*, 39(14):4057–4074, 2015.

[118] J. W. McPherson. *Reliability physics and engineering: time-to-failure modeling*. Springer Science & Business Media, 2013.

[119] L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen. Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, 18(2):195–222, 2009.

[120] J. C. Mora, S. Pérez, and A. Dvorzhak. Application of a semi-empirical dynamic model to forecast the propagation of the covid-19 epidemics in spain. *Forecasting*, 2(4):452–469, 2020.

[121] H. G. Müller and U. Stadtmüller. Generalized functional linear models. *Annals of Statistics*, 33(2):774–805, 2005.

[122] M. F. Neuts. *Probability distributions of phase type*. Liber Amicorum Prof. Emeritus H. Florin, 1975.

[123] M. F. Neuts. A versatile markovian point process. *Journal of Applied Probability*, 16(4):764–779, 1979.

[124] M. F. Neuts. Matrix geometric solutions in stochastic models: An algorithmic approach. In *Probability Distributions of Phase Type*. John Hopkins University Press, Baltimore, 1981.

[125] F. A. Ocaña, A. M. Aguilera, and M. Escabias. Computational considerations in functional principal component analysis. *Computational Statistics*, 22(3):449–465, 2007.

[126] F. A. Ocaña, A. M. Aguilera, and M. J. Valderrama. Functional Principal Components Analysis by Choice of Norm. *Journal of Multivariate Analysis*, 71(2):262–276, 1999.

[127] H. Oja. *Multivariate Nonparametric Methods with R*. Springer Science & Business Media, 2010.

[128] F. Pan, S. Gao, C. Chen, C. Song, and F. Zeng. Recent progress in resistive random access memories: materials, switching mechanisms and performance. *Materials Science and Engineering*, 83:1–59, 2014.

[129] A. Park, S. Guillas, and I. Petropavlovskikh. Trends in stratospheric ozone profiles using functional mixed models. *Atmospheric Chemistry and Physics*, 13(22):11473–11501, 2013.

[130] R. Picos, J. B. Roldán, M. N. Al Chawa, F. Jiménez-Molinos, M. A. Villena, and E. García-Moreno. Exploring ReRAM-based memristors in the charge-flux domain, a modeling approach. In *Proceedings of International Conference on Memristive Systems, MEMRISYS'2015*, 2015.

[131] C. Preda and G. Saporta. Pls regression on a stochastic process. *Computational Statistics & Data Analysis*, 48(1):149–158, 2005.

[132] C. Preda, G. Saporta, and C. Leveder. Pls classification of functional data. *Computational Statistics*, 22:223–235, 2007.

[133] F. M. Puglisi. Noise in resistive random access memory devices. In *Noise in Nanoscale Semiconductor Devices*, pages 87–133. Springer International Publishing, 2020.

[134] F. M. Puglisi, L. Larcher, A. Padovani, and P. Pavan. A complete statistical investigation of rtn in hfo2-based rram in high resistive state. *IEEE Transactions on Electron Devices*, 62(8):2606–2613, 2015.

[135] F. M. Puglisi, N. Zagni, L. Larcher, and P. Pavan. Random telegraph noise in resistive random access memories: Compact modeling and advanced circuit design. *IEEE Transactions on Electron Devices*, 65(7):2964–2972, 2018.

[136] H. Qi, S. Xiao, R. Shi, M. P. Ward, Y. Chen, W. Tu, Q. Su, W. Wang, X. Wang, and Z. Zhang. Covid-19 transmission in mainland china is associated with temperature and humidity: A time-series analysis. *Science of The Total Environment*, 728:138778, 2020.

[137] X. Qi and R. Luo. Function-on-function regression with thousands of predictive curves. *Journal of Multivariate Analysis*, 163:51–66, 2018.

[138] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[139] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[140] P. Ramírez-Cobo, R. E. Lillo, and M. P. Wiper. Nonidentifiability of the two-state markovian arrival process. *Journal of Applied Probability*, 47(3):630–649, 2010.

[141] J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer-Verlag, 2009.

[142] J. O. Ramsay and B. W. Silverman. *Functional data analysis.* Springer-Verlag, 1997.

[143] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: Methods and case studies.* Springer-Verlag, 2002.

[144] J. O. Ramsay and B. W. Silverman. *Functional data analysis (Second Edition).* Springer-Verlag, 2005.

[145] A. R. Rao and M. Reimherr. Modern multiple imputation with functional data. *Stat*, 10(1):e331, 2021.

[146] A. C. Rencher and W. F. Christensen. *Methods of multivariate analysis (Third Edition).* Wiley, 2012.

[147] J. B. Roldán, E. Miranda, G. González-Cordero, P. García-Fernández, R. Romero-Zaliz, P. González-Rodelas, A. M. Aguilera, M. B. González, and F. Jiménez-Molinos. Multivariate analysis and extraction of parameters in resistive rams using the quantum point contact model. *Journal of Applied Physics*, 123(1):014501, 2018.

[148] S. M. Ross, J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, ..., and V. L. Bristow. *Stochastic processes (Vol. 2).* New York: Wiley, 1996.

[149] J. E. Ruiz-Castro. Preventive maintenance of a multi-state device subject to internal failure and damage due to external shocks. *IEEE Transactions on Reliability*, 63(2):646–660, 2014.

[150] J. E. Ruiz-Castro. A preventive maintenance policy for a standby system subject to internal failures and external shocks with loss of units. *International Journal of Systems Science*, 46(9):1600–1613, 2015.

[151] J. E. Ruiz-Castro. Complex multi-state systems modelled through marked markovian arrival processes. *European Journal of Operational Research*, 252(3):852–865, 2016.

[152] J. E. Ruiz-Castro. A complex multi-state k-out-of-n: G system with preventive maintenance and loss of units. *Reliability Engineering and System Safety*, 197:106797, 2020.

[153] J. E. Ruiz-Castro. Optimizing a multi-state cold-standby system with multiple vacations in the repair and loss of units. *Mathematics*, 9(8):913, 2021.

[154] J. E. Ruiz-Castro, C. Acal, and A. M. Aguilera. Phase-type distributions: computational aspects and applications in electronics. *Boletín de Estadística e Investigación Operativa*, 37:3–18, 2021.

[155] J. E. Ruiz-Castro, M. Dawabsha, and F. J. Alonso. Discrete-time markovian arrival processes to model multi-state complex systems with loss of units and an indeterminate variable number of repairpersons. *Reliability Engineering & System Safety*, 174:114–127, 2018.

[156] J. E. Ruiz-Castro and Q. L. Li. Algorithm for a general discrete k-out-of-n: G system subject to several types of failure with an indefinite number of repairpersons. *European Journal of Operational Research*, 211(1):97–111, 2011.

[157] J. E. Ruiz-Castro and Dawabsha M. A multi-state warm standby system with preventive maintenance, loss of units and an indeterminate multiple number of repairpersons. *Computers and Industrial Engineering*, 142:106348, 2020.

[158] J. E. Ruiz-Castro and M. Zenga. A general piecewise multi-state survival model: application to breast cancer. *Statistical Methods and Applications*, 29:813–843, 2020.

[159] D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757, 2002.

[160] L. Santamaría and J. Hortal. Covid-19 effective reproduction number dropped during spain's nationwide dropdown, then spiked at lower-incidence regions. *Science of The Total Environment*, 751:142257, 2021.

[161] A. Schmutz, J. Jacques, C. Bouveyron, L. Cheze, and P. Martin. Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 35:1101–1131, 2020.

[162] Q. Shen and J. Faraway. An f test for linear models with functional responses. *Statistica Sinica*, 14:1239–1257, 2004.

[163] S. Shin, K. Kim, and S. M. Kang. Compact models for memristors based on charge-flux constitutive relationships. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 29(4):590–598, 2010.

[164] B. W. Silverman. Smoothed functional principal component analysis by choice of norm. *Annal Statistics*, 24(1):1–24, 1996.

[165] E. Simoen and C. L. Claeys. *Random telegraph signals in semiconductor devices*. Bristol: IOP Publishing, 2016.

[166] C. A. D. S. Siqueira, Y. N. L. D. Freitas, M. D. C. Cancela, M. Carvalho, A. Oliveras-Fabregas, and D. L. B. de Souza. The effect of lockdown on the outcomes of covid-19 in spain: An ecological study. *PLOS ONE*, 15(7):1–13, 2020.

[167] L. Smaga. Repeated measures analysis for functional data using box-type approximation – with applications. *REVSTAT-Statistical Journal*, 17(4):523–549, 2019.

[168] L. Smaga. A note on repeated measures analysis for functional data. *AStA Advances in Statistical Analysis*, 104:117–139, 2020.

[169] A. J. Suárez and S. Ghosal. Bayesian estimation of principal components for functional data. *Bayesian Analysis*, 12(2):311–333, 2017.

[170] C. Tang, T. Wang, and P. Zhang. Functional data analysis: An application to covid-19 data in the united states, 2020.

[171] H. M. Taylor and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, 1994.

[172] A. Tobias. Evaluation of the lockdowns for the sars-cov-2 epidemic in italy and spain after one month follow up. *Science of The Total Environment*, 725:138539, 2020.

[173] P. Todorovic. *An Introduction to Stochastic Processes and Their Applications*. Springer-Verlag New York, 1992.

[174] S. Tokushige, H. Yadohisa, and K. Inada. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22:1–16, 2007.

[175] C. Torres-Martín, C. Acal, M. El-Homrani, and A. C. Mingorance-Estrada. Impact on the virtual learning environment due to covid-19. *Sustainability*, 13(2), 2021.

[176] A. Torres-Signes, M. P. Frías, and M. D. Ruiz-Medina. Covid-19 mortality analysis from soft-data multivariate curve regression and machine learning, 2021.

[177] T. Tsuruoka, K. Terabe, T. Hasegawa, and M. Aono. Forming and switching mechanisms of a cation-migration-based oxide resistive memory. *Nanotechnology*, 21(42):425205, 2010.

[178] M. J. Valderrama, F. A. Ocaña, A. M. Aguilera, and F. M. Ocaña-Peinado. Forecasting pollen concentration by a two–step functional model. *Biometrics*, 66(2):578–585, 2010.

[179] A. Van der Linde. Variational bayesian functional PCA. *Computational Statistics and Data Analysis*, 53(2):517–533, 2008.

[180] M. A. Villena, J. B. Roldán, M. B. González, P. González-Rodelas, F. Jiménez-Molinos, F. Campabadal, and D. Barrera. A new parameter to characterize the charge transport regime in Ni/HfO2/Si-n+ based rrams. *Solid-State Electronics*, 118:56–60, 2016.

[181] M. A. Villena, J. B. Roldán, F. Jiménez-Molinos, E. Miranda, J. Suñé, and M. Lanza. Sim^{2}rram: A physical model for rram devices simulation. *Journal of Computational Electronics*, 16(4):1095–1120, 2017.

[182] R. Waser. *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Devices (Third Edition).* Wiley, 2012.

[183] R. Waser and M. Aono. Nanoionics-based resistive switching memories. *Nature materials*, 6(11):833–840, 2007.

[184] F. Yao, H. G. Müller, and J. L. Wang. Functional data analysis for sparse longitudinal data. *Journal of American Statistical Association*, 100(470):577–590, 2005.

[185] N. Young. *An introduction to Hilbert space.* Cambridge university press, 1988.

[186] M. A. Zambrano-Monserrate, M. A. Ruano, and L. Sanchez-Alcalde. Indirect effects of covid-19 on the environment. *Science of The Total Environment*, 728:138813, 2020.

[187] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun. Deep learning methods for forecasting covid-19 time-series data: A comparative study. *Chaos, Solitons & Fractals*, 140:110121, 2020.

[188] J. T. Zhang. *Analysis of Variance for functional data.* CRC Press, 2014.

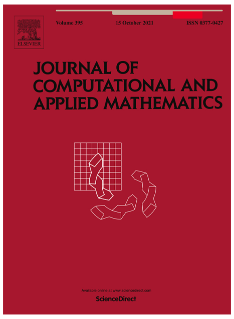
[189] J. T. Zhang, M. Y. Cheng, H. T. Wu, and B. Zhou. A new test for functional one-way anova with applications to ischemic heart screening. *Computational Statistics & Data Analysis*, 132:3–17, 2019.

[190] X. Zhao, J. S. Marron, and M. T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14(3):789–808, 2004.

Appendices

A1 Phase-type distributions for studying variability in resistve memories

- Acal, Christian; Ruiz-Castro, Juan Eloy; Aguilera, Ana M.; Jimenez-Molinos, Francisco; Roldan, Juan B. (2019)
- Phase-type distributions for studying variability in resistve memories
- *Journal of Computational and Applied Mathematics*, vol. 345, pp. 23-32
- DOI: `https://doi.org/10.1016/j.cam.2018.06.010`



Volume 395 15 October 2021 ISSN 0377-0427

ELSEVIER

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Available online at www.sciencedirect.com

ScienceDirect

Mathematics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	2.037	43/261	Q1

Abstract

A new statistical approach has been developed to analyze Resistive Random Access Memory (RRAM) variability. The stochastic nature of the physical processes behind the operation of resistive memories makes variability one of the key issues to solve from the industrial viewpoint of these new devices. The statistical features of variability have been usually studied making use of Weibull distribution. However, this probability distribution does not work correctly for some resistive memories, in particular for those based on the Ni/HfO2/Si structure that has been employed in this work. A completely new approach based on phase-type modelling is proposed in this paper to characterize the randomness of resistive memories operation. An in-depth comparison with experimental results shows that the fitted phase-type distribution works better than the Weibull distribution and also helps to understand the physics of the resistive memories

1 Introduction

In the context of applications for non-volatile memories, several emerging technologies are gaining momentum in the electronic industry. Among the new devices considered, both at industry and academia, RRAMs have shown an incomparable potential because they have good scalability, low power operation, fast speed and outstanding possibilities for fabrication in the current CMOS technology [1, 2, 3, 4, 5, 6].

The physics and internal properties of these devices have been studied by means of profound experimental studies [1, 7, 8, 9], and modelling and simulation studies [10, 11, 12, 13, 14, 15, 16]. A unified mathematical framework capable of describing and simulating the complex and interplaying electro-thermo-chemical processes that occur in this type of new emerging technologies in semiconductor device industry was recently developed [17]. Nevertheless, there are issues, such as variability, that have to be addressed prior to RRAM massive industrialization.

RRAMs operation is based on the stochastic nature of resistive switching (RS) processes that, in most cases, create (set process) and rupture (reset process) a conductive filament that changes drastically the device resistance [1, 5, 18]. There is a great need to analyse the statistics behind RRAM variability that is translated to different resistances, voltages and currents related to set and reset processes for each RS cycle (a cycle consist of a set process followed by a reset process) within a long series of cycles. Because of this, the choice of the right statistical model to describe the distribution of switching parameters (forming, set and reset voltages) is a critical requirement for RRAMs that ensures a robust design of the circuit and reliable data storage unit.

The usual statistical analysis performed on experimental data measured in these devices makes use of the Weibull distribution (WD) [1, 19, 20]; nevertheless, sometimes its fit with experimental data is not very accurate. In the next section, it will be shown that WD does not work correctly for the devices under consideration in this manuscript. In fact, recent dielectric breakdown studies have shown that the WD does not describe the stochastic trends well enough, more so in downscaled structures at the low and high percentile regions. The validity of a defect clustering model for RRAM switching parameters was recently examined [21].

Therefore, another statistical approach is needed. Apart from an accurate statistical description of experimental data, the interpretation of the parameters extracted by the application of the statistical analysis can shed new light to the variability issue and the physics behind RRAM operation.

In order to deepen on this issue, a thorough analysis of the statistical properties of RRAM variability is performed. To do so, phase-type distributions (PHDs) will help us to analyse the possible intermediate states of degradation in the conductive filament destruction processes that lead to a RRAM high resistivity state (the rupture of the conductive filament isolates the electrodes and therefore the RRAM resistance increases).

Phase-type distributions, which were introduced and analyzed in detail by Neuts [22, 23, 24], constitute a class of non-negative distributions that makes it possible to model complex problems with well-structured results, thanks to its matrix-algebraic form. Due to their valuable properties, many varieties of this class of distributions have been considered in diverse branches of science and engineering and applied in reliability studies. Particular cases of PHDs are the exponential, Erlang, generalized Erlang, hyper-geometric and Coxian distributions, among others. In fact, not only very well-known probability distributions are PHDs but also any nonnegative probability distribution can be approximated as needed taking into consideration that the PHD class is dense in the set of probability distributions on the nonnegative half-line [25]. The more essential and important properties of PHDs were reviewed in a recent study [26].

As reported below, the versatility and advantages of the PHD will come up to allow a better analysis of RRAM variability. In this context, it will be shown that the Erlang distribution (ED) (a particular PHD) works much better than WD to describe the experimental data under consideration in this manuscript. The physical interpretation of the fitted parameters from the PHD modelling will shed light on the explanation of RRAM variability.

The fabricated devices and measurement process are described in Section 2, the new statistical approach for modelling RRAMs variability is given in Section 3 and the main results and discussion in Section 4. Finally, the conclusions are given in Section 5.

2 Device description and measurement

The devices employed in this manuscript are unipolar Ni/HfO2/Si-based RRAMs. The fabrication details were given in [27]. A HP-4155B semiconductor parameter analyser was used in the measurement process that consisted of a long series of RS cycles under ramped voltage stress. The Si substrate (bottom electrode) was grounded and a negative voltage was applied to the Ni (top electrode), although for simplicity we have assumed the absolute value of the applied voltage henceforth [27].

Three reset current versus voltage curves of the series of resistive switching cycles (2749 cycles) measured are shown in Figure 1. In the curves plotted, a sudden current drop can be seen corresponding with the rupture of the conductive filament (highlighted as reset point [1, 5, 27, 14, 15]) that connects the electrodes (in this respect, the conductive filament works as a fuse). The corresponding voltages and currents are known as reset voltages and currents respectively [1, 18], they have been explicitly shown in Figure 1 for the sake of clarity. The reset voltage determination has been performed by detecting a 50% current variation at the reset point (when the sudden current drop takes place). Others methods have been proposed in the literature [18]; nevertheless, this one worked well for the devices under consideration here.

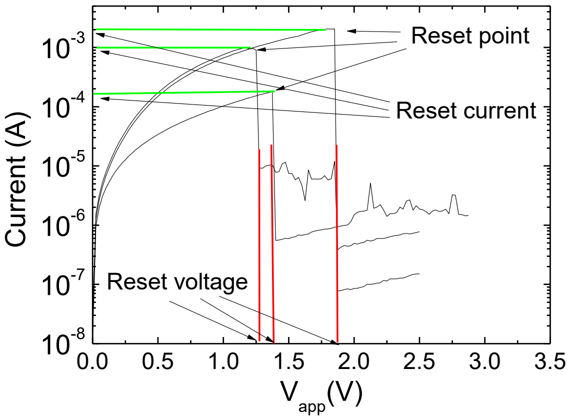


Figure 1. Experimental current versus applied voltage (shown in black lines) for three curves of a 2749 series of continuous resistive switching cycles, including set and reset curves. The reset point and the corresponding reset voltages (Vreset, indicated by vertical red lines) and reset currents (Ireset, indicated by horizontal green lines) are shown for clearness.

As noted in the introduction, the Vreset and Ireset distributions (as well as Vset and Iset distributions) have been subject of a deep statistical analysis for many different RRAMs [20, 28]. The WD was employed to describe the statistical properties of experimental data when the reset and set parameters were extracted. The Weibull model has been successfully employed along with a geometrical cell-based model which was connected with the percolation model for oxide breakdown for SiO2-based devices [29]. In addition, WD has been widely employed in the context of reliability physics and engineering [30]. Its use in the context of the statistical analysis of RRAM makes sense since it is a weakest-link type distribution, i.e., the failure of the whole is dominated by the degradation rate for the weakest element.

The cumulative distribution function for the WD is given in Equation 1 [8, 30]

$$F(v) = 1 - \exp\left(-\left(\frac{v}{\alpha}\right)^{\beta}\right) \tag{1}$$

For the devices reported above the statistical analysis based on the WD has been performed. On the one hand, the Vreset and Ireset for all the reset curves of the 2749 cycles under consideration in our RS series were computed [8, 31]. On the other hand, the typical Weibits, calculated as ln[-ln(1 – F)], have been obtained. If Weibits are plotted versus ln(Vreset), a linear plot should be obtained if experimental data follow a WD, where the slope corresponds to the β parameter in Equation 1 (β measures the statistical dispersion [8, 30]). The results obtained for our devices are shown in Figure 2.

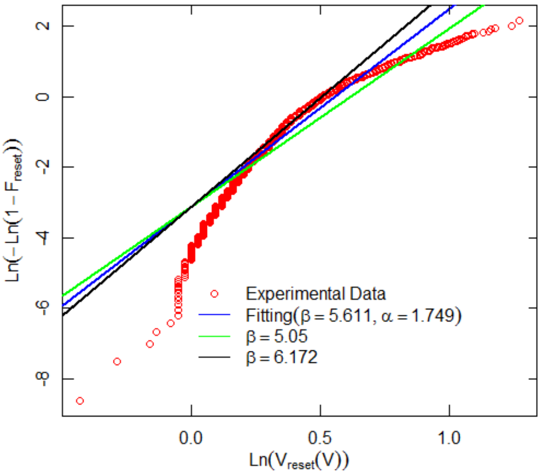


Figure 2. WD linear fit of Vreset for 2749 RS cycles. The best fit is obtained with the blue line, a 10% reduction (increase) in the beta parameter was assumed in the green (black) line.

Other analytical distributions (Equation 1) were also included making use of a β parameter with a 10% variation with respect to the best fit obtained in the statistical analysis. As can be seen, the Weibits of the experimental data are not linear. Therefore, although a rough approximation could be performed in the WD context, it seems reasonable to try other distributions. We do so in the following section and we call the reader's attention to the fact that a much better fit can be obtained with a Phase-type distribution modelling. After a progressive analysis based on step-by-step estimation of phase-type distributions [22, 23], it will be shown that the ED provides the best fit. Other previous statistical analysis on RRAMs might work much better making use of the ED; however, in this manuscript the study is limited to our experimental data.

3 Theory and methods

As shown in previous section, the logarithm of the experimental cumulative hazard rate versus $\ln(V_{reset})$ is not linear and therefore Weibull distribution seems not to be the appropriate distribution for fitting to V_{reset}. Then, the aim of this section is to find out what is the best probability distribution that describes the RRAM variability.

3.1 Phase-type distributions

One class of non-negative probability distributions with very interesting properties that allows to describe the main associated measures in an algorithmic form and to interpret the results is the phase-type distribution class. Phase type distributions were introduced and described in detail by [22] and [23].

The flexibility of the phase-type distribution makes it a good candidate to try a better fit since it generalizes a great number of distributions. The usefulness of this distribution class has been proved in several fields such as queueing theory, renewal processes, reliability and survival [32, 33, 34, 35, 36]. In our case, we can assume that for our devices the conductive filament within the RRAM dielectric pass through different degradation stages before the rupture process takes place (absorption, in the approach we are following). At this point it seems reasonable to figure out the evolution of the conductive filament, i.e., the different stages followed in the destruction process versus Vreset.

A phase type distribution (PHD) is defined as the distribution of the lifetime up to the absorption in an absorbing Markov process (voltage up to the conductive filament failure in the RRAM context).

In the context of RRAMs an absorbing Markov process to model the voltage to the failure of the conductive filament can be assumed. The state space is given by a general number of m transient degradation stages, where the probability of being initially in stage i is given by αi and one absorbing state, m+1, which is the conductive filament failure. In addition, the transition intensity from the transient stage *i* to the transient stage *j* is given by q_{ij} for *i≠j* and if *i=j* then $q_{ii} = -\sum_{\substack{j=1 \\ j\neq i}}^{m+1} q_{ij}$. The voltage up to failure is PHD distributed with representation $(\boldsymbol{\alpha}, \mathbf{T})$ being

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \quad ; \quad \mathbf{T} = (q_{ij})_{i,j=1,\ldots,m} \quad .$$

A PHD is a non-negative probability distribution whose cumulative distribution function is given in Equation 2

$$F(v) = 1 - \boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{e} \quad , v \geq 0, \tag{2}$$

where **e** is a column vector of ones with appropriate order. The density function associated to this distribution is

$$f(v) = -\boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{Te} = \boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{T}^0 \quad , v \geq 0,$$

with $\mathbf{T}^0 = -\mathbf{Te}$ being the transition intensity vector from a transient state up to one absorbing state.

It can be seen that if **α** is the scalar 1 and **T** is the scalar −λ, the exponential distribution is achieved. The reliability function, R(*v*), describes the probability that at voltage *v* the conductive filament is not broken, and it is given by Equation 3

$$R(v) = 1 - F(v) = \boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{e}, \, v \geq 0. \tag{3}$$

Thus, the cumulative hazard rate is given by Equation 4

$$H(v) = -\ln(1 - F(v)) = -\ln(\boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{e}), \tag{4}$$

and then the hazard rate is

$$h(v) = \frac{\partial H(v)}{\partial v} = \frac{f(v)}{1 - F(v)} = \frac{\boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{T}^0}{\boldsymbol{\alpha} \exp(\mathbf{T}v)\mathbf{e}}, \, v \geq 0.$$

3.2 Some PHD Properties

Phase-type distributions are important not only because of their structure but also for the good properties which enable to ease the applicability and interpretation of results.

Many well known distributions, in addition to the exponential distribution mentioned above, are PHD. Next, some of these are exposed with the corresponding PHD representation.

1. The Erlang distribution $F(v)=1-\sum_{j=0}^{m-1}e^{-\lambda v}(\lambda v)^{j}/j!$ for $v \geq 0$, $m \geq 1$ and $\lambda > 0$,

$$\boldsymbol{\alpha}=(1,\ldots,0,0),\ \mathbf{T}=\begin{pmatrix} -\lambda & \lambda & & \\ & -\lambda & \ddots & \\ & & \ddots & \lambda \\ & & & -\lambda \end{pmatrix}_{mxm}.$$

2. Hypo-exponential distribution $F(v)=1-\sum_{x=0}^{v}\sum_{i=1}^{m}\lambda_i e^{-\lambda_i x}\left(\prod_{\substack{j=1\\ j\neq i}}^{m}\frac{\lambda_j}{\lambda_j-\lambda_i}\right)$ for

   $v \geq 0$, $\lambda_i \neq \lambda_j$ for $i \neq j$,

$$\boldsymbol{\alpha}=(1,0,\ldots,0),\ \mathbf{T}=\begin{pmatrix} -\lambda_1 & \lambda_1 & & \\ & -\lambda_2 & \ddots & \\ & & \ddots & \lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}.$$

3. Hyper-exponential distribution $F(v)=1-\sum_{i=1}^{m}\alpha_i\left(1-e^{-\lambda_i v}\right)$ for $v \geq 0$,

$$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_m),\ \mathbf{T}=\begin{pmatrix} -\lambda_1 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_m \end{pmatrix}$$

4. Coxian distribution

$$\boldsymbol{\alpha}=(1,0,\ldots,0),\ \mathbf{T}=\begin{pmatrix} -\lambda_1 & g_1\lambda_1 & & \\ & -\lambda_2 & g_2\lambda_2 & \\ & & \ddots & g_{m-1}\lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}$$

5. Generalized Coxian distribution

$$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_m),\ \mathbf{T}=\begin{pmatrix} -\lambda_1 & g_1\lambda_1 & & \\ & -\lambda_2 & g_2\lambda_2 & \\ & & \ddots & g_{m-1}\lambda_{m-1} \\ & & & -\lambda_m \end{pmatrix}$$

In general, the following result [6] describes when a non-negative probability distribution is PHD. Then, a non-negative probability distribution is a phase-type distribution if and only if it is either the point mass at zero or; it has a strictly positive continuous density on the positive real numbers and it has a rational Laplace-Stieltjes transform with a unique pole of maximal real part.

One essential property that verifies the phase-type distribution class is that this class is dense in the set of probability distributions on the non-negative half-line [25]. In this way, PHDs can be considered general distributions with a well-structured matrix algorithmic form. On the other hand, any non-negative probability distribution can be approximated as much as desired by a phase-type-distribution.

Other properties for phase-type distributions are the closure properties. The PHD class is closed under a number of operations such as minimum, maximum and addition.

3.3 Estimating phase-type distributions: the EM algorithm

Fitting PHDs is a difficult optimization problem given that the representation of a PHD is highly redundant in general. One usual technique used to estimate the parameters of a PHD is the Expectation Maximization (EM) algorithm. The first EM algorithm was developed by [38] and assumed by [39].

Let $v_1,\dots,v_n$ be a sequence of n observed variable values. In our case, the value v_i is the voltage up to the absorption (rupture of the filament). The set $\{v_1,\dots,v_n\}$ includes the outcomes of n independent replications of a PHD with representation $\left(\boldsymbol{\alpha},\mathbf{T}\right)$ associated to an absorbing Markov process. The likelihood function to be optimized in the EM algorithm can be expressed as

$$L\left(\boldsymbol{\alpha},\mathbf{T}\right)=\prod_{i=1}^{m}\boldsymbol{\alpha}_i^{N_i}\prod_{i=1}^{m}e^{x_i\mathbf{T}_{ii}}\prod_{i=1}^{m}\prod_{\substack{j=1\\ j\neq i}}^{m+1}\mathbf{T}_{ij}^{N_{ij}}$$

where N_i is the number of times that the Markov process started in phase i, x_i is the total time spent in phase i and N_{ij} is the total observed number of jumps between both states, i and j.

If the current estimate of the PHD is $\left(\boldsymbol{\alpha},\mathbf{T}\right)$, then the conditional expectations of N_i, x_i and N_{ij} (E-step) are given by

$$E_{(\boldsymbol{\alpha},\mathbf{T})}\left[N_i\right]=\frac{1}{n}\sum_{k=1}^{n}\frac{\boldsymbol{\alpha}_i\left(e^{\mathbf{T}\nu_k}\mathbf{T}^0\right)_i}{\boldsymbol{\alpha}e^{\mathbf{T}\nu_k}\mathbf{T}^0},\quad E_{(\boldsymbol{\alpha},\mathbf{T})}\left[x_i\right]=\frac{1}{n}\sum_{k=1}^{n}\frac{\left[\int_0^{\nu_k}\left(\boldsymbol{\alpha}e^{\mathbf{T}(\nu_k-u)}\right)'\left(e^{\mathbf{T}u}\mathbf{T}^0\right)'du\right]_{ii}}{\boldsymbol{\alpha}e^{\mathbf{T}\nu_k}\mathbf{T}^0},$$

$$E_{(\boldsymbol{\alpha},\mathbf{T})}\left[N_{ij}\right]=\frac{1}{n}\sum_{k=1}^{n}\frac{\left[\int_0^{\nu_k}\left(\boldsymbol{\alpha}e^{\mathbf{T}(\nu_k-u)}\right)'\left(e^{\mathbf{T}u}\mathbf{T}^0\right)'du\right]_{ij}\mathbf{T}_{ij}}{\boldsymbol{\alpha}e^{\mathbf{T}\nu_k}\mathbf{T}^0},\quad E_{(\boldsymbol{\alpha},\mathbf{T})}\left[N_{i,m+1}\right]=\frac{1}{n}\sum_{k=1}^{n}\frac{\left(\boldsymbol{\alpha}e^{\mathbf{T}\nu_k}\right)_i\mathbf{T}_i^0}{\boldsymbol{\alpha}e^{\mathbf{T}\nu_k}\mathbf{T}^0}$$

Then, the M-step results in the estimation of new parameters

$$\hat{\boldsymbol{\alpha}}_i=E_{(\boldsymbol{\alpha},\mathbf{T})}\left[N_i\right];\hat{\mathbf{T}}_{ij}=\frac{E_{(\boldsymbol{\alpha},\mathbf{T})}\left[N_{ij}\right]}{E_{(\boldsymbol{\alpha},\mathbf{T})}\left[x_i\right]}\quad,\quad i\neq j;\qquad\hat{\mathbf{T}}_i^0=\frac{E_{(\boldsymbol{\alpha},\mathbf{T})}\left[N_{i,m+1}\right]}{E_{(\boldsymbol{\alpha},\mathbf{T})}\left[x_i\right]};\qquad\hat{\mathbf{T}}_{ii}=-\left(\hat{\mathbf{T}}_i^0+\sum_{\substack{j=1\\ j\neq i}}^{m}\hat{\mathbf{T}}_{ij}\right)$$

4 Results and discussion

To analyze the behavior of V$_{\text{reset}}$, the classical methodology based on the Weibull distributions has been used as it can be seen in Section 2. The results are not as good as desirable. Phase-type distributions with their corresponding properties have been introduced in the section above. One interesting property of PHD is that this class of distributions is dense in the non-negative probability distributions set. Thus, PHD are going to be assumed to estimate V$_{\text{reset}}$ distribution.

The voltage up to the conductive filament failure has been fitted by considering multiple general PHDs with different stages by using the EM algorithm [38] described in Subsection 3.3. The computations have been made by using the program EMpht for fitting phase-type distributions to data [40], in addition to developing own code with the software R and using the R project for Statistical Computing [41].

Thirty PHD with m transient stages, for m = 1,…,30, have been fitted to our data set by using the *EM*-algorithm. In total n=2749 V$_{\text{reset}}$ were observed. We have assumed any internal structure for matrix **T** (transition intensities), therefore for each one we have estimated $m(m–1)+m$ parameters. After this analysis, we have observed that the internal structure of the PHD representation depends only of one parameter, fixed m, for all cases. This structure can be expressed as follows

$$\boldsymbol{\alpha} = (1,0,\ldots,0) \text{ and } \mathbf{T} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ 0 & -\lambda & \lambda & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -\lambda & \lambda \\ 0 & \ldots & \ldots & 0 & -\lambda \end{pmatrix}. \tag{5}$$

It is known that a PHD with the structure described above is a Erlang distribution with representation $E(m, \lambda)$ as it can be seen in section 3.2. The cumulative distribution function of an Erlang distribution is given by

$$F(v) = 1 - \sum_{j=0}^{m-1} e^{-\lambda v} (\lambda v)^j / j!. \tag{6}$$

Thus, for the representation described in (5), both distributions: phase-type and Erlang, are equivalent, but now Erlang distribution is expressed in an algorithmic way through PH distributions. Therefore, functions (2) and (6) are the same.

Taking into account the previous results, we can conclude from the PHD analysis that the voltage up to the rupture process (V$_{\text{reset}}$) is Erlang distributed; however, the PHD structure is considered. The physical structure of the Erlang distribution and its parameters can be interpreted as follows: the conductive filament always begins in stage 1 (after a successful set process where its formation has been achieved) and it undergoes a sequential degradation thorough m distinct and well differenced stages (the number of stages is characterized by parameter m*)* where the mean V$_{\text{reset}}$ in each stage is equal to $1/\lambda$ (inverse of the second parameter of the Erlang distribution).

The Erlang distribution parameters have been estimated. Table 1 shows the λ estimates after applying the EM algorithm for different stages (whose number is described by parameter m).

Number of stages	Iterations EM algorithm	LogL	Estimate λ
15	1700	–1066.427	9.279325
14	1700	–1096.136	8.660703
13	1300	–1133.043	8.042082
12	1000	–1178.317	7.423459
11	1000	–1233.436	6.804838
10	800	–1300.308	6.186217
9	600	–1381.457	5.567595
8	600	–1480.314	4.948974
7	600	–1601.724	4.330352
6	500	–1752.833	3.711730
5	200	–1944.833	3.093108
4	200	–2196.729	2.474486
3	200	–2544.869	1.855864

Table 1. Parameter λ estimated by using the EM algorithm depending on the number of stages.

The optimum value is reached for 15 stages with $\hat{\lambda} = 9.279325$. Therefore, the estimated mean V_{reset} in each stage is equal to 0.1078. Finally, the mean estimated V_{reset} from the beginning up to the conductive filament failure is 1.6165.

The experimental cumulative hazard rate estimated by the Erlang and Weibull distributions have been plotted and compared graphically (see Figure 3).

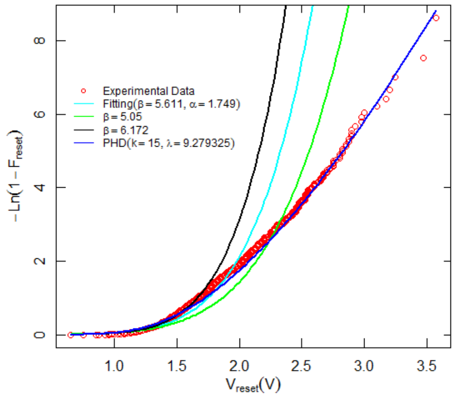


Figure 3. Cumulative hazard rate of V_{reset} for 2749 RS cycles and the corresponding Weibull and PH distribution fit.

The best result is achieved when the Erlang distribution is considered and the accuracy of the fit is remarkable, as shown in Figure 3.

Once the statistical analysis has been developed, a detailed study of the devices considered here is developed. To do so, an analysis based on the data screened by means of the Low Resistance State

(LRS) of the device resistance, R, is performed. The device LRS resistance is measured just after a set process is over, when the conductive filament is fully formed. Usually, the resistance at this point is the lowest value found all along a complete resistive switching cycle.

If V_{reset} values are sorted out by considering the LRS resistance, a better fit is obtained by means of WD, although the fitting is not accurate, as can be seen in Figure 4a. In particular, for R<20 kΩ the fitting is not very good. Nevertheless, if a PHD is employed instead of a WD a reasonable accuracy is achieved. In this respect, the PHD appropriateness to deal with our experimental data is noteworthy at the sight of Figure 4b, and this is for all the resistance range under consideration.

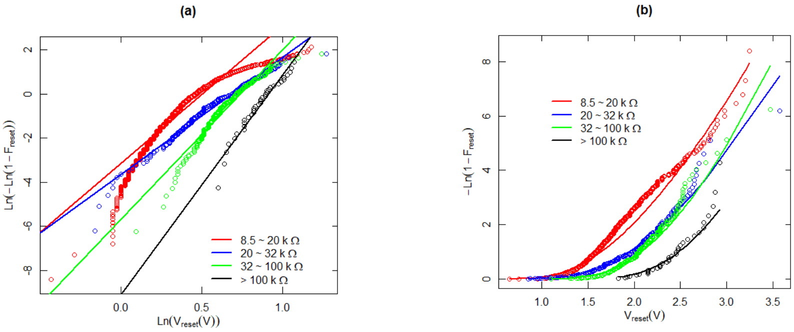


Figure 4. a) WD Weibits versus Ln(V_{reset}) for the experimental data under consideration screened for different LRS resistances are plotted in symbols. The analytical WD best fit is also shown in solid lines, b) hazard rate for the screened experimental data (symbols) and PHD (solid lines).

The reliability function, i.e., the survival function, as it is known in scientific branches not related to engineering, is interesting to analyze the statistical properties of the data we are dealing with. Since V_{reset} can be considered as the failure voltage for the memories under study (RRAMs), the reliability functions portrays the probability that a memory state change will not be produced; i.e., the conductive filament will not be broken and the memory resistance state will not be switched. From another viewpoint, the reliability function describes the probability that the conductive filament will not be ruptured for voltages below the failure voltage.

The reliability function has been plotted in Figure 5 for the experimental data, WDs and PHDs. Although no distribution shows a close reproduction of the experimental values for low voltages, at medium-high voltages the PHD works better than WD and achieves a reasonably good performance.

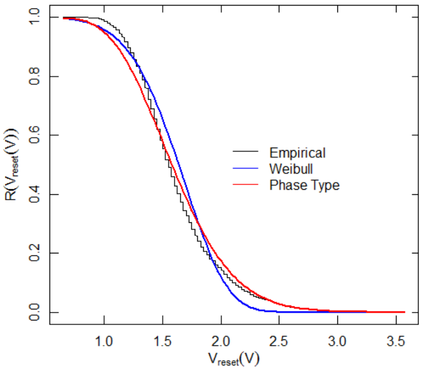


Figure 5. Reliability function versus V_{reset} for experimental data and Weibull and Phase-type distributions.

In order to further characterize the correctness of our approach, the hazard rate function should also be considered since it describes the failure rate in a voltage interval (V_{reset},V_{reset}+dV_{reset}). It could also be interpreted as the device degradation velocity at a certain voltage. This function has been plotted in Figure 6. Again, as can be seen and it was expected from previous results, PHD works better than WD.

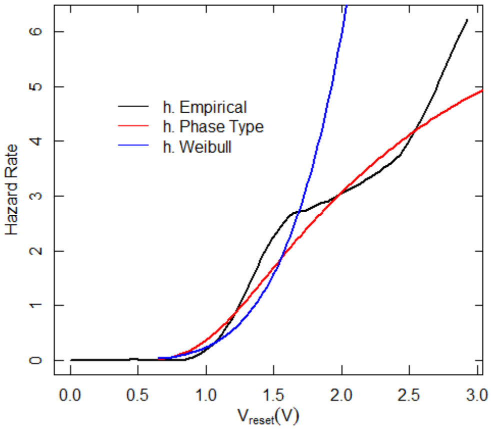


Figure 6. Hazard rate versus V_{reset} for experimental data, phase-type and Weibull distributions.

5 Conclusions

The usual statistical analysis performed on RRAM experimental data in order to characterize the device variability makes use of Weibull distribution. Nevertheless, sometimes the fit obtained to measured data is not accurate. This fact suggests that other statistical distributions could work in a better manner. In this respect, a new methodology is developed in our manuscript by considering phase-type distributions to fit the V_{reset} distribution, the voltage corresponding to the reset processes where RRAM conductive filaments rupture happens. The phase-type distribution class employed can be considered a general class, given that any non-negative distribution can be approximated as needed through a phase-type distribution. From the general phase-type distribution parameter estimation performed on RRAM experimental measurements, the best fit is obtained and it was found that Erlang distribution, a distribution belonging to phase-type distribution class, is particularly appropriate. The phase-type parameters were estimated from experimental data and interpreted from the physically based viewpoint. The first parameter is the number of sequential degradation stages up to the reset and the inverse of the second parameter is the mean Vreset in each stage. In addition, the fit is compared with the usual Weibull distribution to shed light on the issue.

References

[1] F. Pan, S. Gao, C. Chen, C. Song, F. Zeng, Recent progress in resistive random access memories: materials, switching mechanisms and performance, Materials Science and Engineering 83 (2014) 1-59.

[2] D. Ielmini, R. Waser. Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications, Wiley-VCH, 2015.

[3] R. Waser, M. Aono, Nanoionics-based resistive switching, Nature materials 6 (2007) 833–840.

[4] M. Lanza, G. Bersuker, M. Porti, E. Miranda, M. Nafría, X. Aymerich, Resistive switching in hafnium dioxide layers: Local phenomenon at grain boundaries, Applied Physics Letters 101 (2012) 193502.

[5] M. Lanza, A Review on Resistive Switching in High-k Dielectrics: A Nanoscale point of View Using Conductive Atomic Force Microscope, Materials 7 (2014) 2155-2182.

[6] R. Waser, Nanoelectronics and Information Technology, Third ed., Wiley, 2012.

[7] S. Long, C. Cagli, D. Ielmini, M. Liu, J. Suñé, Reset statistics of NiO-based resistive switching memories, IEEE Electron Device Lett. 32*(11)* (2011) 1570–1572.

[8] S. Long, X. Lian, T. Ye, C. Cagli, L. Perniola, E. Miranda, M. Liu, J. Suñé, Cycle-to-cycle intrinsic RESET statistics in HfO_2-based unipolar RRAM devices, IEEE Electron Device Lett. 34*(5)* (2013) 623–625.

[9] T. Tsuruoka, K. Terabe, T. Hasegawa, M. Aono, Forming and switching mechanisms of a cation-migration-based oxide resistive memory, Nanotechnology 21*(42)* (2010).

[10] A. Padovani, Member, IEEE, L. Larcher, Member, IEEE, O. Pirrotta, L. Vandelli, G. Bersuker, Member, IEEE, Microscopic Modeling of HfO x RRAM Operations: From Forming to Switching, IEEE Transactions on Electron Devices 62*(6)* (2015) 1998-2006.

[11] S. Aldana, P. García-Fernández, A. Rodríguez-Fernández, R. Romero-Zaliz, M.B. González, F. Jiménez-Molinos, F. Campabadal, F. Gómez-Campos, J.B. Roldán, A 3D Kinetic Monte Carlo

simulation study of Resistive Switching processes in Ni/HfO2/Si-n+-based RRAMs, Journal of Physics D: Applied Physics, 50 (2017) 33510.

[12] J. Guy, G. Molas, P. Blaise, M. Bernard, A. Roule, G. Le Carval, V. Delaye, A. Toffoli, G. Ghibaudo, *Fellow, IEEE*, F. Clermidy, B. De Salvo, L. Perniola , Investigation of Forming, SET, and Data Retention of Conductive-Bridge Random-Access Memory for Stack Optimization, IEEE Transactions on Electron Devices 62*(11)* (2015) 3482-3489.

[13] F. Jiménez-Molinos, M.A. Villena, J.B. Roldán, A.M. Roldán, A SPICE Compact Model for Unipolar RRAM Reset Process Analysis, IEEE Transactions on Electron Devices 62 (2015) 955-962.

[14] M.A. Villena, M.B. González, J.B. Roldán, F. Campabadal, F. Jiménez-Molinos, F.M. Gómez-Campos, J. Suñé, An in-depth study of thermal effects in reset transitions in HfO2 based RRAMs, Solid State Electronics 111 (2015) 47-51.

[15] M.A. Villena, J.B. Roldán, M.B. González, P. González-Rodelas, F. Jiménez-Molinos, F. Campabadal, D. Barrera, A new parameter to characterize the charge transport regime in Ni/HfO_2/Si-n^+-based RRAMs, Solid State Electronics 118 (2016) 56-60.

[16] J.B. Roldán, E. Miranda, G. González-Cordero, P. García-Fernández, R. Romero-Zaliz, P. González-Rodelas, A. M. Aguilera, M.B. González, F. Jiménez-Molinos, Multivariate analysis and extraction of parameters in resistive RAMs using the Quantum Point Contact model, Journal of Applied Physics 123 (2018) 014501.

[17] A. Mauri, R. Sacco, M. Verri, Electro-thermo-chemical computational models for 3D heterogeneous semiconductor device simulation, Applied Mathematical Modelling 39 (2015) 4057–4074.

[18] M.A. Villena, J.B. Roldán, F. Jiménez-Molinos, E. Miranda, J. Suñé, M. Lanza, *SIM²RRAM*: A physical model for RRAM devices simulation, Journal of Computational Electronics 16*(4)* (2017) 1095-1120.

[19] G. González-Cordero, J.B. Roldan, F. Jiménez-Molinos, J. Suñé, S. Long, M. Liu, A new model for bipolar RRAMs based on truncated cone conductive filaments, a Verilog-A approach, Semiconductor Science and Technology 31 (2016) 115013.

[20] W.C, Luo, J.C. Liu, H.T. Feng, Y.C. Lin, J.J. Huang, K.L. Lin, T.H. Hou, RRAM SET speed-disturb dilemma and rapid statistical perdiction methodololy, in: Proceedings of International Eelectron Device Meeting, 2012, 9.5.1-9.5.4.

[21] N. Raghavan, Application of the defect clustering model for forming, SET and RESETstatistics in RRAM devices, Microelectronics Reliability 64 (2016) 54–58.

[22] M.F. Neuts, Probability distributions of phase type, in: Liber Amicorum Professor Emeritus H. Florin, Belgium: University of Louvain, 1975, pp. 173–206.

[23] M.F. Neuts, Matrix geometric solutions in stochastic models. An algorithmic approach, in Probability Distributions of Phase Type, Baltimore: John Hopkins University Press, 1981.

[24] M.F. Neuts, Phase-type probability distributions, in S. L. Gass and M. C. Fu eds, Encyclopedia of Operations Research and Management Sciences, 2013.

[25] S. Asmussen, Ruin probabilities, World Scientific, Hong Kong, 2000.

[26] Q.M. He, Fundamentals of Matrix-Analytic Methods, Springer Science+Business Media, New York, 2014.

[27] M. B. González, J. M. Rafí, O. Beldarrain, M. Zabala, F. Campabadal, Analysis of the switching variability in Ni/HfO2-based RRAM devices, IEEE Trans. Device Mater. Reliab. 14*(2)* (2014) 769–771.

[28] S. Long, C. Cagli, D. Ielmini, M. Liu, J. Suñé, Analysis and modeling of resistive switching statistics, J. Appl. Phys. 111 (2012) 074508.

[29] J. Suñé, New physics-based analytic approach to the thin-oxide breakdown statistics, IEEE Electron Device Lett. 22*(6)* (2001) 296-298.

[30] J. W. McPherson, Reliability Physics and Engineering. Time-to-Failure Modeling, second ed., Springer, 2013.

[31] M.A. Villena, J.B. Roldán, F. Jiménez-Molinos, J. Suñé, S. Long, E. Miranda, M. Liu, A comprehensive analysis on progressive reset transitions in RRAMs, Journal of Physics D: applied physics 7 (2014) 205102.

[32] M.C. Segovia, P.E. Labeau, Reliability of a multi-state system subject to shocks using phase-type distributions, Applied Mathematical Modelling 37 (2013) 4883–4904.

[33] M. Yu, Tang, W. Wu, J. Zhou, Optimal order-replacement policy for a phase-type geometric process model with extreme shocks, Applied Mathematical Modelling 38 (2014) 4323–4332.

[34] [33] J.E. Ruiz-Castro, Complex multi-state systems modelled through Marked Markovian Arrival Processes, European Journal of Operational Research 252*(3)* (2016) 852-865.

[35] J.E. Ruiz-Castro, Markov counting and reward processes for analyzing the performance of a complex system subject to random inspections, Reliability Engineering and System Safety 145 (2016) 155-168.

[36] S. R. Chakravarthy, A catastrophic queueing model with delayed action, Applied Mathematical Modelling 46 (2017) 631–649.

[37] O'Cinneide, Characterization of phase-type distributions, Stochastic Models 6 (1990) 1-57.

[38] S. Asmussen, O. Nerman, M. Olsson, Fitting Phase-Type distributions via the EM algorithm, Scandinavian Journal of Statistics 23*(4)* (1996) 419-441.

[39] P. Buchholz, J. Kriege, I. Felko, Input modeling with phase-type distributions and Markov models, Theory and Applications, Springer Cham Heidelberg New York Dordrecht London, 2014.

[40] The R-project for Statistical Computing. https://www.r-project.org, 2018 (accessed 28 February 20018).

[41] The EMpht program. http://home.math.au.dk/asmus/pspapers.html, 2018 (accessed 28 February 20018).

A2 A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories

- Ruiz-Castro, Juan Eloy; Acal, Christian; Aguilera, Ana M.; Roldan, Juan B. (2021)
- A Complex Model via Phase-Type Distributions to Study Random Telegraph Noise in Resistive Memories
- *Mathematics*, vol. 9, num. 4, pp. 390
- DOI: `https://doi.org/10.3390/math9040390`

Mathematics			
JCR Year	Impact Factor	Rank	Quartile
2019	1.747	28/235	Q1

Abstract

A new stochastic process is developed by considering the internal performance of macro-states where the sojourn time in each one is phase-type distributed depending on time. The stationary distribution is calculated through matrix-algorithmic methods and multiple interesting measures are worked out. The number of visits distribution to a determine macro-state is analyzed from the respective differential equations and Laplace transform. The mean number of visits to a macro-state between any two times is given. The results have been implemented computationally and they have been successfully applied to study the Random Telegraph Noise (RTN) in resistive memories. RTN is an important concern in Resistive Random Access Memory (RRAM) operation. On the one hand, it could limit some of the technological applications of these devices and on the other hand, RTN can be used for the physical characterization. Therefore, an in-depth statistical analysis to model the behavior of these devices is of essential importance.

1 Introduction

In several fields, such as computing and electronics engineering, is of great interest to analyze complex devices with several macro-states that evolve by time. It is usual to consider Markov processes for this analysis but in multiple occasions the spent times in each macro-state are not exponentially distributed. In this context, a new approach is to consider that the spent time in each macro-state is phase-type (PH) distributed. For this case, it is assumed that each macro-state is composed of internal performance states which have a Markovian behavior. One interesting aspect is that the modeling of the stochastic process when only the macro-state process can be observed. For this new process the Markovianity is lost.

A phase-type distributions is defined as the distribution of the absorption time in an absorbing Markov chain. These probability distributions constitute a class of distributions on the positive real axis which seems to strike a balance between generality and tractability thanks to its good properties. This class of distributions was introduced by Neuts [1,2] and allows to model complex problems in an algorithmic and computational way. It has been widely applied in fields such as engineering to model complex systems [3-5]; queueing theory [6]; risk theory [7] and electronics engineering [8].

The good properties of PH distributions through appealing probabilistic arguments constitute their main feature of being mathematically tractable. Several well-known probability distributions; e.g. exponential, Erlang, Erlang generalized, hyper-geometric and Coxian distributions, among others, are particular cases of PH. One of the most important PH properties is that whatever nonnegative probability distribution can be approximated as much as desired through a PH, accounting for the fact that the PH class is dense in the set of probability distributions on the nonnegative half-line [9]. It allows considering general distributions through PH approximations. Exact solutions to many complex problems in stochastic modeling can be obtained either explicitly or numerically by using matrix-analytic methods. The main features of PH distributions are described in depth in Ref. [10].

A macro-state stochastic process is built here to model the behavior of different Random Telegraph Noise (RTN) signals. Given that the behavior of the different macro-states (levels) is not exponentially distributed, we assume each level to be composed of multiple internal phases which can be related with other levels. We show that if the internal process, by considering internal states inside the macro-states, has a Markovian behavior, the process between levels is not. For the latter, the sojourn time in each level is phase-type distributed depending on the initial time. This fact is of great interest and gives us more information about the device internal process.

Advanced statistical techniques are key tools to model complex physical and engineering problems in many different areas of expertise. In this context, a new statistical methodology is presented; we will concentrate in the study of resistive memories [11-12], also known as Resistive Random Access Memories (RRAMs), a subgroup of a wide class of electron devices named memristors [13]. They constitute a promising technology with great potential in several applications in the integrated circuit realm.

Circuit design is a tough task because of the high number of electronics components included in integrated circuits, therefore, Electronic Design Automation (EDA) tools are essential. These software tools need compact models that represent the physics of the devices in order to fulfil their role in aiding circuit design. Although there are many works in compact modeling in the RRAM literature [14, 15] there is a lot to be done, in particular in certain areas related to variability and noise. Variability is essential in modeling due to the RRAM operation inherent stochasticity, since the physics is linked to random physical mechanisms [16, 17]. Another issue of great importance in these devices is noise. Among the different types of noise, Random Telegraph Noise is a great concern [18]. The disturbances produced by one or several traps (physical active defects within the dielectric) inside the conductive filament or close to it alter charge transport mechanisms and consequently current fluctuations can show up (see Figure 1) that lead to RTN [14, 15, 19, 20]. This noise can affect the correct device operation in applications linked to memory cell arrays and neuromorphic hardware [14, 16, 21], posing important hurdles to the use of this technology in highly-scaled integrated circuits. Nevertheless, RTN fluctuations can also be beneficial, for instance, when used as entropy sources in random number generators, an employment of most interest in cryptography [22, 23].

The application in this issue is dealt in the current work, in particular with the statistical description of RTN signals in RRAMs. This study, in addition to physically characterize the devices, can be used for compact modeling and, therefore, as explained above, for developing the software tools needed for circuit design. Accounting for the aforementioned intrinsic stochasticity of these devices, the choice of a correct statistical strategy to attack this problem is essential in the analysis. In this respect, we use the PH distributions, which have already been employed to depict some facets of RRAM variability [24]; nonetheless, as far as we know, they have not been used in RTN analysis.

In our study we deal with devices whose structure is based on the Ni/HfO$_2$/Si stack. The fabrication details as well as their electric characterization are given in Ref. [25]. RTN measurements were recorded in the time domain by means of a HP-4155B semiconductor parameter analyzer (see Figure 1). The current fluctuation between levels (see the marked levels in red in Figure 1), ΔI, and the spent time in each of the levels are key parameters to analyze. The associated time with the current low level is known as emission time, τ_e, and is linked to the time the active defect keeps a "captured" electron trapped till it releases it (through this time it is said that the defect is occupied). On the other hand, the capture time, τ_c, represents the time taken to capture an electron by an empty defect.

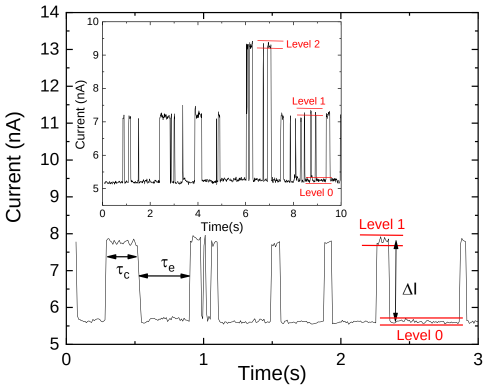


Figure 1: Current versus time trace for a Ni/HfO$_2$/Si device in the HRS, the RTN noise can be clearly seen. A two level signal is shown, the two different current levels are marked with the corresponding current thresholds. Another RTN trace is shown in the inset, measured for the same device. Three current levels are seen in this case.

The manuscript is organized as follows: in Section 2 the proposed statistical procedure is described, in Section 3 we deepen on the measures associated, and the number of visits to a determined macro-state is presented in Section 4. The parameter estimation is unfolded in Section 5 and the application of the methodology developed here in Section 6. Finally, conclusions related with the main contributions of this work are presented in Section 7.

2 Statistical Methodology

Different RTN signals have been employed here such as the ones shown in Figure 1. We have determined the number of current levels in order to be able to calculate the emission and capture times for a certain time interval. The current thresholds (red lines in Figure 1) were chosen to let the extraction algorithm to calculate the time intervals corresponding to the levels defined previously.

In order to explain these issues, a new model is firstly built in transient and stationary regimes.

2.1 The Model

We assume a stochastic process $\{X(t);\ t \geq 0\}$ with macro-state space $E = \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{r}\}$. Each macro-state $\mathbf{k}$, level $\mathbf{k}$, is composed of n_k internal phases or states denoted as i_h^k for $h = 1,\ldots, n_k$. We assume that the device internal performance is governed by a Markov process, $\{J(t);\ t \geq 0\}$, with state space $E = \left\{i_1^1,\ldots,i_{n_1}^1, i_1^2,\ldots,i_{n_2}^2,\ldots,i_1^r,\ldots,i_{n_r}^r\right\}$, and with an initial distribution $\boldsymbol{\theta}$, and the following generator matrix expressed by blocks,

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1k} & \cdots & \mathbf{Q}_{1r} \\ & \ddots & & & \\ \vdots & & \mathbf{Q}_{kk} & & \vdots \\ & & & \ddots & \\ \mathbf{Q}_{r1} & \cdots & \mathbf{Q}_{rk} & \cdots & \mathbf{Q}_{rr} \end{pmatrix},$$

where the matrix block $\mathbf{Q}_{ij}$ contains the transition intensities between the states of the macro-states from **i** to **j**.

Throughout this work we will denote as $\mathbf{Q}_k$ to the matrix with zeros except the column matrix block k of **Q**, and $\mathbf{Q}_{-k}$ the matrix **Q** with zeros in the *k-th* block column respectively. These matrices give the transition intensities to macro-state **k** and to any macro-state different to **k,** respectively.

Thus,

$$\mathbf{Q}_k = \begin{pmatrix} \mathbf{0} & \dots & \mathbf{0} & \mathbf{Q}_{1k} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{Q}_{rk} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix}, \qquad \mathbf{Q}_{-k} = \begin{pmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1,k-1} & \mathbf{0} & \mathbf{Q}_{1,k+1} & \cdots & \mathbf{Q}_{1r} \\ \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ \mathbf{Q}_{r1} & \cdots & \mathbf{Q}_{r,k-1} & \mathbf{0} & \mathbf{Q}_{r,k+1} & \cdots & \mathbf{Q}_{rr} \end{pmatrix}.$$

Clearly, it can be seen that the following equality holds, $\mathbf{Q} = \mathbf{Q}_k + \mathbf{Q}_{-k}$.

Given the Q-matrix of the Markov process, the transition probability matrix for the process $\{J(t);\ t \geq 0\}$ is given by $\mathbf{P}(t)=\exp\{\mathbf{Q}t\}$. From this and from the initial distribution $\boldsymbol{\theta}$, the transient distribution at time t is given by $(\mathrm{P}\{X(t)=i_1^1\}, \mathrm{P}\{X(t)=i_2^1\},\dots, \mathrm{P}\{X(t)=i_{n_r}^r\}) = \mathbf{a}(t) = \boldsymbol{\theta}\cdot\mathrm{P}(t)$. The order of this vector is $1 \text{ x } \sum_{i=1}^{r} n_i$.

The transition probabilities for the process $\{X(t);\ t \geq 0\}$ can be obtained from the transition probabilities of the Markov process $\{J(t);\ t \geq 0\}$. If it is denoted as $\mathbf{H}(\cdot,\cdot)$ where the transition between macro-states **i**→**j** is given by

$$h_{\mathbf{ij}}(s,t) = P\{X(t)=\mathbf{j} \mid X(s)=\mathbf{i}\} = \frac{\sum_{h\in\mathbf{j}}\sum_{k\in\mathbf{i}} P\{J(t)=h \mid J(s)=k\} P\{J(s)=k\}}{\sum_{k\in\mathbf{i}} P\{J(s)=k\}}$$

$$= \frac{\sum_{k\in\mathbf{i}} \left[a_k(s) \sum_{h\in\mathbf{j}} p_{kh}(t-s) \right]}{\sum_{k\in\mathbf{i}} a_k(s)}.$$

This transition probability can be expressed in a matrix-algorithmic form as follows. Denoting by $\mathbf{A}_\mathbf{k}$ a matrix of zeros with order $\sum_{i=1}^{r} n_i$ x $\sum_{i=1}^{r} n_i$ except the block corresponding to the macro-state **k**, which is the identity matrix, then,

$$h_{\mathbf{ij}}(s,t) = P\{X(t)=\mathbf{j} \mid X(s)=\mathbf{i}\} = \frac{\mathbf{a}(s)\mathbf{A}_\mathbf{i}\mathbf{P}(t-s)\mathbf{A}_\mathbf{j}\cdot\mathbf{e}}{\mathbf{a}(s)\mathbf{A}_\mathbf{i}\cdot\mathbf{e}},$$

where $\mathbf{a}(t)\mathbf{A}_\mathbf{i}\mathbf{e}$ is the probability of being in the macro-state **i** at time t with **e** being a column vector of ones with appropriate order. A similar reasoning can be done with the embedded jumping probability.

Note that the matrix **H** is a stochastic matrix, and therefore it is a transition matrix of a Markov process. However, an important remark is that the Markov process related to the matrix **H** does not correspond to the process defined in this section. In fact, this process is non-homogeneous and not Markovian.

2.2 Stationary distribution

The stationary distribution for the process $\{X(t);\ t \geq 0\}$ can be calculated by blocks from the stationary distribution for the process $\{J(t);\ t \geq 0\}$. It is well-known that this distribution verifies the balance equations $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ and the normalization equation $\boldsymbol{\pi}\mathbf{e} = 1$. We denote as $\boldsymbol{\pi}_k$ to the block of $\boldsymbol{\pi}$ corresponding to the macro-state **k**. It is denoted as $\boldsymbol{\pi}_k\,\mathbf{e}$ (**k** = 1,..., **r**) to the stationary distribution for the macro-state **k** of the process $\{X(t);\ t \geq 0\}$.

The stationary distribution has been worked out by blocks in order to reduce the computational cost, according to the macro-states, applying matrix-analytic methods.

As it has been said above, the stationary distribution verifies the following system:

$$\begin{aligned}
&\boldsymbol{\pi}_1\mathbf{Q}_{11}+\boldsymbol{\pi}_2\mathbf{Q}_{21}+\boldsymbol{\pi}_3\mathbf{Q}_{31}+\cdots+\boldsymbol{\pi}_r\mathbf{Q}_{r1}=\mathbf{0}\\
&\boldsymbol{\pi}_1\mathbf{Q}_{12}+\boldsymbol{\pi}_2\mathbf{Q}_{22}+\boldsymbol{\pi}_3\mathbf{Q}_{32}+\cdots+\boldsymbol{\pi}_r\mathbf{Q}_{r2}=\mathbf{0}\\
&\boldsymbol{\pi}_1\mathbf{Q}_{13}+\boldsymbol{\pi}_2\mathbf{Q}_{23}+\boldsymbol{\pi}_3\mathbf{Q}_{33}+\cdots+\boldsymbol{\pi}_r\mathbf{Q}_{r3}=\mathbf{0}\\
&\cdots\\
&\boldsymbol{\pi}_1\mathbf{Q}_{1r}+\boldsymbol{\pi}_2\mathbf{Q}_{2r}+\boldsymbol{\pi}_3\mathbf{Q}_{3r}+\cdots+\boldsymbol{\pi}_r\mathbf{Q}_{rr}=\mathbf{0}\\
&\boldsymbol{\pi}\cdot\mathbf{e}=1.
\end{aligned}$$

It has been solved by using matrix-analytics methods and the solution is given by

$\boldsymbol{\pi}_j=-\sum_{i=1}^{j-1}\boldsymbol{\pi}_i\mathbf{R}_{ij}^{r-j+1}$; $j = 2,\dots, r$,

where $\mathbf{R}_{ij}^{1}=\mathbf{Q}_{ij}\mathbf{Q}_{jj}^{-1}$ for any $i, j = 1,\dots, r$ with $i \neq j$,

$\mathbf{R}_{ij}^{r-j+1}=\left(\mathbf{R}_{ij}^{1}+\mathbf{H}_{ij}^{r-j+1}\right)\left(\mathbf{I}+\mathbf{H}_{jj}^{r-j+1}\right)^{-1}$ for $1 \leq i < j < r$, being

$$\begin{aligned}
\mathbf{H}_{ij}^{r-j+1}=&-\sum_{j<u_1\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1j}^{1}+\sum_{j<u_1<u_2\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\mathbf{R}_{u_2j}^{1}\\
&-\sum_{j<u_1<u_2<u_3\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\mathbf{R}_{u_2u_3}^{r-u_3+1}\mathbf{R}_{u_3j}^{1}\\
&+\sum_{j<u_1<u_2<u_3<u_4\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\mathbf{R}_{u_2u_3}^{r-u_3+1}\mathbf{R}_{u_3u_4}^{r-u_4+1}\mathbf{R}_{u_4j}^{1}-\dots\\
&\pm\sum_{j<u_1<u_2<\cdots<u_{r-j}\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\cdots\mathbf{R}_{u_{r-3}u_{r-2}}^{r-u_{r-2}+1}\mathbf{R}_{u_{r-2}j}^{1}\\
&\mp I_{\{j<r-1\text{ or } i\neq j\}}\mathbf{R}_{1,j+1}^{r-j}\mathbf{R}_{j+1,j+2}^{r-j}\cdots\mathbf{R}_{r-1,r}^{1}\mathbf{R}_{rj}^{1}.
\end{aligned}$$

for $1 \leq i \leq j < r$.

The vector $\boldsymbol{\pi}_1$ is worked out as follows:

$$\boldsymbol{\pi}_1=(1,\mathbf{0})\left[\mathbf{B}\middle|\left(\mathbf{I}+\mathbf{H}_{11}^{r}\right)^{*}\right]^{-1},$$

where given a matrix $\mathbf{A}$, $\mathbf{A}^{*}$ is the matrix $\mathbf{A}$ without the last column and

$$\begin{aligned}
\mathbf{B}=\mathbf{e}&-\sum_{j<u_1\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{e}+\sum_{j<u_1<u_2\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\mathbf{e}-\sum_{j<u_1<u_2<u_3\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\mathbf{R}_{u_2u_3}^{r-u_3+1}\mathbf{e}\\
&+\sum_{j<u_1<u_2<u_3<u_4\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\mathbf{R}_{u_2u_3}^{r-u_3+1}\mathbf{R}_{u_3u_4}^{r-u_4+1}\mathbf{e}-\dots\\
&\pm\sum_{j<u_1<u_2<\cdots<u_{r-j}\leq r}\mathbf{R}_{iu_1}^{r-u_1+1}\mathbf{R}_{u_1u_2}^{r-u_2+1}\cdots\mathbf{R}_{u_{r-3}u_{r-2}}^{r-u_{r-2}+1}\mathbf{e}.
\end{aligned}$$

Finally, the stationary distribution of the process $\{X(t);\ t \geq 0\}$ is given by $\left(\boldsymbol{\pi}_1\cdot\mathbf{e},\boldsymbol{\pi}_2\cdot\mathbf{e},\cdots,\boldsymbol{\pi}_r\cdot\mathbf{e}\right)$.

Algorithm to calculate the stationary distribution

```
Step 1. For $i$, $j$ =1,..., $r$ and $i \neq j$
    Compute $\mathbf{R}_{ij}^{1}$
Step 2. For $j$ = $r$–1,..., 2 {
        For $i$ = 1,..., $j$ {
            Compute $\mathbf{H}_{ij}^{r-j+1}$ }
            For $i = 1,\ldots, j$ {
        Compute $\mathbf{R}_{ij}^{r-j+1}$ $(i \neq j)$}}
Step 3. Compute $\mathbf{H}_{11}^{r}$ and B
Step 4. Compute $\boldsymbol{\pi}_1$
Step 5. Compute $\boldsymbol{\pi}_2,\ldots,\boldsymbol{\pi}_r$ and $\boldsymbol{\pi}_1 \cdot \mathbf{e},\ldots,\boldsymbol{\pi}_r \cdot \mathbf{e}$
```

Example. Case r = 4

Step 1.

$$\mathbf{R}_{12}^{1} = \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \;\; ; \;\; \mathbf{R}_{13}^{1} = \mathbf{Q}_{13}\mathbf{Q}_{33}^{-1} \;\; ; \;\; \mathbf{R}_{14}^{1} = \mathbf{Q}_{14}\mathbf{Q}_{44}^{-1} \;\; ;$$

$$\mathbf{R}_{21}^{1} = \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} \;\; ; \;\; \mathbf{R}_{23}^{1} = \mathbf{Q}_{23}\mathbf{Q}_{33}^{-1} \;\; ; \;\; \mathbf{R}_{24}^{1} = \mathbf{Q}_{24}\mathbf{Q}_{44}^{-1} \;\; ;$$

$$\mathbf{R}_{31}^{1} = \mathbf{Q}_{31}\mathbf{Q}_{11}^{-1} \;\; ; \;\; \mathbf{R}_{32}^{1} = \mathbf{Q}_{32}\mathbf{Q}_{22}^{-1} \;\; ; \mathbf{R}_{34}^{1} = \mathbf{Q}_{34}\mathbf{Q}_{44}^{-1} \;\; ;$$

Step 2.

$$\mathbf{H}_{13}^{2} = -\mathbf{R}_{14}^{1}\mathbf{R}_{43}^{1} \;\; ; \;\; \mathbf{H}_{23}^{2} = -\mathbf{R}_{24}^{1}\mathbf{R}_{43}^{1} \;\; ; \;\; \mathbf{H}_{33}^{2} = -\mathbf{R}_{34}^{1}\mathbf{R}_{43}^{1} \;\; ;$$

$$\mathbf{R}_{13}^{2} = \left(\mathbf{R}_{13}^{1} + \mathbf{H}_{13}^{2}\right)\left(\mathbf{I} + \mathbf{H}_{33}^{2}\right)^{-1} \;\; ; \;\; \mathbf{R}_{23}^{2} = \left(\mathbf{R}_{23}^{1} + \mathbf{H}_{23}^{2}\right)\left(\mathbf{I} + \mathbf{H}_{33}^{2}\right)^{-1};$$

$$\mathbf{H}_{12}^{3} = -\mathbf{R}_{13}^{2}\mathbf{R}_{32}^{1} - \mathbf{R}_{14}^{1}\mathbf{R}_{42}^{1} + \mathbf{R}_{13}^{2}\mathbf{R}_{34}^{1}\mathbf{R}_{42}^{1};$$

$$\mathbf{H}_{22}^{3} = -\mathbf{R}_{23}^{2}\mathbf{R}_{32}^{1} - \mathbf{R}_{24}^{1}\mathbf{R}_{42}^{1} + \mathbf{R}_{23}^{2}\mathbf{R}_{34}^{1}\mathbf{R}_{42}^{1};$$

$$\mathbf{R}_{12}^{3} = \left(\mathbf{R}_{12}^{1} + \mathbf{H}_{12}^{3}\right)\left(\mathbf{I} + \mathbf{H}_{22}^{3}\right)^{-1}.$$

Step 3.

$$\mathbf{H}_{11}^{4} = -\mathbf{R}_{12}^{3}\mathbf{R}_{21}^{1} - \mathbf{R}_{13}^{2}\mathbf{R}_{32}^{1} - \mathbf{R}_{14}^{1}\mathbf{R}_{41}^{1} + \mathbf{R}_{12}^{3}\mathbf{R}_{23}^{2}\mathbf{R}_{31}^{1} + \mathbf{R}_{12}^{3}\mathbf{R}_{24}^{1}\mathbf{R}_{31}^{1} + \mathbf{R}_{13}^{2}\mathbf{R}_{34}^{1}\mathbf{R}_{41}^{1} - \mathbf{R}_{12}^{3}\mathbf{R}_{23}^{2}\mathbf{R}_{34}^{1}\mathbf{R}_{41}^{1}$$

$$\mathbf{B} = \mathbf{e} - \mathbf{R}_{12}^{3} \cdot \mathbf{e} - \mathbf{R}_{13}^{2} \cdot \mathbf{e} - \mathbf{R}_{14}^{1} \cdot \mathbf{e} + \mathbf{R}_{12}^{3}\mathbf{R}_{23}^{2} \cdot \mathbf{e} + \mathbf{R}_{12}^{3}\mathbf{R}_{24}^{1} \cdot \mathbf{e} + \mathbf{R}_{13}^{2}\mathbf{R}_{34}^{1} \cdot \mathbf{e} - \mathbf{R}_{12}^{3}\mathbf{R}_{23}^{2}\mathbf{R}_{34}^{1} \cdot \mathbf{e}$$

Step 4.

$$\boldsymbol{\pi}_1 = (1, \mathbf{0})\left[\mathbf{B}\middle|\left(\mathbf{I} + \mathbf{H}_{11}^{4}\right)^{*}\right]^{-1}$$

Step 5.

$$\boldsymbol{\pi}_2 = -\boldsymbol{\pi}_1\mathbf{R}_{12}^{3} \;\; ; \;\; \boldsymbol{\pi}_3 = -\boldsymbol{\pi}_1\mathbf{R}_{13}^{2} - \boldsymbol{\pi}_2\mathbf{R}_{23}^{2} \;\; ; \;\; \boldsymbol{\pi}_4 = -\boldsymbol{\pi}_1\mathbf{R}_{14}^{1} - \boldsymbol{\pi}_2\mathbf{R}_{24}^{1} - \boldsymbol{\pi}_3\mathbf{R}_{34}^{1}$$

3 Associated measures

Several associated measures such as the sojourn time in each level (macro-state) and the number of visits to each macro-state by time have been worked out.

3.1 Sojourn time. Phase-type distribution

One interesting question to answer at this point is what is the probability distribution of the sojourn time in a macro-state. It is well known that for the Markov process $J(t)$, the sojourn time in any state is exponentially distributed. For the stochastic process $X(t)$ is different.

We denote as $T(X(s))$ to the random sojourn time in macro-state $X(s)$ from time s. If the macro-state is known at time s, then

$$P\left\{T\left(X\left(s\right)\right) > t \middle| X\left(s\right) = \mathbf{i}\right\} = \frac{\mathbf{a}\left(s\right)\mathbf{A}_{\mathbf{i}} \cdot \exp\left\{\mathbf{A}_{\mathbf{i}}\mathbf{Q}\mathbf{A}_{\mathbf{i}}t\right\} \cdot \mathbf{e}}{\mathbf{a}\left(s\right)\mathbf{A}_{\mathbf{i}} \cdot \mathbf{e}}.$$

Therefore, the probability distribution of $T\left(X\left(s\right)\right)\middle|X\left(s\right)=\mathbf{i}$ for any s is phase-type distributed for any s with representation $\left(\frac{\mathbf{a}\left(s\right)\mathbf{A_i}}{\mathbf{a}\left(s\right)\mathbf{A_i}\cdot\mathbf{e}},\mathbf{A_i}\mathbf{Q}\mathbf{A_i}\right)$. Obviously, this distribution is the same than $\left(\frac{\mathbf{b}\left(s\right)}{\mathbf{b}\left(s\right)\cdot\mathbf{e}},\mathbf{Q_{ii}}\right)$, with b($s$) being the vector $\boldsymbol{\theta}\mathbf{P}\left(s\right)$ restricted to the states of the macro-state **i**.

If the process $\{X(t);\ t \geq 0\}$ has reached the stationary regime then, if the process is in macro-state **i**, the probability distribution of the sojourn time is PH with representation $\left(\frac{\boldsymbol{\pi}\mathbf{A_i}}{\boldsymbol{\pi}\mathbf{A_i}\cdot\mathbf{e}},\mathbf{A_i}\mathbf{Q}\mathbf{A_i}\right)$.

3.2 First step time

It is well-known that the first step time distribution for the process $\{J(t);\ t \geq 0\}$ from state l (out of macro-state **k**) to a macro-state **k** is phase-type distributed with representation $\left(\boldsymbol{\theta}_l,\mathbf{Q_{-2k}}\right)$ with $\boldsymbol{\theta}_l=\left(0,\ldots,0,1,0,\ldots,0\right)$, where the value 1 corresponds to state l and $\mathbf{Q_{-2k}}$ is the matrix Q with row and column blocks of zeros corresponding to the macro-state **k**. If we denote as $T_{\mathbf{h}}\left(s,\mathbf{k}\right)$ to the first step time from macro state **h**, at time s, to macro-state **k** then,

$$\begin{aligned}P\left(T_{\mathbf{h}}\left(s,\mathbf{k}\right)>t\right)&=\frac{\sum_{j\in\mathbf{k}}\sum_{i\in\mathbf{h}}P\left(T\left(j\right)>t\mid J\left(s\right)=i\right)P\left(J\left(s\right)=i\right)}{\sum_{i\in\mathbf{h}}P\left(J\left(s\right)=i\right)}\\&=\frac{\mathbf{a}\left(s\right)\mathbf{A_h}\cdot\exp\left\{\mathbf{Q_{-2k}}t\right\}\cdot\mathbf{e}}{\mathbf{a}\left(s\right)\mathbf{A_h}\mathbf{e}}.\end{aligned}$$

4 Number of visits to a macro-state

One interesting measure is the number of visits to a determined state between any two times s and t. We denote as $N_k\left(s,t\right)$ to the number of visits to the macro-state **k** from time s up to time t. We denote as $\mathbf{p}_k\left(n,s,t\right)$ to the matrix whose (i, j) element is

$$\begin{aligned}\left[\mathbf{p}_k\left(n,s,t\right)\right]_{ij}&=P\left\{N_k\left(s,t\right)=n,J\left(t\right)=j\mid J\left(s\right)=i\right\}\\&=P\left\{N_k\left(t-s\right)=n,J\left(t-s\right)=j\mid J\left(0\right)=i\right\}\\&=\left[\mathbf{p}_k\left(n,t-s\right)\right]_{ij}.\end{aligned}$$

The probability matrix verifies the following differential equations:

$$\mathbf{p}_k^{'}\left(n,t\right)=\mathbf{p}_k\left(n,t\right)\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)+\mathbf{p}_k\left(n-1,t\right)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right);\qquad n\geq1,$$

with initial condition $\mathbf{p}_k\left(n,0\right)=\mathbf{0}$, with $\tilde{\mathbf{Q}}_{kk}$ a matrix of zeros, with the same order than Q, except for the matrix block $\mathbf{Q}_{kk}$.

For n=0,

$$\mathbf{p}_k^{'}\left(0,t\right)=\mathbf{p}_k\left(0,t\right)\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)$$

$$\mathbf{p}_k\left(0,0\right)=\mathbf{I},$$

where, clearly, from the last two expressions we have that,

$$\mathbf{p}_k\left(0,t\right)=\exp\left\{\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)t\right\}.$$

To get the probability matrix we use Laplace transforms. It is well known that given a locally-integrable function $f(t)$, its Laplace transform is defined as $f^{*}(u)=\int_{0}^{\infty}e^{-ut}f(t)dt$. Therefore,

$$u\mathbf{p}_k^{*}(0,u)-\mathbf{I}=\mathbf{p}_k^{*}(0,u)\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right),$$

$$u\mathbf{p}_k^{*}(n,u)=\mathbf{p}_k^{*}(n,u)\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)+\mathbf{p}_k^{*}(n-1,u)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right);\qquad n\geq 1,$$

Then,

$$\mathbf{p}_k^{*}(0,u)=\left[u\mathbf{I}-\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)\right]^{-1},$$

$$\mathbf{p}_k^{*}(n,u)=\mathbf{p}_k^{*}(n-1,u)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)\left[u\mathbf{I}-\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)\right]^{-1};\qquad n\geq 1.$$

Thus, it can be proved that

$$\mathbf{p}_k^{*}(n,u)=\left[u\mathbf{I}-\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)\right]^{-1}\mathbf{A}^{n}(u);\qquad n\geq 0,$$

with A(u) being

$$\mathbf{A}(u)=\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)\left[u\mathbf{I}-\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)\right]^{-1}.$$

Taking inverse Laplace transform, the function $\mathbf{p}_k(n,t)$ is achieved and for the non-homogeneous Markov process $\{X(t);\ t\geq 0\}$ we have,

$$P\{N_{\mathbf{k}}(n,s,t)=n,X(t)=\mathbf{j}\,|\,X(s)=\mathbf{i}\}=\frac{\boldsymbol{\theta}\mathbf{P}(s)\mathbf{A}_{\mathbf{i}}\mathbf{p}_{\mathbf{k}}(n,t-s)\mathbf{A}_{\mathbf{j}}\cdot\mathbf{e}}{\boldsymbol{\theta}\mathbf{P}(s)\mathbf{A}_{\mathbf{i}}\cdot\mathbf{e}}.$$

Therefore,

$$P\{N_k(s,t)=n\}=\boldsymbol{\theta}\cdot\mathbf{P}(s)\cdot\mathbf{p}_k(n,t-s)\cdot\mathbf{e}$$, for n = 0, 1, 2,...

4.1 Expected number of visits to a determine macro-state

The mean number of visits to the macro-state **k** can be got from Section 4 as follows. This measure is given by

$$E\{N_k(t)\}=\boldsymbol{\theta}\cdot\sum_{n=0}^{\infty}n\cdot\mathbf{p}_k(n,t)\cdot\mathbf{e}=\boldsymbol{\theta}\cdot\mathbf{M}_k(t)\cdot\mathbf{e}.$$

From the differential equations above we have

$$\sum_{n=1}^{\infty}n\cdot\mathbf{p}_k^{'}(n,t)=\sum_{n=1}^{\infty}n\cdot\mathbf{p}_k(n,t)\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)+\sum_{n=1}^{\infty}n\cdot\mathbf{p}_k(n-1,t)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right),$$

$$\sum_{n=1}^{\infty}n\cdot\mathbf{p}_k^{'}(n,t)=\sum_{n=1}^{\infty}n\cdot\mathbf{p}_k(n,t)\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)+\sum_{n=1}^{\infty}n\cdot\mathbf{p}_k(n,t)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)+\mathbf{P}(t)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)$$

$$\mathbf{M}_k^{'}(t)=\mathbf{M}_k(t)\left[\left(\mathbf{Q}_{-k}+\tilde{\mathbf{Q}}_{kk}\right)+\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)\right]+\mathbf{P}(t)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right),$$

$$\mathbf{M}_k^{'}(t)=\mathbf{M}_k(t)\mathbf{Q}+\mathbf{P}(t)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right).$$

Given that **Q** is a conservative matrix then,

$$\mathbf{M}_k^{'}(t)\cdot\mathbf{e}=\mathbf{P}(t)\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)\cdot\mathbf{e},$$

with initial condition $\mathbf{M}_k(0)=\mathbf{0}$ if the initial state is not considered or $\mathbf{M}_k(0)=\mathbf{A}_k$ if the initial state is counted.

Therefore, for the first and second case we have, respectively

$$E\left[N_k(t)\right]=\boldsymbol{\theta}\cdot\mathbf{M}_k(t)\cdot\mathbf{e}=\boldsymbol{\theta}\cdot\int_0^t\mathbf{P}(u)du\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)\cdot\mathbf{e},$$

or

$$E\left[N_k(t)\right]=\boldsymbol{\theta}\cdot\mathbf{A}_k\cdot\mathbf{e}+\boldsymbol{\theta}\cdot\int_0^t\mathbf{P}(u)du\left(\mathbf{Q}_k-\tilde{\mathbf{Q}}_{kk}\right)\cdot\mathbf{e}.$$

5 Parameter estimation

The likelihood function when the exact change time between macro-states is known is achieved. For a device *l* we observe m_l transition times denoted as,

$0 = t_0^l, t_1^l, t_2^l, \ldots, t_{m_l-1}^l, t_{m_l}^l$,

where the last time is a complete or a censoring time.

The time t_a^l corresponds to the transition from x_a^l to x_{a+1}^l. These macro-states are $x_0^l, x_1^l, x_2^l, \ldots, x_{m_l-1}^l, x_{m_l}^l$.

This device contributes to the likelihood function

$$L_l = \boldsymbol{\alpha}_{x_0^l} \left[\prod_{a=0}^{m_l-1} e^{\mathbf{Q}_{x_a^l x_a^l}\left(t_{a+1}^l - t_a^l\right)} \left(\mathbf{Q}_{x_a^l x_{a+1}^l}\right)^{\tau_l} \right] \mathbf{e} ,$$

where τ_l is zero if the last time is a censoring time and one otherwise.

When *m* independent devices are considered,

$$L = \prod_{l=1}^{m} L_l = \prod_{l=1}^{m} \boldsymbol{\alpha}_{x_0^l} \left[\prod_{a=0}^{m_l-1} e^{\mathbf{Q}_{x_a^l x_a^l}\left(t_{a+1}^l - t_a^l\right)} \left(\mathbf{Q}_{x_a^l x_{a+1}^l}\right)^{\tau_l} \right] \mathbf{e} .$$

This function is maximized by considering that $\mathbf{Q}_{aa}$ is a square sub-stochastic matrix whose main diagonal is negative and the rest positive values and $\mathbf{Q}_{ab}$ a matrix with positive elements with $\sum_{b=1}^{r} \mathbf{Q}_{ab} \mathbf{e} = \mathbf{0}$ for any *a* and *b*.

Other log-likelihood function is given from the transition probabilities of the process

$$\log L = \sum_{l=1}^{m} \sum_{a=0}^{m_l-1} \log\left(h_{x_a^l x_{a+1}^l}\left(t_a^l, t_{a+1}^l\right)\right) .$$

6 Application of the developed methodology

We have made use of measurements of current-time traces in a unipolar RRAM. The RTN signals have an easy pattern, so different features of the device can be studied. Current traces show, for instance, the number of current levels in the signal and also the frequency at which each of these levels is active. The developed methodology allows modeling the internal states under an approach based on hidden Markov processes that produces the observed output (RTN signal measured). By analyzing the signals shown in Figure 1 we are able to determine these states and characterize them probabilistically speaking in order to understand their nature. Besides, it is possible to reproduce similar signals in case they are needed for circuit simulation or physical characterization of the traps that help to generate the signals.

Specifically, we are going to analyze the behaviour of several different current-time traces RTN25, RTN26, RTN27 that shows RTN for the described devices in the introduction. In addition, a long (more than 3 hours with millions of measured data) RTN current-time trace was measured and used here for the same device. Due to the fact that all signals come from the devices with the same characteristics of fabrication, the difference between them lies in the applied voltage that produces the variations of electric current. Besides, naturally, the measurement time for the long RTN is different. These measurements have previously been characterized in different ways [19, 20]; however, in this new approach, the internal Markov chain that leads to the observed data set is identified and the model, whose methodology is developed in this work, is estimated. In particular, in this application is shown that the short signals (RTN25, RTN26 and RTN27) have a Markovian internal behavior, whereas the developed new methodology will be applied to the long RTN trace for its modeling.

6.1 Series RTN25-26-27

As stated above, the signals RTN25, RTN26 and RTN27 are produced by devices whose structure is based on the Ni/HfO$_2$/Si stack. Nevertheless, different voltages were applied: 0.34 volts, 0.35 volts and 0.36 volts, respectively. On the other hand, the measurement time was similar for each one of these series.

Hidden Markov Models and the latent Markov chain for RTN25-26-27 traces

To study the number of possible latent levels hidden into the signals, we have considered hidden Markov models (HMM) [26, 27]. Whereas in simple Markov models the state of the device is visible to the observer at each time, in HMM only the output of the device is visible, while the state that leads to a determined output is hidden. Each state is associated to a set of transition probabilities (one per each state) defining how likely is for the system, being in a given state at a given instant of time, to switch to another of the possible states (including a transition to the same state) at the successive instants of time.

We have analysed different data sets for RTN25, RTN26, RTN27 and the hidden states from the corresponding signals have been discriminated. The best fit is achieved for 2 and 3 latent levels for RTN25, RTN26, RTN27. Figure 2 shows the original observed RTN signals and the corresponding latent levels given by the model for both cases. This analysis has been performed by using the package depmixS4 of R-cran [28].

The series RTN25-26-27 has been analyzed. The proportional number of times that the chain is in the latent states for the different proposed models is shown in Table 1.

Table 1. Proportional number of times that the chain is in the latent states.

Signal	Model	Latent state 1	Latent state 2	Latent state 3
RTN25	2 latent states model	0.784	---	0.216
	3 latent states model	0.762	0.037	0.201
RTN26	2 latent states model	0.756	---	0.244
	3 latent states model	0.753	0.0305	0.2165
RTN27	2 latent states model	0.7665	---	0.2335
	3 latent states model	0.7575	0.0255	0.2170

For each model, the continuous Markov chain associated to the latent states has been estimated and the corresponding stationary distribution is achieved. It is shown in Table 2.

Table 2. Stationary distribution for the different traces.

Signal	Model	Latent state 1	Latent state 2	Latent state 3
RTN25	2 latent states model	0.7974	---	0.2026
	3 latent states model	0.7847	0.0185	0.1968
RTN26	2 latent states model	0.7708	---	0.2292
	3 latent states model	0.7746	0.0165	0.2089
RTN27	2 latent states model	0.7736	---	0.2264
	3 latent states model	0.7785	0.0026	0.219

These values can be interpreted as the proportional time that the device is in each latent state in the stationary regime for the embedded continuous Markov chain. We can observe that these values are very close to zero for the RTN25-26-27 devices. We have tested this fact and it cannot be rejected.

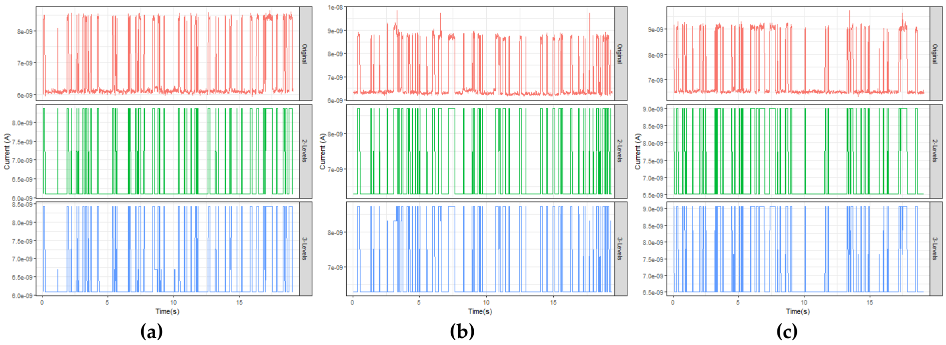


Figure 2. Fit obtained with the HMM for multiple latent states for the different RTN signals under study: (**a**) RTN25; (**b**) RTN26; (**c**) RTN27.

Therefore, we will assume that the internal performance of the devices is behaved as a Markov model with 2 latent states for the RTN25-26-27 traces. The Markov chains for 2 latent states for the different cases have been estimated. The exponentiality of the sojourn time in each state cannot be rejected according to the p-values obtained by means of Kolmogorov-Smirnov test, and the expected number of visits up to a certain time as explained in Section 4.1 has also been estimated. These estimations are shown in Table 3.

Table 3. Expected number of visits to each level for the different traces.

Signal	Level	t=5	t=10	t=15	t=20
RTN25	Level 1	12.3497	23.8609	35.3721	46.8833
	Level 2	11.5523	23.0635	34.5747	46.0859
RTN26	Level 1	12.8198	24.8162	36.8127	48.8092
	Level 2	12.0490	24.0455	36.0419	48.0384
RTN27	Level 1	11.5645	22.5305	33.4965	44.4626
	Level 2	11.7909	22.7569	33.7229	44.6890

6.2 Long RTN trace

A similar exploratory analysis for this trace has been carried out. For this signal it was supplied 0.5 volts and its behaviour was measured during more than 3 hours. To study the number of possible latent levels hidden into the signal, we have again considered hidden Markov chains. The best fit is achieved for 3 and 4 latent levels for this long RTN trace. Figure 3 shows the original observed signal (a time interval of the whole signal has been considered for the sake of computing feasibility) and the corresponding latent levels given by the model for both cases. This analysis has also been performed by using the package depmixS4 of R-cran.

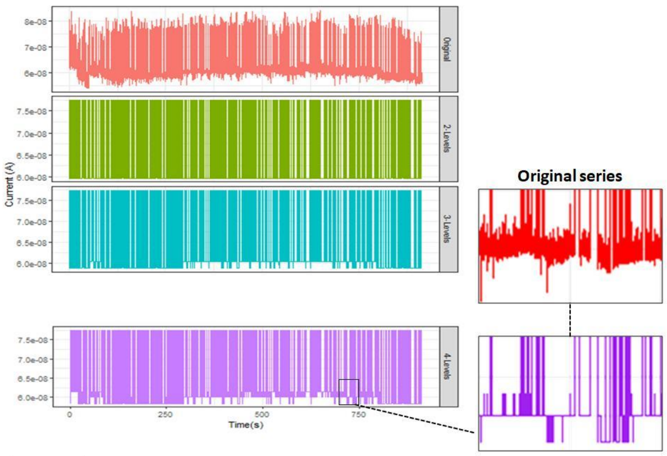


Figure 3. Fit obtained with the HMM for multiple latent states for the RTN different signals under study.

We have focused on the four latent states case. If the HMM is considered the proportional time that the process is in each latent states is (0.3243048, 0.1219429, 0.1635429, 0.3902095). If more latent states were assumed, negligible proportions would appear. Therefore, 4 different latent states are assumed. Before applying the new model to the data set, a statistical analysis was performed considering classical techniques. For these latent states, the exponentiality of the sojourn time has been tested and rejected through Kolmogorov-Smirnov test for any latent state by obtaining the following p-values; 0.00027, 0.0045, 0.0000 and 0.0001, respectively. Next, we have studied if these times could be described as PH distributions. After multiple analysis, the best PH distributions for the latent states have 2, 2, 4 and 3 internal states, respectively. The structures of these PH distributions are generalized Coxian/Erlang distributions. Then, the internal behaviour for each latent state passes across of multiple internal states in a sequential way one-by-one. The Anderson-Darling test has been applied to test the goodness of fit obtaining p-values 0.8972, 0.4405, 0.0752 and 0.9876, respectively, for each latent state.

Generator of the internal Markov Process

Given the previous analysis, we have observed that the sojourn time distribution in each macro-state (latent state) is phase-type distributed with a sequential degradation (generalized Coxian degradation). That is, the macro-states **1**, **2** are composed of two phases (internal states), the macro-state **3** of four phases and the macro-state **4** of three states. Thus, we assume the sojourn time in a macro-state **i** (level **i**) is PH distributed with representation $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$ with the following structure,

Macro-state 1	**Macro-state 2**
$\boldsymbol{\alpha}_1 = \left(\boldsymbol{\alpha}_1^1, 1-\boldsymbol{\alpha}_1^1\right)$	$\boldsymbol{\alpha}_2 = \left(\boldsymbol{\alpha}_1^2, 1-\boldsymbol{\alpha}_1^2\right)$
$\mathbf{T}_1 = \begin{pmatrix} -t_{12}^1 & t_{12}^1 \\ 0 & -t_{13}^1 \end{pmatrix}$	$\mathbf{T}_2 = \begin{pmatrix} -t_{12}^2 & t_{12}^2 \\ 0 & -t_{13}^2 \end{pmatrix}$
$\mathbf{T}_1^0 = \left(0, t_{13}^1\right)'$	$\mathbf{T}_2^0 = \left(0, t_{13}^2\right)'$

Macro-state 3	**Macro-state 4**
$\boldsymbol{\alpha}_3=\left(\boldsymbol{\alpha}_1^3,\boldsymbol{\alpha}_2^3,\boldsymbol{\alpha}_3^3,1-\boldsymbol{\alpha}_1^3-\boldsymbol{\alpha}_2^3-\boldsymbol{\alpha}_3^3\right)$	$\boldsymbol{\alpha}_4=\left(\boldsymbol{\alpha}_1^4,\boldsymbol{\alpha}_2^4,1-\boldsymbol{\alpha}_1^4-\boldsymbol{\alpha}_2^4\right)$
$\mathbf{T}_3=\begin{pmatrix}-t_{12}^3 & t_{12}^3 & 0 & 0\\ 0 & -t_{23}^3 & t_{23}^3 & 0\\ 0 & 0 & -t_{34}^3 & t_{34}^3\\ 0 & 0 & 0 & -t_{35}^3\end{pmatrix}$	$\mathbf{T}_4=\begin{pmatrix}-t_{12}^4 & t_{12}^4 & 0\\ 0 & -t_{23}^4 & t_{23}^4\\ 0 & 0 & -t_{34}^4\end{pmatrix}$
$\mathbf{T}_3^0=\left(0,0,0,t_{35}^3\right)'$	$\mathbf{T}_4^0=\left(0,0,0,t_{34}^4\right)'$

We assume that when the device leaves the macro-state **i**, it goes to macro-state **j** with a probability of p_{ij}, where $p_{ii}=0$ and $p_{14}=1-\sum_{i=1}^{3}p_{ij}$, and the sojourn time in this new macro-state begins with the corresponding initial distribution $\boldsymbol{\alpha}_j$.

Thus, the behaviour of the device is governed by a stochastic process $\{X(t);\ t\geq 0\}$ with an embedded Markov process $\{J(t);\ t\geq 0\}$ as described in Section 2.1. The macro-state space is $\{\mathbf{1},\mathbf{2},\mathbf{3},\mathbf{4}\}$ and the state space $\left\{1=i_1^1,2=i_2^1;3=i_1^2,4=i_2^2;5=i_1^3,6=i_2^3,7=i_3^3,8=i_4^3;9=i_1^4,10=i_2^4,11=i_3^4\right\}$.

For this case, the matrix blocks are $\mathbf{Q}_{ii}=\mathbf{T}_i$ and $\mathbf{Q}_{ij}=p_{ij}\mathbf{T}_i^0\otimes\boldsymbol{\alpha}_j$ for i, j = 1,...,r and $i\neq j$. Therefore, the matrix generator is given by

$$\mathbf{Q}=\begin{pmatrix}\mathbf{T}_1 & p_{12}\mathbf{T}_1^0\otimes\boldsymbol{\alpha}_2 & p_{13}\mathbf{T}_1^0\otimes\boldsymbol{\alpha}_3 & p_{14}\mathbf{T}_1^0\otimes\boldsymbol{\alpha}_4\\ p_{21}\mathbf{T}_2^0\otimes\boldsymbol{\alpha}_1 & \mathbf{T}_2 & p_{23}\mathbf{T}_2^0\otimes\boldsymbol{\alpha}_3 & p_{24}\mathbf{T}_2^0\otimes\boldsymbol{\alpha}_4\\ p_{31}\mathbf{T}_3^0\otimes\boldsymbol{\alpha}_1 & p_{32}\mathbf{T}_3^0\otimes\boldsymbol{\alpha}_2 & \mathbf{T}_3 & p_{34}\mathbf{T}_3^0\otimes\boldsymbol{\alpha}_4\\ p_{41}\mathbf{T}_4^0\otimes\boldsymbol{\alpha}_1 & p_{42}\mathbf{T}_4^0\otimes\boldsymbol{\alpha}_2 & p_{43}\mathbf{T}_4^0\otimes\boldsymbol{\alpha}_3 & \mathbf{T}_4\end{pmatrix}.$$

The parameters have been estimated by considering the likelihood function built in Section 5. The estimated parameters with a value of the logL= −4066.118120 are

$$\mathbf{P}=\begin{pmatrix}0 & 0.6667 & 0.0407 & 0.2926\\ 0.3870 & 0 & 0.1969 & 0.4161\\ 0.0296 & 0.2238 & 0 & 0.7466\\ 0.1605 & 0.3395 & 0.5 & 0\end{pmatrix};$$

$$\boldsymbol{\alpha}_1=\left(0.5730374,0.4269626\right);\mathbf{T}_1=\begin{pmatrix}-0.6790043 & 0.6790043\\ 0 & -4.1343018\end{pmatrix};\boldsymbol{\alpha}_2=\left(0.4699825,0.5300175\right);$$

$$\mathbf{T}_2=\begin{pmatrix}-3.249849 & 3.249849\\ 0 & -10.426533\end{pmatrix};\boldsymbol{\alpha}_3=\left(0.0741494,0.4258142,0.5000364,0\right);$$

$$\mathbf{T}_3=\begin{pmatrix}-0.5533471 & 0.5533471 & 0 & 0\\ 0 & -2.2755462 & 2.2755462 & 0\\ 0 & 0 & -29.535695 & 29.535695\\ 0 & 0 & 0 & -242.7465\end{pmatrix}$$

$$\boldsymbol{\alpha}_4=\left(0.3593538,0.6406462,0.0000\right)\ ;\ \mathbf{T}_4=\begin{pmatrix}-0.964899 & 0.964899 & 0\\ 0 & -4.112655 & 4.112655\\ 0 & 0 & -28.638854\end{pmatrix}.$$

Therefore, the generator of the Markov process $\{J(t);\ t\geq 0\}$ is

$$\hat{\mathbf{Q}} = \left(\begin{array}{cc|cc|cccc|ccc}
-0.6790 & 0.6790 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -4.1343 & 1.2954 & 1.4609 & 0.0125 & 0.0717 & 0.0841 & 0 & 0.4347 & 0.7750 & 0 \\
\hline
0 & 0 & -3.2498 & 3.2498 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2.3122 & 1.7228 & 0 & -10.4265 & 0.1522 & 0.8742 & 1.0266 & 0 & 1.5590 & 2.7794 & 0 \\
\hline
0 & 0 & 0 & 0 & -0.5533 & 0.5533 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -2.2755 & 2.2755 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -29.5357 & 29.5357 & 0 & 0 & 0 \\
4.1174 & 3.0679 & 25.5326 & 28.7941 & 0 & 0 & 0 & -242.7465 & 65.1273 & 116.1072 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.9649 & 0.9649 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4.1127 & 4.1127 \\
2.6340 & 1.9625 & 4.5696 & 5.1533 & 1.0618 & 6.0974 & 7.1602 & 0 & 0 & 0 & -28.6389
\end{array}\right)$$

with initial distribution $\hat{\boldsymbol{\theta}} = (0,0,0,0,0.0741494,0.4258142,0.5000364,0,0,0,0)$.

The stationary distribution has been estimated for the process $\{X(t);\ t \geq 0\}$ (given in Table 4) from Section 2.2. It can be interpreted as the proportional time in each macro-state (long-run).

Table 4. Stationary distribution for the process $\{X(t);\ t \geq 0\}$

	Level 1	Level 2	Level 3	Level 4
Selected interval in the long RTN trace	0.3273	0.1197	0.1612	0.3919

Finally, the mean number of visits to each macro-state up to a certain time t has been worked out following Section 4.1. It is shown in Table 5.

Table 5. Expected number of visits up to a certain time for different times

Time	Level 1	Level 2	Level 3	Level 4
t =50	16.0207	25.0716	20.3974	30.0827
t =100	31.0837	49.9364	40.9591	60.1877
t =200	61.1925	99.6404	82.0612	120.3666
t =500	151.4018	248.5475	205.1981	300.6553

7 Conclusions

A real problem motivates the construction of a new stochastic process accounting for the internal performance of different macro-states by considering that follows an internal Markovian behaviour. It is proved that the homogeneity and Markovianity is lost for the developed new macro-state model. The sojourn time in each macro-state is phase-type distributed depending on the initial observed time. The stationary distribution has been calculated through matrix-algorithmic methods and the number of visits distribution to a determined macro-state between any two different times is calculated from Laplace transform. The mean number of visits depending on times is worked out explicitly.

The new developed methodology enables to model complex systems in an algorithmic way solving classic calculus problems. Also, thanks to the proposed development, the results and measures are worked out in an easier way and they can be interpreted. Not only matrix analysis and Laplace transform techniques are used to determine the properties of the model, but algorithms to obtain quantitative results are provided. Given that everything is carried out algorithmically, it is implemented computationally and is successfully applied to study different Random Telegraph Noise signals measured for unipolar resistive memories.

Resistive memory Random Telegraph Noise signals have been analyzed in depth in order to characterize them from the probabilistic point of view. These signals are essential since they can pose a limit in the performance of certain applications; in addition, this type of noise can be used for good as an entropy source for the design of random number generators for cryptography. From the proposed model, we prove that a latent state of a resistive memory RTN long signal is composed of multiple internal states. Of course, the applications of a phase-type model, as given in this work, are not restricted to the RRAM context.

References

[1] Neuts, M.F. *Probability distributions of phase type;* Liber amicorum professor emeritus H. Florin. Belgium: Department of Mathematics, University of Louvain, 183206, 1975.

[2] Neuts, M.F. *Matrix geometric solutions in stochastic models. An algorithmic approach*; Baltimore: John Hopkins University Press, 1981.

[3] Ruiz-Castro; J.E. Complex multi-state systems modelled through Marked Markovian Arrival Processes. *Eur. J. Oper. Res.* **2016**, *252,* 852-865. [DOI]

[4] Ruiz-Castro, J.E. A complex multi-state k-out-of-n: G system with preventive maintenance and loss of units. *Reliab. Eng. Syst. Safe.* **2020**, *197*, 106797. [DOI]

[5] Ruiz-Castro, J.E.; Dawabsha, M. A multi-state warm standby system with preventive maintenance, loss of units and an indeterminate multiple number of repairpersons. *Comput. Ind. Eng.* **2020**, *142*, 106348. [DOI]

[6] Artalejo, J.R.; Chakravarthy, S.R. Algorithmic Analysis of the MAP/PH/1 Retrial Queue. *Top* **2006**, *14*, 293-332. [DOI]

[7] Asmussen, S.; Bladt, M. Phase-type distribution and risk processes with state-dependent premiums. *Scand. Actuar. J.* **1996**, *1*, 19-36. [DOI]

[8] Ruiz-Castro, J.E.; Acal, C.; Aguilera, A.M.; Aguilera-Morillo, M.C.; Roldán, J.B. Linear-Phase-Type probability modelling of functional PCA with applications to resistive memories. *Math. Comput. Simulat.* in press. [DOI]

[9] Asmussen, S. *Ruin probabilities*; World Scientific, Chinese, 2000.

[10] He, Q.M. *Fundamentals of Matrix-Analytic Methods;* Springer, USA, 2014.

[11] Carboni, R.; Ielmini, D. Stochastic Memory Devices for Security and Computing. *Adv. Electron. Mater.* **2019**, *5*, 1900198. [DOI]

[12] Aldana, S.; Roldán, J.B.; García-Fernández, P.; Suñe, J.; Romero-Zaliz, R.; Jiménez-Molinos, F.; Long, S.; Gómez-Campos, F.; Liu, M. An in-depth description of bipolar resistive switching in Cu/HfOx/Pt devices, a 3D Kinetic Monte Carlo simulation approach. *J. Appl. Phys.* **2018**, *123*, 154501. [DOI]

[13] Chual L.O. Memristor-the missing circuit element. *IEEE T. Circuits Syst.* **1971**, *18*, 507-519. [DOI]

[14] Puglisi, F.M.; Zagni, N.; Larcher, L.; Pavan, P. Random Telegraph Noise in Resistive Random Access Memories: Compact Modeling and Advanced Circuit Design. *IEEE T. Electron Dev.* **2018**, *65*, 2964-2972. [DOI]

[15] Puglisi, F.M.; Larcher, L.; Padovani, A.; Pavan, P. A Complete Statistical Investigation of RTN in HfO2-Based RRAM in High Resistive State. *IEEE T. Electron Dev.* **2015**, *62*, 2606-2613. [DOI]

[16] Alonso, F.J.; Maldonado, D.; Aguilera, A.M.; Roldán, J.B. Memristor variability and stochastic physical properties modeling from a multivariate time series approach. *Chaos Soliton Fract.* **2021**, *143*, 110461. [DOI]

[17] Aguilera-Morillo, M.C.; Aguilera, A.M.; Jiménez-Molinos, F., Roldán, J.B. Stochastic modeling of Random Access Memories reset transitions. *Math. Comput. Simulat.* **2019**, *159*, 197–209. [DOI]

[18] Simoen, E.; Claeys, C. *Random Telegraph Signals in Semiconductor Devices*; IOP Publishing, 2017.

[19] González-Cordero, G.; González, M.B.; Jiménez-Molinos, F.; Campabadal, F.; Roldán, J.B. New method to analyze random telegraph signals in resistive random access memories. *J. Vac. Sci. Technol. B* **2019**, *37*, 012203. [DOI]

[20] González-Cordero, G.; González, M.B.; Morell, A.; Jiménez-Molinos, F.; Campabadal, F.; Roldán, J.B. Neural network based analysis of Random Telegraph Noise in Resistive Random Access Memories. *Semicond. Sci. Tech.* **2020**, *35*, 025021. [DOI]

[21] Grasser, T. *Noise in Nanoscale Semiconductor Devices*; Springer, 2020.

[22] Wei, Z.; Katoh, Y.; Ogasahara, S.; Yoshimoto, Y.; Kawai, K.; Ikeda, Y.; Eriguchi, K.; Ohmori, K.; Yoneda, S. True random number generator using current difference based on a fractional stochastic model in 40-nm embedded ReRAM. 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, 2016, 4.8.1-4.8.4. [DOI]

[23] Chen, X; Wang, L.; Li, B.; Wang, Y.; Li, X.; Liu, Y.; Yang, H. Modeling Random Telegraph Noise as a Randomness Source and its Application in True Random Number Generation. *IEEE T. Comput. Aid. D.* **2016**, 35, 1435-1448. [DOI]

[24] Acal, C.; Ruiz-Castro, J.E.; Aguilera, A.M.; Jiménez-Molinos, F.; Roldán, J.B. Phase-type distributions for studying variability in resistive memories. *J. Comput. Appl. Math.* **2019**, *345*, 23-32. [DOI]

[25] González, M.B.; Rafí, J.M.; Beldarrain, O.; Zabala, M.; Campabadal, F. Analysis of the Switching Variability in Ni/HfO2 -Based RRAM Devices. *IEEE T. Device Mat. Re.* **2014**, *14*, 769-771. [DOI]

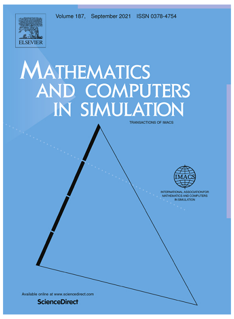
[26] Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *P IEEE* **1989,** *77*, 257-286. [DOI]

[27] Puglisi, F.M.; Pavan, P. Factorial Hidden Markov Model analysis of Random Telegraph Noise in Resistive Random Access Memories. *ECTI T. Ele. Eng. Electron. Commun.* **2014**, *12*, 24-29.

[28] Visser, I.; Speekenbrink, M. depmixS4: An R Package for Hidden Markov Models. *J. Stat. Softw.* **2010**, *36,* 1-21. [DOI]

A3 Linear Phase-Type probability modelling of functional PCA with applications to resistive memories

- Ruiz-Castro, Juan Eloy; Acal, Christian; Aguilera, Ana M.; Aguilera-Morillo, M. Carmen; Roldan, Juan B. (2021)
- Linear Phase-Type probability modelling of functional PCA with applications to resistive memories
- *Mathematics and Computers in Simulation*, vol. 186, pp. 71-79
- DOI: `https://doi.org/10.1016/j.matcom.2020.07.006`


Volume 187, September 2021 ISSN 0378-4754
ELSEVIER
MATHEMATICS AND COMPUTERS IN SIMULATION
TRANSACTIONS OF IMACS
IMACS
Available online at www.sciencedirect.com
ScienceDirect

Mathematics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	1.620	68/261	Q2

Abstract

Functional principal component analysis (FPCA) based on Karhunen-Loève (K-L) expansion allows to describe the stochastic evolution of the main characteristics associated to multiple systems and devices. Identifying the probability distribution of the principal component scores is fundamental to characterize the whole process. The aim of this work is to consider a family of statistical distributions that could be accurately adjusted to a previous transformation. Then, a new class of distributions, the linear-phase-type, is introduced to model the principal components. This class is studied in detail in order to prove, through the K-L expansion, that certain linear transformations of the process at each time point are phase-type distributed. This way, the one-dimensional distributions of the process are in the same linear-phase-type class. Finally, an application to model the reset process associated with resistive memories is developed and explained.

1 Introduction

Among the electron devices with greater potential in the current microelectronic industry landscape are Resistive Random Access Memories (RRAMs). The number of indexed publications in this field has skyrocketed and therefore, the attention of the academic community as well as the electronics companies' development teams is fixed on them. The applications of these new devices range from non-volatile memory circuits, security modules for cryptography and neuromorphic computation [10].

The stochastic nature of the physical mechanisms behind RRAM resistive switching (RS) operation makes the statistical modelling of the inherent device stochasticity essential. The key issue here rests upon the need to correctly explain variability in the current/voltage curves associated with long series of successive RS cycles [3, 11, 13, 1], i.e., cycles of continuous reset and set processes. If the device charge conduction is filamentary, the most common case, RS cycles get translated into rupture and rejuvenation of conductive filaments that dramatically changes the device resistance [7]. The modelling of the current versus voltage curves in these devices is of most importance for circuit design. Therefore, in this context, and taking into consideration that the experimental data we have are curves, an approach based on functional data analysis (FDA) can be applied in order to accurately model resistive memory characteristics.

A deep description of the main FDA methods with applications in different fields was developed in [12]. Functional principal component analysis (FPCA) based on Karhunen-Loève (K-L) expansion provides an orthogonal representation of an stochastic process in terms of uncorrelated random variables, called principal components (p.c.'s). The K-L expansion can be truncated so that the process is approximated in terms of the most explicative p.c.'s [2]. A three step algorithm for estimating FPCA from the reset curves (current versus voltage curves) of a sample of RRAM cycles was proposed in [3]. This new type of modelling can be very attractive from the circuit simulation viewpoint because it allows to describe the main characteristics of these devices, such as variability. Making use of this technique, the implementation of variability in compact models for RRAMs can be greatly simplified.

Nevertheless, identifying the probability distribution of the principal components is fundamental to characterize the whole process through the K-L expansion. In previous studies, several authors have considered different transformations and used them as a starting point, they have fitted different distributions successfully. However, to find the appropriate transformation and its probability distribution is not an easy task. The aim of this work is to consider a family of statistical distributions that could be accurately adjusted for any transformation. In this respect, a new methodology is developed by considering phase-type distributions (PH) that were applied in [1, 11] for modelling the reliability functions associated to RRAM reset points, among others parameters. This class of distributions have been also considered in other multiple science fields such as queueing theory and reliability ([16], [14], [15]). The properties of this distribution class are very interesting and allow to achieve results in a well structured form. The developments and results can be expressed in an matrix-algorithmic and computational way. One of the main advantages of this class is that any non-negative distribution can be approximated as needed through a PH distribution [9]. In order to fit this distribution, the p.c.'s scores should be transformed previously to positive values. In fact, in this research, it is proved that for several transformations, the fit obtained is more accurate by considering PH distributions than any other distribution. A new class of distributions are introduced, the linear-phase-type distributions (LPH) defined as variables for which there is a linear transformation that is PH distributed. This class is studied in detail in order to prove, through K-L expansion, that certain linear transformations of the process at each time point is PH distributed too.

In addition to this introduction, the paper has three other sections. The new LPH distributions and their main properties are studied in detail in Section 2. Then, the one-dimensional LPH distributions of the process are obtained from the LPH distributions of the p.c.'s through the K-L expansion. Finally, the proposed methodology is applied on different samples of current/voltage curves associated to RRAM devices in Section 4.

2 LPH modelling

In reliability, computer and electronic engineering, physics, queues theory and other fields, multiple probability distributions are frequently used, including the exponential, Erlang and Weibull distributions. Most of them involve calculations that may become unmanageable, due to the analytic expressions required. PH play an important role in this respect. This type of distributions enables us to express the main results in an algorithmic and computational way. This class of distribution was described in detail in [9].

2.1 PH distributions

Definition 1 A nonnegative random variable X is a PH distribution if its reliability function is given by

$$R\left(x\right)=P\left\{X>x\right\}=\boldsymbol{\alpha}e^{\mathbf{T}x}\mathbf{e}\quad;\quad x\geq 0,$$

where $\boldsymbol{\alpha}$ is a substochastic vector of order m, $\mathbf{T}$ a subgenerator of order m (matrix $m\times m$ where all diagonal elements are negative, all off-diagonal elements are non-negative, invertible and all row sums are non-positive) and, throughout the paper, $\mathbf{e}$ is a column vector of ones with appropriate order.

A PH distribution can be defined as the time up to the absorption in an absorbent Markov chain with initial distribution and generator for the transient states $\boldsymbol{\alpha}$ and $\mathbf{T}$, respectively. In this case, $(\boldsymbol{\alpha},\mathbf{T})$ is called the representation of the PH distribution.

Multiple good properties of these distributions are described in [9]. One of the main properties is that of PH distributions can approximate arbitrarily closely any probability distribution defined on the nonnegative real line.

2.2 LPH distributions

A new probability distribution class is defined in this subsection. This class is called the linear-phase-type distribution class (LPH). A LPH distribution is defined as follows.

Definition 2 A random variable X follows a LPH distribution if $Y=a+bX$ is PH distributed for a and b ($b\neq 0$) in $\mathbb{R}$.

If the representation of Y is $(\boldsymbol{\alpha},\mathbf{T})$ then the reliability function of X (LPH) is

$$R_X\left(x\right)=P\left(X>x\right)=\left\{\begin{array}{lll}\boldsymbol{\beta}e^{\mathbf{S}x}\mathbf{e} & ; & \text{for } x>\frac{-a}{b}; b>0\\ 1-\boldsymbol{\beta}e^{\mathbf{S}x}\mathbf{e} & ; & \text{for } x<\frac{-a}{b}; b<0\end{array}\right.,$$

where $\boldsymbol{\beta}=\boldsymbol{\alpha}e^{\mathbf{T}a}$, $\mathbf{S}=b\mathbf{T}$ and $\mathbf{e}$ is a column vector with appropriate order. In this case, will we denote the 4-tuple $(a,b,\boldsymbol{\beta},\mathbf{S})$ as the representation of the corresponding LPH.

The density function of this class of distributions is given by

$$f_X\left(x\right)=\left\{\begin{array}{lll}-\boldsymbol{\beta}e^{\mathbf{S}x}\mathbf{S}^0 & ; & \text{for } x>\frac{-a}{b}; b>0\\ \boldsymbol{\beta}e^{\mathbf{S}x}\mathbf{S}^0 & ; & \text{for } x<\frac{-a}{b}; b<0\end{array}\right.,$$

[1] where $\mathbf{S}^0$ is the column vector -$\mathbf{Se}$=-b$\mathbf{Te}$=b$\mathbf{T}^0$.

The moment-generating function is given by $M_X(t) = -\boldsymbol{\beta}(\mathbf{S}+\mathbf{I}t)^{-1}e^{-(\mathbf{S}+\mathbf{I}t)a/b}\mathbf{S}^0$, and then $E[X^n] = \left.\frac{\partial^n M_X(t)}{\partial t^n}\right|_{t=0}$.

From this expression the first and second moments are

$$E[X] = -\boldsymbol{\beta}e^{-\mathbf{S}a/b}\mathbf{S}^{-1}\mathbf{e} - \tfrac{a}{b}$$

$$E\left[X^2\right] = \tfrac{1}{b^2}\left[2\boldsymbol{\beta}\mathbf{e}^{-\mathbf{S}a/b}\mathbf{S}^{-1}\left(\tfrac{1}{b^2}\mathbf{S}^{-1}+\tfrac{a}{b}\mathbf{I}\right)\mathbf{e}+a^2\right].$$

Consequently,

$$Var(X) = \frac{1}{b^4}\left[2\boldsymbol{\beta}e^{-\mathbf{S}a/b}\mathbf{S}^{-2}\mathbf{e} - \left(\boldsymbol{\beta}e^{-\mathbf{S}a/b}\mathbf{S}^{-1}\mathbf{e}\right)^2\right].$$

Let's see that the finite addition of independent PH distributions or homothecy of PH distributions is PH distributed.

Result 1 (Summation of independent PH distributions)
Let $\{Y_i; i = 1, \ldots, n\}$ be a finite sequence of independent PH distributions with representation $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$ for i=1,...,n. Then, the variable $W_n = \sum_{i=1}^n Y_i$ is PH distributed with representation $(\boldsymbol{\rho}_n, \mathbf{L}_n)$ given by $\boldsymbol{\rho}_n = (\boldsymbol{\alpha}_1, \mathbf{0})$ and

$$\mathbf{L}_n = \begin{pmatrix} \mathbf{T}_1 & \mathbf{T}_1^0 \otimes \boldsymbol{\alpha}_2 & & & & & \\ & \mathbf{T}_2 & \mathbf{T}_2^0 \otimes \boldsymbol{\alpha}_3 & & & & \\ & & \mathbf{T}_3 & \mathbf{T}_3^0 \otimes \boldsymbol{\alpha}_4 & & & \\ & & & \ddots & \ddots & & \\ & & & & \mathbf{T}_{n-1} & \mathbf{T}_{n-1}^0 \otimes \boldsymbol{\alpha}_n \\ & & & & & \mathbf{T}_n \end{pmatrix},$$

where $\otimes$ is the Kronecker product defined as follows. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices with order $m \times n$ and $k \times l$ respectively. Then, $\mathbf{A} \otimes \mathbf{B}$ is a matrix with order $mk \times nl$ defined as $(a_{ij}\mathbf{B})$.

Proof.
The proof of this result is developed through induction. It is well known that the distribution of W_2 is given by the convolution of Y_1 and Y_2. If we denote to the distribution function of Y_i as F_i then the distribution function of W_2, convolution of F_1 and F_2, denoted by $*$, is

[1] Throughout the paper, if $\mathbf{A}$ is a matrix then $\mathbf{A}^0$= -$\mathbf{Ae}$ being $\mathbf{e}$ a column vector of ones with appropriate order

$$W_2\left(t\right) = F_1 * F_2\left(t\right) = \int_0^{\infty} F_1\left(du\right) F_2\left(t-u\right) du.$$

It is well-known that the Laplace-Stieltjes transform of the convolution is the product of the Laplace-Stieltjes transforms and that there is a biunivocal relationship between the original distribution and its Laplace-Transform.

Given the distribution function of a PH distribution with representation $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$, then its Laplace-Stieltjes transform is given by

$$F_i^*\left(s\right) = \boldsymbol{\alpha}_i (s\mathbf{I} - \mathbf{T}_i)^{-1} \mathbf{T}_i^0.$$

Then,

$$\begin{aligned} W_2^*\left(s\right) &= \boldsymbol{\rho}_2 (s\mathbf{I} - \mathbf{L}_2)^{-1} \mathbf{L}_2^0 = (\boldsymbol{\alpha}_1, \mathbf{0}) \begin{pmatrix} s\mathbf{I} - \mathbf{T}_1 & \mathbf{T}_1^0 \otimes \boldsymbol{\alpha}_2 \\ \mathbf{0} & s\mathbf{I} - \mathbf{T}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{T}_2^0 \end{pmatrix} \\ &= (\boldsymbol{\alpha}_1, \mathbf{0}) \begin{pmatrix} (s\mathbf{I} - \mathbf{T}_1)^{-1} & (s\mathbf{I} - \mathbf{T}_1)^{-1} \mathbf{T}_1^0 \boldsymbol{\alpha}_2 (s\mathbf{I} - \mathbf{T}_2)^{-1} \\ \mathbf{0} & (s\mathbf{I} - \mathbf{T}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{T}_2^0 \end{pmatrix} \\ &= \boldsymbol{\alpha}_1 (s\mathbf{I} - \mathbf{T}_1)^{-1} \mathbf{T}_1^0 \cdot \boldsymbol{\alpha}_2 (s\mathbf{I} - \mathbf{T}_2)^{-1} \mathbf{T}_2^0 = F_1^*\left(s\right) \cdot F_2^*\left(s\right). \end{aligned}$$

We assume that $W_{n-1} = \sum\limits_{i=1}^{n-1} Y_i$ is PH-distributed with representation $\left(\boldsymbol{\rho}_{n-1}, \mathbf{L}_{n-1}\right)$. Given that $W_n = W_{n-1} + Y_n$ and $\boldsymbol{\rho}_n = \left(\boldsymbol{\rho}_{n-1}, \mathbf{0}\right)$ and $\mathbf{L}_n = \begin{pmatrix} \mathbf{L}_{n-1} & \mathbf{L}_{n-1}^0 \\ \mathbf{0} & \mathbf{T}_n^0 \end{pmatrix}$, then

$$W_n^*\left(s\right) = \boldsymbol{\rho}_n (s\mathbf{I} - \mathbf{L}_n)^{-1} \mathbf{L}_n^0 =$$

$$\boldsymbol{\rho}_{n-1} (s\mathbf{I} - \mathbf{L}_{n-1})^{-1} \mathbf{L}_{n-1}^0 \cdot \boldsymbol{\alpha}_n (s\mathbf{I} - \mathbf{T}_n)^{-1} \mathbf{T}_n^0 = W_{n-1}^*\left(s\right) \cdot F_n^*\left(s\right).$$

Corollary 1

Let $\{X_i; i = 1, \ldots, n\}$ be a finite sequence of independent LPH distributions with PH-distributions associated given by $\{Y_i = a_i + bX_i; i = 1, \ldots, n\}$ with representation $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$ for i=1,...,n. Then, the variable $\Lambda_n = \sum_{i=1}^n X_i$ is LPH distributed with representation $\left(\sum\limits_{i=1}^n a_i, b, \boldsymbol{\rho}_n e^{\mathbf{L}_n \sum\limits_{i=1}^n a_i}, b\mathbf{L}_n\right)$.

Proof.

From result 1, $\sum\limits_{i=1}^n Y_i = b\Lambda_n + \sum\limits_{i=1}^n a_i$ is PH with representation $(\boldsymbol{\rho}_n, \mathbf{L}_n)$. Then, $\Lambda_n = \frac{1}{b} \sum\limits_{i=1}^n Y_i - \frac{1}{b} \sum\limits_{i=1}^n a_i$ is LPH with representation $\left(\sum\limits_{i=1}^n a_i, b, \boldsymbol{\rho}_n e^{\mathbf{L}_n \sum\limits_{i=1}^n a_i}, b\mathbf{L}_n\right)$.

Next, we show that a positive homothecy of a PH distribution is also PH distributed.

Result 2
Let Y be a PH distribution with representation $(\boldsymbol{\alpha}, \mathbf{T})$ then the variable γY is PH distributed with representation $(\boldsymbol{\alpha}, \frac{1}{\gamma}\mathbf{T})$, being γ a non-negative real number.
The proof of this result is immediate. Thus,

$$P(\gamma Y > t) = P(Y > t/\gamma) = \boldsymbol{\alpha} e^{\frac{1}{\gamma}\mathbf{T}t}\mathbf{e} \quad ; \quad t > 0.$$

Corollary 2
Let X be a LPH distribution with representation $(a, b, \boldsymbol{\beta}, \mathbf{S})$, then the variable γX is LPH with representation $\left(|\gamma|\, a, b \cdot sgn(\gamma), \boldsymbol{\beta}, \frac{1}{\gamma}\boldsymbol{S}\right)$, being γ a non-zero real number, $|\cdot|$ the absolute value function and $sgn(\cdot)$ the sign function.

Proof.
If X is a LPH distribution with representation $(a, b, \boldsymbol{\beta}, \mathbf{S})$, then there exist a and b such that $Y = a + bX$ is PH$(\boldsymbol{\alpha}, \mathbf{T})$ where $\boldsymbol{\beta} = \boldsymbol{\alpha} e^{\mathbf{T}a}$ and $\mathbf{S} = b\mathbf{T}$.

Then, from *Result 2* we have that any homothecy of a LPH distribution is also LPH distributed.

- If $\gamma > 0$, $\gamma Y = \gamma a + b\gamma X$ is PH$\left(\boldsymbol{\alpha}, \frac{1}{\gamma}\mathbf{T}\right)$.
  Then,
  γX is LPH with representation $\left(\gamma a, b, \boldsymbol{\alpha} e^{\mathbf{T}a}, \frac{b}{\gamma}\mathbf{T}\right) \equiv \left(\gamma a, b, \boldsymbol{\beta}, \frac{1}{\gamma}\mathbf{S}\right)$.

- If $\gamma < 0$, $-\gamma Y = -\gamma a - b\,(\gamma X)$ is PH$\left(\boldsymbol{\alpha}, \frac{-1}{\gamma}\mathbf{T}\right)$.
  Then,
  γX is LPH with representation $\left(-\gamma a, -b, \boldsymbol{\alpha} e^{\mathbf{T}a}, \frac{b}{\gamma}\mathbf{T}\right) \equiv \left(-\gamma a, -b, \boldsymbol{\beta}, \frac{1}{\gamma}\mathbf{S}\right)$.

Therefore γX is LPH distributed with representation $\left(|\gamma|\, a, b \cdot sgn(\gamma), \boldsymbol{\beta}, \frac{1}{\gamma}\boldsymbol{S}\right)$.

Result 3 (Density of the LPH class)
The set of LPH distributions is dense in the set of probability distributions defined on any half-line of real numbers.
Proof.
This theorem is proved from the classical result for PH distributions: the set of PH distributions is dense in the set of probability distributions on the non-negative half-line. Let W be a random variable defined on $w > c$ for any real

number c. It is immediate that $W-c$ is defined on the nonnegative half-line. Then, there exists a variable Y, PH distributed, so closed as desirable to $W-c$. Therefore, the variable $X=Y+c$, which is LPH, approximates to the initial variable W.

A similar reasoning can be done for the case when W is defined on $w<c$ for any real number c. In this case $-W+c$ is defined on $\mathbb{R}^{+}$. Then, there exists a variable Y, PH distributed, so closed as desirable to $-W+c$. For this case, the variable $X=-Y+c$, which is LPH, approximates the initial variable W.

3 LPH modelling of functional PCA

Let X be a functional variable whose observed values are curves, and let us assume that $X=\{X\left(t\right):t\in T\}$ is a second order stochastic process, continuous in quadratic mean, whose sample functions belong to the Hilbert space $L^{2}\left(T\right)$ of square integrable functions with the usual inner product $\langle f,g\rangle=\int_{T}f\left(t\right)g\left(t\right)dt,\ \forall f,g\in L^{2}\left(T\right).$

In order to reduce the infinite dimension of a functional variable and to explain its dependence structure by a reduced set of uncorrelated variables, multivariate PCA was extended to the functional case [6]. The functional principal components (p.c.'s) are obtained as uncorrelated generalized linear combinations of the process variables with maximum variance (Var). Then, the $j-th$ p.c. score is given by $\xi_{j}=\int_{T}\left(X\left(t\right)-\mu(t)\right)f_{j}\left(t\right)dt$, where the weight function or loading f_{j} is the value of the argument $f(t)$ that maximizes de objective function with the corresponding constraints

$$\begin{cases} Var\left[\int_{T}\left(X\left(t\right)-\mu(t)\right)f\left(t\right)dt\right] \\ \text{subject to } \|f\|^{2}=1 \text{ and } \int f_{\ell}\left(t\right)f\left(t\right)dt=0,\ \ \ell=1,\ldots,j-1. \end{cases}$$

It can be shown that the weight functions are the eigenfunctions of the covariance operator C. That is, the solutions to the eigenequation $C(f_{j})(t)=\int C\left(t,s\right)f_{j}\left(s\right)ds=\lambda_{j}f_{j}(t)$, where $C\left(t,s\right)$ is the covariance function and $\lambda_{j}=Var[\xi_{j}]$. Then, the process admits the following orthogonal representation (K-L expansion):

$$X\left(t\right)=\mu(t)+\sum_{j=1}^{\infty}\xi_{j}f_{j}\left(t\right),$$

with $\mu(t)$ being the mean function. This principal component decomposition can be truncated providing the best linear approximation of the sample curves in the least squares sense $X^{q}\left(t\right)=\mu(t)+\sum_{j=1}^{q}\xi_{j}f_{j}\left(t\right),$ whose explained variance is given by $\sum_{j=1}^{q}\lambda_{j}$.

There are three main groups of rules for choosing the number of principal components.

The first one consists of ad hoc rules-of-thumb that work very well in practice. The most used chooses a cut-off of total variability, somewhere between $90 - 95\%$, and selects the smallest value of components for which this chosen percentage is exceeded. A graphical procedure, named scree-graph, consists of ploting the number of components against the eigenvalues and retaining the number of components defining an ‘elbow’ in the graph.

The second type of rules is based on formal tests of hypothesis and makes distributional assumptions, as multivariate normality, that are often unrealistic. The Bartlett’s test to decide if the last eigenvalues are equal can be sequentially used to find the number of components that are not noise.

The third group consists of statistically based rules, most of which do not require distributional assumptions, based on computationally intensive methods such as cross-validation and bootstrapping.

A detailed study on principal components selection rules can be seen in Chapter 6 in [8].

The main objective of this work is to model the whole process from the random principal components. Given that PH distributions are dense in the non-negative probability distributions, we show that if the principal components are LPH distributed with the same scale parameter, then the one-dimensional distributions of the process are also LPH.

Corollary 3

Let us assume that each principal component ξ_j is LPH distributed with representation $\left(a_j, b\cdot sgn\left(f_j(t)\right), \boldsymbol{\beta}_j, \mathbf{S}_j\right)$ for a real number t and $j = 1, \ldots, q$. Then, the centered process $X(t) - \mu(t)$ is also LPH distributed with representation

$$\left(\sum_{j=1}^{q} \left|f_j\left(t\right)\right| a_j, b, \boldsymbol{\rho}_j e^{\mathbf{L}_q \sum_{j=1}^{q} |f_j(t)| a_j}, b\mathbf{L}_q\right),$$

with $\boldsymbol{\rho}_q = (\boldsymbol{\alpha}_1, \mathbf{0})$ and

$$\mathbf{L}_q = \begin{pmatrix} \frac{1}{|f_1(t)|}\mathbf{T}_1 & \frac{1}{|f_1(t)|}\mathbf{T}_1^0 \otimes \boldsymbol{\alpha}_2 & & & \\ & \frac{1}{|f_2(t)|}\mathbf{T}_2 & \frac{1}{|f_2(t)|}\mathbf{T}_2^0 \otimes \boldsymbol{\alpha}_3 & & \\ & & \ddots & \ddots & \\ & & & \frac{1}{|f_{q-1}(t)|}\mathbf{T}_{q-1} & \frac{1}{|f_{q-1}(t)|}\mathbf{T}_{q-1}^0 \otimes \boldsymbol{\alpha}_q \\ & & & & \frac{1}{|f_q(t)|}\mathbf{T}_q \end{pmatrix},$$

where $|f_j(t)|$ is the absolute value of $f_j(t)$.

Proof.

From Corollary 2, it is deduced that for a real number t,

if $f_j\left(t\right) > 0$ then $f_j\left(t\right)\xi_j$ is LPH with representation $\left(f_j\left(t\right)a_j, b, \boldsymbol{\beta}_j, \frac{1}{f_j(t)}\mathbf{S}_j\right)$,
if $f_j\left(t\right) < 0$ then $f_j\left(t\right)\xi_j$ is LPH with representation $\left(-f_j\left(t\right)a_j, b, \boldsymbol{\beta}_j, \frac{-1}{f_j(t)}\mathbf{S}_j\right)$.
Then, from Corollary 1, $\sum_{j=1}^{q}\xi_j f_j\left(t\right)$ is also LPH.

4 Application

The devices employed in this paper are composed of a metal-oxide-semiconductor stack whose metal electrode used was copper (200 nm thick), a dielectric 10 nm thick (HfO_2) and a bottom electrode made of /Si-n^+. The resistive memories were fabricated and measured at the Institute of Microelectronics of Barcelona (CNM-CSIC). The variability of these devices is generated by an inherent stochastic process that changes extremely the inner resistance of the device by means of resistive switching physical mechanisms. The experimental data consist of a sample of current-voltage curves corresponding to the reset-set cycles associated with the formation and rupture of a conductive filament that shorts the electrodes and changes drastically the device resistance. From the mathematical viewpoint, the main objective here is to determine the current probability distribution at each voltage in the reset process by means of the K-L expansion and the LPH distributions previously introduced .

In this study, we have 232 reset curves denoted by $\{I_i(v) \;:\; v \in [0, V_{i-reset}], i = 1, \ldots, 232\}$ with $V_{i-reset}$ being the reset voltage. Before applying FPCA to characterize the whole process through the K-L expansion, we must carry out some important previous steps proposed in [3]. Briefly, this approach consists in synchronising all curves in the same interval due to the reset voltage is different for each curve, and using P-spline smoothing to reconstruct all reset curves since we only have discrete observations at a finite set of current values until the voltage reset for each curve. In this paper, the initial domain was transformed in the interval [0,1] and a cubic B-Spline basis of dimension 20 with 17 equally spaced knots and penalty parameter $\lambda = 0.5$ was considered. Figure 1 shows all the smoothed registered curves in the interval [0,1], denoted by $\{I_i^*(u) \;:\; u \in [0,1], i = 1, ..., n\}$, and the estimation of the mean function (red line).

Then, FPCA is estimated and the percentages of variance explained by the first four p.c.'s are 99.42, 0.44, 0.08 and 0.04, respectively. Let us observe that only the first p.c. explains more than 99% of the total variability of the process. Hence, by considering the K-L expansion, principal component decomposition of the registered reset curves can be truncated in the first term as follows: ${I^*}^1(u) = \overline{I}^*(u) + \xi_1^* f_1^*(u),\; u \in [0,1]$. This approach can be used for circuit simulation in

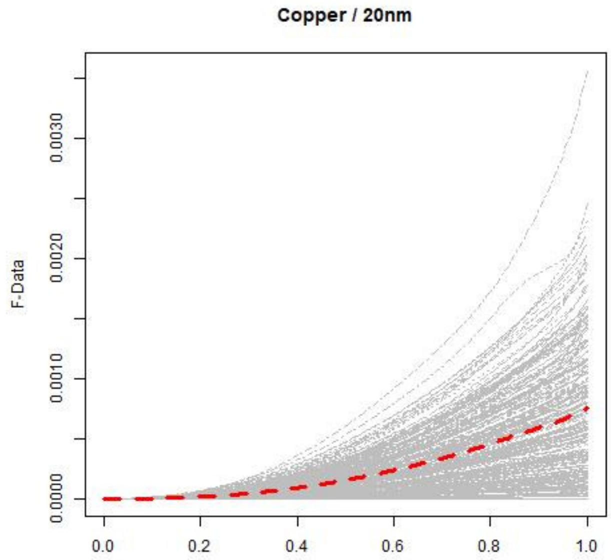


Figure 1: Sample group mean function (dashed red line) and all the P-spline smoothed registered curves.

Distribution	p-value K-S	p-value Anderson-Darling	LogL
PHD	0.11	0.054	18.11
Weibull	0.004	0.004	0.78
Normal	0.02	0.006	7.79
Cauchy	< 0.001	< 0.001	-42.30

Table 1: Comparison among all distributions is considered. P-values of the Kolmogorov-Smirnov and Anderson-Darling tests, and the value of the maximum log-likelihood are showed for each distribution.

this type of devices. Nevertheless, the probability distribution of the first p.c. is unknown.

In order to fit a probability model to the scores of the first p.c. different distributions were employed but none of them could be accepted (p-value associated with Kolmogorov-Smirnov test was < 0.01 in all of them). Then, some transformation is necessary. In this study, the LPH distributions associated with the linear transformation $1 + 1000 \times \xi_1^*$ is considered. The p.c. is multiplied by 1000 because the standard variation of the principal component is very small, with minimum and maximum values of the component equal to $3.2e^{-04}$ and $7e^{-04}$, respectively. These facts produce a great number of phases in the corresponding PH distribution estimated (more than 1000 produced exploding effects). After that, all values were in $[-1, 1]$ and then 1 is added (to consider a PH distribution). Although the constant and slope values could be calculated by maximum likelihood (we are working on it), in this paper they were found *ad hoc,* taking into account that PHD are non-negative variables (the values of the first p.c. are positives and negatives).

The EM algorithm was used for estimating the parameters of a PHD with m transient stages and any internal structure for matrix $\mathbf{T}$ ([4][5]). This methodology has also been applied to estimate the parameters of the PH distributions embedded in the study of the variability in resistive memories. The algorithm is described in [1]. The optimum value was reached for 21 stages. Besides, in order to prove that PHD is better than any other distribution, Weibull, Normal and Cauchy distributions were fitted as well. Their estimation by maximum likelihood are $W(\beta = 4.4344, \lambda = 1.0897)$, $N(\mu = 0.9958, \sigma = 0.234)$ and $C(\gamma = 0.9252, \delta = 0.1505)$, respectively. The results provided by all of them are given and compared in Table 1. Thus, taking into account the logL value and the p-values of the K-S and the Anderson-Darling tests, the best distribution to get an accurate fit of the first p.c. score is the PH distribution. In fact, at 5% significance level, only the PH distribution can be accepted to model the first p.c. score according to the p-values provided by the Kolmogorov-Smirnov and Anderson-Darling tests. This conclusion can be achieved graphically. The cumulative hazard rate (topleft), the density function (topright), the cumulative distribution function (bottomleft) and the reliability function (bottomright) of data with the fitting by means of PH, Weibull, Normal and Cauchy distributions are displayed in Figure 2. In order to sum up, we have proved that the considered linear transformation of the first p.c. is PH distributed with representation $(\boldsymbol{\alpha}, \mathbf{T})$. Therefore, the first p.c. score can be modelled through a LPH distribution with representation $(1, 1000, \boldsymbol{\beta}, \mathbf{S})$. Finally, the reset process ${I^*}^1(u)$ is LPH distributed as well with representation

$$\left(|f_1^*(u)| - 1000\overline{I}^*(u) sgn\left(f_1^*(u)\right), 1000 sgn\left(f_1^*(u)\right), \boldsymbol{\alpha} e^{\boldsymbol{T}\left(1-\frac{1000}{f_1^*(u)}\overline{I}^*(u)\right)}, \frac{1000}{f_1^*(u)}\boldsymbol{T}\right).$$

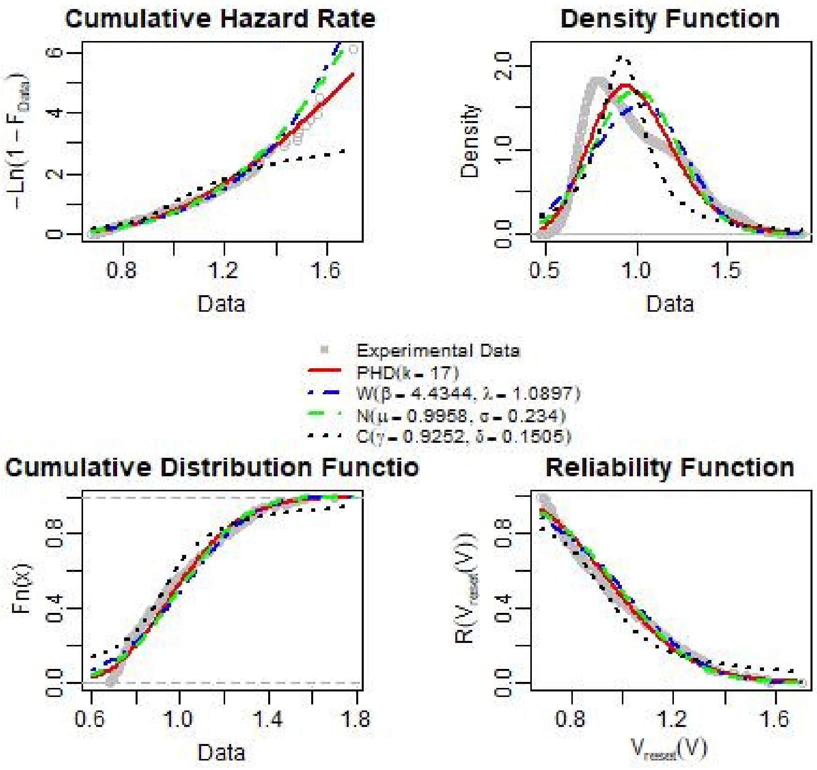


Figure 2: The cumulative hazard rate (topleft), the density function (topright), the cumulative distribution function (bottomleft) and the reliability function (bottomright) of experimental data with the fitting by means of PH, Weibull, Normal and Cauchy distributions.

5 Conclusions

A new probability distribution class with good properties, the LPH class, has been introduced to model the principal components in a matrix and algorithmic form. Multiple properties of this distribution class are developed, including that the LPH class is dense in the probability distribution class defined on any half-line of real numbers. Functional principal components analysis provides a representation of a stochastic process through uncorrelated random variables called principal components. It is of great interest identifying the probability distribution of these components to analyse the random behaviour of the process. In this work, it has also been proved that the process, characterized through the K-L expansion, follows a LPH distribution at each point. The results have been applied to model the stochastic behaviour of resistive memories. In this case, one principal component is considered and the explicit representation of the LPH is given for the stochastic process at each point.

References

[1] C. Acal, J. E. Ruiz-Castro, A.M. Aguilera, F. Jiménez-Molinos, and J.B. Roldán. Phase-type distributions for studying variability in resistive memories. *Journal of Computational and Applied Mathematics*, 345(1):23–32, 2019.

[2] A. M. Aguilera and M. C. Aguilera-Morillo. Penalized PCA approaches for B-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219(14):7805–7819, 2013.

[3] M. C. Aguilera-Morillo, A.M. Aguilera, F. Jiménez-Molinos, and J.B. Roldán. Stochastic modeling of random access memories reset transitions. *Mathematics and Computers in Simulation*, 159(1):197–209, 2019.

[4] S. Asmussen. *Ruin probabilities.* World Scientific, 2000.

[5] P. Buchholz, J. Kriege, and I. Felko. *Input modeling with phase-type distributions and Markov models, Theory and Applications.* Springer, 2014.

[6] J. C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, 15:3–101, 1974.

[7] G. González-Cordero, M.B. González, H. García, F. Campabadal, S. Dueñas, H. Castán, F. Jiménez-Molinos, and J.B. Roldán. A physically based model for resistive memories including a detailed temperature and variability description. *Microelectronic Engineering*, 178(1):26–29, 2017.

[8] I.T. Joliffe. *Principal Component Analysis.* Springer in Statistics. Springer, New York., 2002.

[9] M.F. Neuts. *Matrix geometric solutions in stochastic models. An algorithmic approach, in Probability Distributions of Phase Type.* Baltimore: John Hopkins University Press, 1981.

[10] F Pan, S Gao, C Chen, C Song, and F Zeng. Recent progress in resistive random access memories: materials, switching mechanisms and performance. *Materials Science and Engineering*, 83:1–59, 2014.

[11] E. Pérez, D. Maldonado, C. Acal, J. E. Ruiz-Castro, F. J. Alonso, A. M. Aguilera, F. Jiménez-Molinos, C. Wenger, and J. B. Roldán. Analysis of the statistics of device-to-device and cycle-to-cycle variability in tin/ti/al:hfo2/tin rrams. *Microelectronics Engineering*, 214(1):104–109, 2019.

[12] J. O. Ramsay and B. W. Silverman. *Functional data analysis (Second Edition).* Springer-Verlag, 2005.

[13] J.B. Roldán, F.J. Alonso, A.M. Aguilera, D. Maldonado, and M. Lanza. Time series statistical analysis: a powerful tool to evaluate the variability of resistive switching memories. *Journal of Applied Physics*, 125(1):174504, 2019.

[14] J. E. Ruiz-Castro. A complex multi-state k-out-of-n: g system with preventive maintenance and loss of units. *Reliability Engineering & System Safety*, 197(106797):1–18, 2020.

[15] J. E. Ruiz-Castro and M. Dawabsha. A multi-state warm standby system with preventive maintenance, loss of units and an indeterminate multiple number of repairpersons. *Computers & Industrial Engineering*, 142(106348):1–16, 2020.

[16] J. E. Ruiz-Castro, M. Dawabsha, and F. J. Alonso. Discrete-time markovian arrival processes to model multi-state complex systems with loss of units and an indeterminate variable number of repairpersons. *Reliability Engineering & System Safety*, 174(1):114–127, 2018.

A4 New Modeling Approaches Based on Varimax Rotation of Functional Principal Components

- Acal, Christian; Aguilera, Ana M.; Escabias, Manuel (2020)
- New Modeling Approaches Based on Varimax Rotation of Functional Principal Components
- *Mathematics*, vol. 8, num. 11, pp. 2085
- DOI: `https://doi.org/10.3390/math8112085`

mathematics
an Open Access Journal by MDPI

Mathematics			
JCR Year	Impact Factor	Rank	Quartile
2019	1.747	28/235	Q1

Abstract

Functional Principal Component Analysis (FPCA) is an important dimension reduction technique to interpret the main modes of functional data variation in terms of a small set of uncorrelated variables. The principal components can not always be simply interpreted and rotation is one of the main solutions to improve the interpretation. In this paper, two new functional Varimax rotation approaches are introduced. They are based on the equivalence between FPCA of basis expansion of the sample curves and Principal Component Analysis (PCA) of a transformation of the matrix of basis coefficients. The first approach consists of a rotation of the eigenvectors that preserves the orthogonality between the eigenfunctions but the rotated principal component scores are not uncorrelated. The second approach is based on rotation of the loadings of the standardized principal component scores that provides uncorrelated rotated scores but non-orthogonal eigenfunctions. A simulation study and an application with data from the curves of infections by COVID-19 pandemic in Spain are developed to study the performance of these methods by comparing the results with other existing approaches.

1 Introduction

Nowadays, the great advancement of technology makes it common to have high-dimensional data associated with a large number of highly correlated variables. Functional data is a type of high-dimensional data in which a large number of observations of one or more variables are available at a continuous argument, usually time, on a sample of individuals. Therefore, a sample of functional data is a set of functions (curves, surfaces, etc.) that vary in a continuous argument such as time. Examples of data of this type are given in very diverse areas such as life sciences, environment, economics, chemometrics and electronic, among others. Functional Data Analysis (FDA) deals with the statistical modeling of this type of data. A detailed study of the main FDA methodologies as well as relevant applications and computational aspects are described in the books by [26, 25, 24, 10, 14].

The most common FDA technique is Functional Principal Component Analysis (FPCA) introduced by [9] as a generalization of the reduction dimension multivariate technique PCA to the case in which the data are functions instead of vectors. The first papers on this topic were framed in the theory of second order stochastic processes with the Karhunen–Loève (KL) expansion being the main tool. Thanks to this probabilistic result, the sample functions are reconstructed in terms of a small set of uncorrelated variables called principal components, whose interpretation allows to explain the main modes of variation in the functional data set. The theoretical aspects related with the properties, asymptotic theory and inference results of FPCA in the general framework of Hilbertian random functions were deeply studied in [8, 23, 12].

Most of the functional data can not be observed directly so that the latent stochastic process of interest must be reconstructed from discrete observations of each sample curve on a fixed or random time grid, which can be dense or sparse and different for the sample individuals. One usual form of reconstructing the functional form of sample curves is by an expansion in terms of basis functions such as Fourier, B-splines or wavelets [3, 4, 6, 1, 5, 19]. The equivalence between FPCA of basis expansion of functional data and certain multivariate PCA in terms of the basis coefficients data matrix was studied in [23]. On the other hand, different Bayesian approaches to FPCA were considered in [31, 30]. In addition, nonparametric methods to perform functional principal components analysis for the case of irregularly spaced longitudinal data (sparse) were developed [16, 20].

The problem inherent to many applications is that interpreting the components is not always straightforward. It is known that the greatest contribution in the structure of a functional principal component is given by the process variables associated with the greatest values of the corresponding weight curve at certain time points [11]. In some cases the principal components are difficult to interpret because the estimated weight functions have a lot of variability and lack of smoothness. One way to solve this problem is based on penalizing the roughness of the weight functions. Several penalized FPCA approaches were developed to improve the estimation of the principal weight functions in the case of smooth curves observed with error [28, 7, 2]. In other cases, the first principal component explains a very high percentage of the total variance and is a straightforward average or size effect. These problems are usually solved by a rotation of the weight functions that simplifies the component structure and therefore makes the interpretation easier. The main drawback of rotation is that it is not able to retain the two crucial properties of FPCA: uncorrelatedness of the components and orthogonality of the weight functions. The most popular rotation method is Varimax [17]. This criterion has been extended to FPCA in two different way: the first one is based on Varimax rotation of the matrix of basis coefficients of the weight functions, and the other one is based on Varimax rotation of the matrix of values of the weight functions in a grid of equally spaced time points [26]. Varimax criterion could be unhelpful when data have a strong seasonal behaviour leading to a periodic structure as well as trends and isolated features in the weight curves. This is because Varimax rotation does not take into account the dependence structure in functional data at nearby time points. In order to solve this problem, a functional factor rotation based on canonical correlation was introduced in [18] as a means of extracting nearly-periodic directions in the data (principal periodic components). In this paper, two new approaches for rotation of FPCA are introduced. Both are based on the equivalence between FPCA and multivariate PCA of certain transformation of the matrix of basis coefficients of the sample curves [22]. On the one hand, Varimax rotation of the eigenvectors provides orthonormal rotated eigenfunctions but the associated principal components are not uncorrelated anymore. On the other hand, Varimax rotation of the loadings associated with the standardized principal components yields uncorrelated components with non-orthogonal eigenfunctions.

After this introduction, theoretical aspects related with the Varimax functional rotation are developed in Section 2. The behaviour of the proposed rotation methodologies is tested on a simulation study in Section 3, where the results are compared with other functional Varimax approaches previously developed in the literature. An application on COVID-19 infection curves is developed in Section 4. Finally, a detailed discussion of the results is given in Section 5.

2 Rotation in Functional Principal Component Analysis

Let us begin by a brief summary on Varimax rotation of multivariate PCA before introducing the functional Varimax rotation approaches.

2.1 Rotation in PCA

The rotation of principal components has its origin in the Factor Analysis (FA) whose goal is to find out the dependence structure among several variables by expressing them in terms of a small number of non-observable latent variables called factors. The aim of rotation of the matrix of factor loadings (multiplication by an orthogonal matrix R) is to facilitate the interpretation so that each factor is associated with a small block of observed variables. That means that the columns of the rotated loading matrix have high values for several variables and low for the remainder (the most elements either close to zero or far from zero, and with as few as possible values taking intermediate values). This approach gives raise to different criteria for defining the type of rotation which is designed to simplify the structure of loadings. Varimax, quartimax and promax are the most usual orthogonal methods meanwhile oblimax provides oblique factors by allowing R to be not necessarily orthogonal. The contributions of this paper are based on Varimax criterion which is the most applied in practice thanks to its good interpretation results. This type of rotation can be extended to PCA in order to simplify the structure of the problem and to facilitate the interpretation.

Formally, let X be a data matrix associated with a sample of size n of p random variables $(X_1, \ldots, X_p)$. Let us suppose without loss of generality that the variables are centered. PCA can be applied by means of Singular Value Decomposition (SVD), that is, $X = UDV^T$ where U is a $(n \times p)$ unitary matrix, D is a $(p \times p)$ diagonal matrix whose principal diagonal is formed by the singular values and V is a $(p \times p)$ orthogonal matrix whose columns are the eigenvectors of the covariance matrix of X given by $\Sigma_{p\times p} = X^T X/(n-1) = V\Lambda V^T$, with Λ being a diagonal matrix whose elements are the eigenvalues of Σ. Then, the following principal component representation is obtained:

$$X = UD \times V^T = ZV^T,$$

where $Z = UD$ are the principal components (PCs) scores and the columns of matrix V are also called principal directions or axes of the PCA. It is well known that the eigenvectors associated with different PCs are orthogonal ($V^TV = I$) and that all the p unrotated components are uncorrelated $Z^TZ/(n-1) = \Lambda$. On the other hand, the standardized PC scores (uncorrelated scores with unit variance) denoted by $\tilde{Z}$ are given by $\tilde{Z} = Z\Lambda^{-1/2} = ZD^{-1}\sqrt{n-1} = U\sqrt{n-1}$, so that the data matrix is expressed as $X = \tilde{Z}\Delta^T$, with $\Delta = VD/\sqrt{n-1}$ being the loadings associated with the standardized PCs which are eigenvectors scaled by the corresponding singular values.

There are two different ways to perform the rotation that provide different interpretation results. Thus, by considering the first $q < p$ p.c's, X can be approximated by means of SVD as $X^q = U_qD_qV_q^T$ and the orthogonal rotation matrix R can be inserted through the following two possibilities:

1. $X^q = (U_qD_qR)(R^TV_q^T) = Z_q^R {V^T}_q^R$.
2. $X^q = (U_qR)(R^TD_qV_q^T) = \tilde{Z}_q^R {\Delta^T}_q^R$.

One is based on rotating the loadings of PCs (eigenvectors) and the other in rotating the loadings of the standardized PCs (eigenvectors scaled by the singular values). In the first option the new scores provided by the rotation will not be uncorrelated anymore although the axes do will remain orthogonal. This is not how PCA is usually understood and applied. For that reason, it is quite common not to call them anymore rotated PCs but only rotated components. By contrast, in the second option the rotated loadings are not orthogonal axes but the rotated scores continue to be uncorrelated. Any of these approaches can be considered but in order to interpret the results it is important to take these properties into account. In fact, and according to our research, even the experts in this field do not reach an agreement about what method is better or what approach must be considered more often in practice. Therefore, it seems reasonable to conclude that there is not an ideal method for rotating the PCs and any of them can be employed. Another important aspect has to do with the amount of variance explained by the rotated components. After applying the Varimax rotation, the variance explained by the first q components remains unchanged and gets redistributed among the rotated components so that the quantities are not arranged in descending order.

Let us remember that in Varimax rotation the matrix R is computed by maximizing the variance of the coefficients that define the effect of each factor on the observed variables. Then, in PCA R is chosen to maximize the variability of the squares elements of the rotated matrix of eigenvectors/loadings. In any case, the amount of explained variance by each rotated component is determined by the following formula:

$$VT_k^R = \frac{\delta_k}{\sum_{k=1}^{q} \delta_k} \times VT_q,$$

where δ_k is the kth value of the diagonal of $Z^{R^T} Z^R$ and VT_q is the proportion of total variance captured by the first q PCs. Let us observe that the criterion of rotating the loadings provides the same proportion of variance explains by each one of the rotated standardized components. This fact is due to the properties of the matrix U from the SVD analysis.

2.2 Rotation in Functional PCA

For many reasons, FPCA is the basic tool in FDA. It is an extension of PCA which is crucial to reduce the infinite dimension of functional data and to explain the variability and dependence structure of functional variables in terms of a reduce set of uncorrelated variables called functional PCs [9].

Let $\{x_i(t) : t \in T, i = 1, \ldots, n\}$ be a size n sample of curves associated with a second order and quadratic mean functional variable X defined on a probabilistic space $(\Omega, \mathcal{A}, P)$, whose sample curves belong to the space $L^2(T)$ of square integrable functions on a real interval T, with the natural inner product defined as

$$\langle f, g \rangle = \int_T f(t) g(t)\ dt\ , \qquad \forall f, g \in L^2[T].$$

Let us also assume without loss of generality that the functional variable X is centered.

The principal components are uncorrelated generalized linear combinations with maximum variance (Var). In general, the j-th principal component score is given by

$$z_{ij} = \int_T x_i\,(t)\, f_j\,(t)\, dt,\ i = 1, \ldots, n,$$

where the weight function (loading) f_j is obtained by maximizing the variance

$$\begin{cases} Max_f \;\; Var\left[\int_T x_i\,(t)\, f\,(t)\, dt\right] \\ r.t.\ \|f\|^2 = 1 \text{ and } \int f_\ell\,(t)\, f\,(t)\, dt = 0,\;\; \ell = 1, \ldots, j-1. \end{cases}$$

This problem is solved in term of the eigenanalysis of the sample covariance operator C. That is, the solutions to the second order integral equation

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

where $c(t,s)$ is the sample covariance function and $\lambda_j = Var[z_j]$. Then, the following principal component decomposition of the sample curves is obtained: $x_i(t) = \sum_{i=1}^{n-1} z_{ij} f_j(t)$, that can be truncated in the qth term providing the best least squares linear approximation of the sample curves $x_i^q(t) = \sum_{i=1}^{q} z_{ij} f_j(t)$, with explained variance given by $\sum_{i=1}^{q} \lambda_i$. The most usual criterion for choosing the number of PCs consist of selecting the first q components whose proportion of explained variance is close to one (at least 0.75–0.8 in most cases).

In order to estimate the eigenvalues and eigenvectors, it is usual to assume that sample paths belong to a finite-dimension space generated by a basis $\{\phi_1(t), ..., \phi_p(t)\}$, so they can be expressed as

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t) = a_i'\Phi(t),\ i = 1, ..., n,$$

where p must be sufficiently large to get an accurate representation of the curves. The selection of the type and dimension of the basis is a crucial problem that must be solved by keeping in mind the characteristics of the curves. Normally, Fourier basis is used when the curves are periodic, B-spline basis is employed for non-periodic smooth paths and wavelet basis for data with a strong local behaviour. Once the basis is selected, the basis coefficients are commonly approximated by least squares from noisy discrete time observations of each sample curve.

In this context, FPCA is equivalent to multivariate PCA of matrix $A\Psi^{1/2}$, with $A = (a_{ij})_{n\times p}$ being the matrix of basis coefficients and $\Psi^{1/2}$ being the squared root of the matrix of inner products between basis functions $\Psi = (\Psi_{ij})_{p\times p} = \int_T \phi_i(t)\phi_j(t)dt,\ i, j = 1, ..., p$ [22]. Then, the PC weight functions admit the following basis expansion:

$$f_j(t) = \sum_{k=1}^{p} b_{jk}\phi_k(t),$$

where the vector b_j of basis coefficients is given $b_j = \Psi^{-1/2}v_j$ where the v_j are computed as the eigenvectors of the sample covariance matrix of $A\Psi^{1/2}$. Then, $Z = (z_{ij})_{n\times p} = (A\Psi^{1/2})V$ is the matrix whose columns are the PC scores of $A\Psi^{1/2}$ and V the one whose columns are its associated eigenvectors. In matrix form, the basis expansion of weight functions would be $f = B^T\Phi$, with $f = (f_1, \ldots, f_p)^T$ being the vector with the eigenfunctions, B the matrix of basis coefficients $B_{p\times p} = (b_{ij}) = \Psi_{p\times p}^{-1/2} V_{p\times p}$, V the matrix with columns the eigenvectors of the covariace matrix of $A_{n\times p}\Psi_{p\times p}^{1/2}$, and $\Phi = (\phi_1, \ldots, \phi_p)^T$, the vector of basis functions.

2.2.1 Functional Varimax Rotation

Two different ways of functional varimax rotation were proposed so far [26]. One is based on rotating the matrix of basic coefficients of the eigenfunctions and the other, coarser, on rotating the matrix of values of the eigenfunctions in a grid of equally spaced time points. In both cases the rotated component scores are no longer uncorrelated although the weight functions (axes) after rotation are still orthonormal. At this point, the new methodology that we propose for rotating the functional PCs consists of rotating PCA of the matrix $A\Psi^{1/2}$, based on the statement that FPCA is equivalent to multivariate PCA of this matrix. This is the main contribution of the current study in addition to doing an exhaustive revision about different ways of functional Varimax rotation and a comparison study among them. As a natural extension of the multivariate case, our proposal considers two different possibilities depending whether the rotation is done on the eigenvectors or on the loadings of the standardized principal component scores. This way, the rotation of the functional principal components is inspired by the theory of rotation of factor analysis presented in previous subsection by considering the multivariate viewpoint in the FDA context.

More formally, FPCA rotation would consists of rotating the first q PC weight functions as $f_q^{R^T} = f_q^T R$. This way, the vector $n \times 1$ with the sample functions is approximated in terms of the first q PCs as

$$X^q = Z_q f_q = (Z_q R)(R^T f_q) = Z_q^R f_q^R,$$

where the vector of rotated eigenfunctions is expressed as $f_q^{R^T} = \Phi^T B_q R = \Phi^T(\Psi^{-1/2} V_q) R$ with B_q being the matrix of basic coefficients associated with the first q eigenfunctions and V_q the matrix whose columns are the first q eigenvectors. This expression was our inspiration to propose a methodology based on directly rotating the eigenvectors instead of the methodology based on rotating the basic coefficients proposed by [26].

Thus, the chances in order to rotate functional PCA are the following:

R1 Applying the VARIMAX rotation criterion to weight function values.

In this case, the purpose would be to find a matrix R that maximizes the variance of the squares of the elements of the matrix

$$F_q^{R^T} = F_q^T R,$$

where F_q is the $q \times m$ matrix whose elements are the values of the first q eigenfunctions evaluated at a grid of time points $t_1, ..., t_m$, given by $F_q^T = \Gamma^T \Psi^{-1/2} V_q$, with Γ being the $p \times m$ matrix that contains as rows the values of each basis function at the time points.

R2 Applying the VARIMAX rotation criterion to weight function coefficients.

In this occasion, the goal is to calculate a matrix R that maximizes the variability of the squares elements of $B^R = BR = \Psi^{-1/2}VR$. Then, the rotated principal factors are given by

$$f_q^{R^T} = \phi^T B^R.$$

R3 Applying the VARIMAX rotation criterion to PCs by rotating the matrix of eigenvectors.

Here, the objective is to determine a matrix R that maximizes the variability of the squares elements of the rotated matrix of eigenvectors $V_q^R = V_qR$. Then, the rotated principal factors are given by

$$f_q^{R^T} = \phi^T(\Psi^{-1/2}V^R).$$

R4 Applying the VARIMAX rotation criterion to the standardized PCs by rotating the matrix of loadings

Hence, this method consists of computing a matrix R that maximizes the variance of the squares elements of the matrix $\Delta_q^R = \Delta_q R = V_q\Lambda_q^{1/2}R$. Then, the rotated principal factors are given by

$$f_q^{R^T} = \phi^T\left(\Psi^{-1/2}\Delta_q^R\Lambda_q^{-1/2}\right).$$

The two last functional Varimax approaches (R3 and R4) are the main contribution of this paper based on Varimax rotation of the multivariate PCA of $A\Psi^{1/2}$ matrix, which is equivalent to functional PCA of X. On the other hand, methods R1 and R2 are not new and are considered in this paper only for comparison purpose in the simulation study. Let us observe that in the case of orthonormal basis functions, approaches R2 and R3 match. Moreover, with the first three methods the rotated factors are ortonormal but the rotated components are not uncorrelated, meanwhile with the last one the opposite happens.

3 Simulation Study

The good performance of the two functional Varimax approaches introduced in this paper (R3 and R4) is tested on simulated data. The results will be compared with the ones provided by approaches R1 and R2 discussed in the book by [26].

The data are simulated from the approximation of the Wiener process (Brownian motion) given by its Karhunen–Loève (KL) expansion truncated in the qth

term. This is a Gaussian process with covariance function given by $C(t, s) = \sigma^2 min(t, s)$. The KL expansion of this process is given as follows in terms of the eigenvalues and eigenfunctions of the covariance operator:

$$X(t) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k f_k(t), \tag{1}$$

where the PCs ξ_k are independent Gaussian random variables with mean zero and variance one, the eigenvalues are given by $\lambda_k = \frac{\sigma^2}{(k-0.5)^2\pi^2}$ and the eigenfunctions by $f_k(t) = \sqrt{2}\text{sin}((k - 0.5)\pi t)$. In this study, the cut-off $q = 8$ and a dispersion parameter $\sigma = 0.2$ were considered. Then, 500 samples of 150 sample curves of the process $X(t)$ given by Equation (1) were simulated at different number of equally spaced knots in the observed domain [0, 1]. Three different scenarios were considered by defining the time points as $t_k = k/m, k = 0, 1, \ldots, m; m = 25, 50, 100$. Different sample sizes were also considered but the results are not included in the paper because they were quite similar for sample sizes large enough.

First, least squares approximation of each sample curve was performed in terms of a basis of cubic B-splines of dimension 8. The sample curves of one of the simulated samples are displayed in Figure 1. Then, functional PCA and the four considered functional Varimax approaches for rotating the first four components were performed. Table 1 shows an example of the amount of variance explained by the first four PCs and the redistribution of the variances after applying the three type of rotation of the eigenfunctions aforementioned. Let us observe that the criterion of rotating the loadings (R4) is not included in this table because the same proportion of variance is distributed among the rotated standardized components (24.48%). This fact is due to the properties of the matrix U from the SVD analysis.

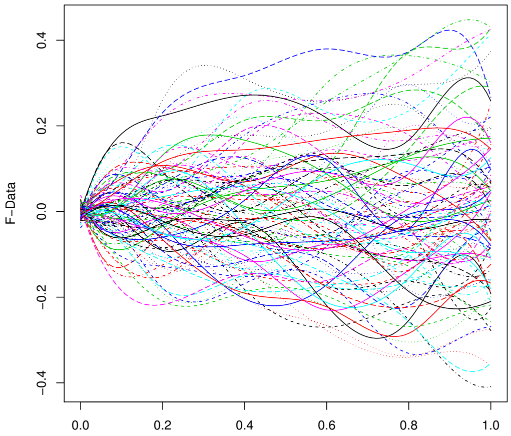


Figure 1: Sample of 150 simulated sample curves of the KL expansion of the Wiener process truncated in the fourth term.

Table 1: Percentages of variance explained by the first four PCs and their redistribution after the three types of Varimax rotation of the eigenfunctions.

PC	**FPCA**	**R1**	**R2**	**R3**
1	80.4	23.2	22.0	23.5
2	10.8	11.6	49.5	9.3
3	4.5	24.9	7.6	44.2
4	2.2	38.2	18.8	20.9

In Figure 2, the estimated eigenfunctions (FPCA) and their functional Varimax rotations by the four considered approaches (R1, R2, R3 and R4) are displayed for one of the simulated samples next to the original rotation of the theoretic values for the first four eigenfunctions. Theoretically, the rotated eigenfunctions with the first three approaches should resemble their corresponding original rotation. In order to draw general conclusions, the integrated mean squares error (MSE) of each rotated eigenfunction with respect to the original rotation is computed as the squared root of

$$\|f_i^R - \hat{f}_i^R\|^2 = \int_T \left[f_i^R(t) - \hat{f}_i^R(t)\right]^2 dt = \int_T \left[\sum_{j=1}^{p} d_{ij}\phi_j(t)\right]^2 dt = d_i'\Psi d_i,$$

where $d_i = (d_{i1}, \ldots, d_{ip})'$ is the vector with the differences between the basis coefficients of each original rotated eigenfunction and the ones of its estimation by using the different type of functional rotations. The boxplots of the MSEs for the rotated eigenfunctions estimated by using R1, R2 and R3 with 26, 51 and 101 observed time points for 500 simulations of the Wiener process were plotted in Figure 3. Rotation R4 is included in these boxplots although the estimated eigenfunctions are not orthogonal and the comparison with the other approaches makes no sense. Let us observe that the new Varimax rotation of the eigenfunctions introduced in this paper (R3) provides the most accurate results, which are also more robust with respect to the number of observation nodes of the sample curves.

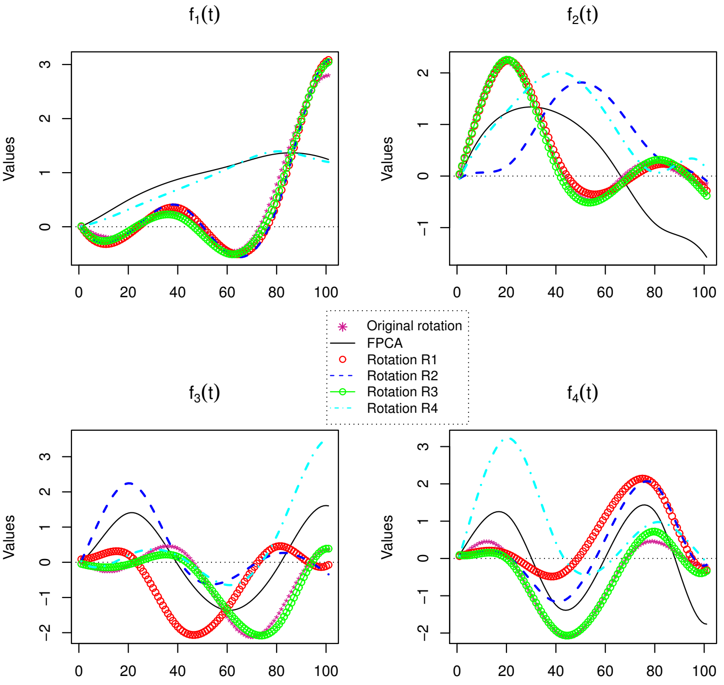


Figure 2: Eigenfunctions after applying FPCA analysis and the four type of rotation explained.

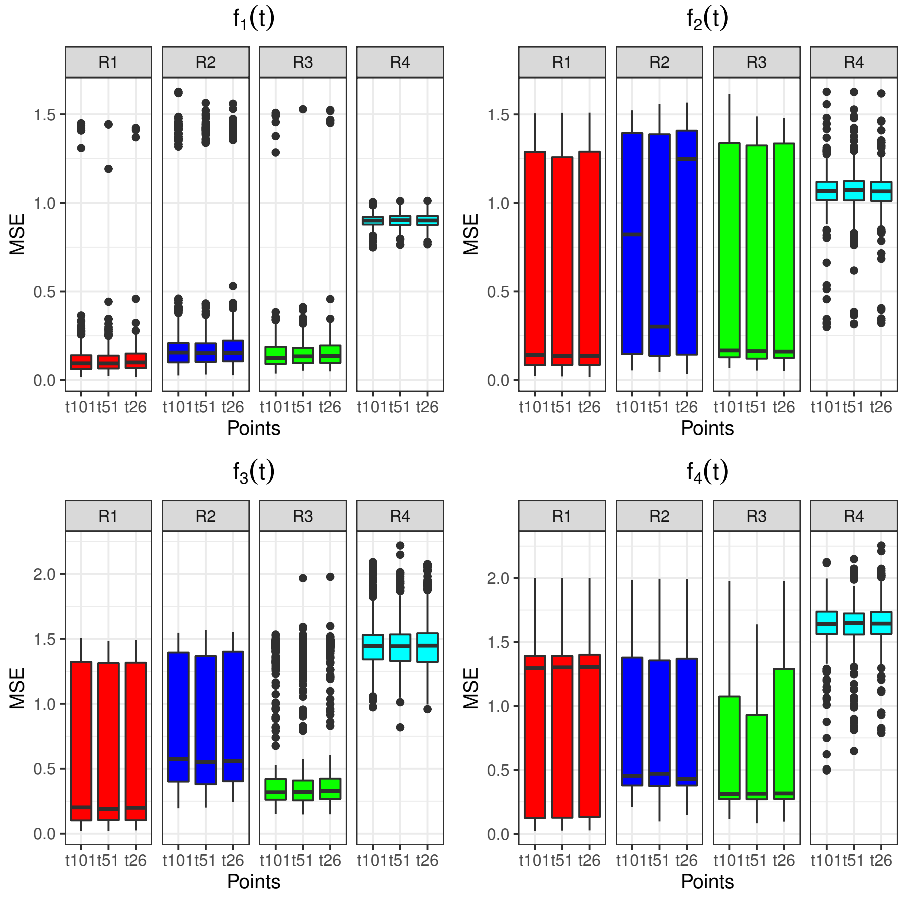


Figure 3: Box plots for the integrated MSEs of the rotated eigenfunctions estimated by using R1, R2, R3 and R4 rotation approaches with 26, 51 and 101 basis knots on 500 simulations of the Wiener process.

4 COVID-19 Data

In order to show up the usefulness of rotation to facilitate the interpretation of the principal components, an application with data from COVID-19 pandemic has been developed. The functional data are the number of daily cumulative informed cases of COVID-19 for seventeen autonomous communities (ACs) in Spain from 20/02/2020 to 27/04/2020 (first wave of COVID-19). Data source: .

The sample curves, denoted by $x_1(t), \ldots, x_{17}(t)$, are daily observed starting the day that at least one case is reported. Therefore, the period of observation and the number of observations are different for each AC. In order to homogenize the data, the number of cases per 10,000 inhabitants is considered and the first observation for each curve corresponds to the day that exceeds by first time the maximum of the first reported values. Then, all the curves were registered in the common interval [0, 1]. A detailed description of basis approaches for functional data registration can be seen in [26].

The first step for estimating FPCA is to approximate the sample curves in terms of an appropriate functional basis by using least squares smoothing. A B-spline basis of dimension 10 with equally spaced knots in the interval [0, 1] was chosen in this paper for the functional representation of each curve. Figure 4 shows all the smoothed sample curves.

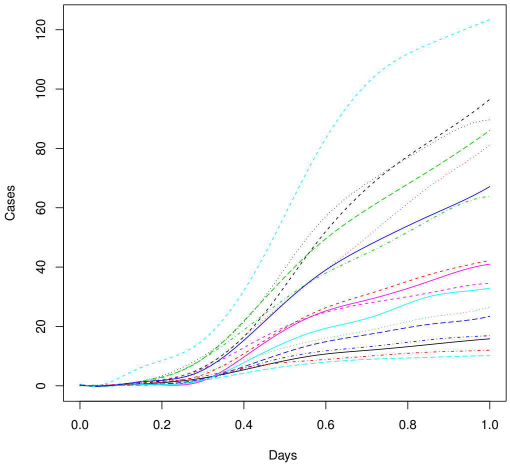


Figure 4: B-spline smoothing of the number of daily cumulative informed cases by COVID-19 per 10,000 inhabitants for seventeen autonomous communities in Spain.

Second, FPCA was performed in order to reduce the dimension of the problem and to explain the different modes of variability in the data. As the first principal component explains more than 99% of the total variability the results are not easy to interpret (Table 2). The estimated first four weight functions are displayed in Figure 5 (black line). Let us observe that the first eigenfunction is positive and strictly increasing through the entire observation period, and in addition, the weight placed on the cases at the end is about two times higher than at the beginning. This could lead to interpret that the most important mode of variation between ACs represents a quick increase in cases as time passed with the infection curve out of control. The rest of the components are difficult to interpret since they account for much smaller and insignificant proportions of the total variation.

Table 2: Percentages of variance explained by the first four PCs of COVID-19 data per 10,000 inhabitants for seventeen autonomous communities in Spain.

PC	**FPCA**	**Rotation R3**
1	99.32	44.36
2	0.52	38.14
3	0.12	0.67
4	0.03	14.82

Third, in order to obtain weight functions and PC scores much easier to interpret, the two Varimax rotation approaches introduced in this paper (R3 and R4) are carried out on the first four PCs. This way, the variability explained by the first four rotated components is divided in different proportions, which can be seen in Table 2. Let us now observe that the first two rotated components explain more than a 82% of the total variability with the main mode of variation accounting a 44% and the second a 38% approximately. The first four rotated eigenfuntions are shown in Figure 5. Taking into account their explained variances, only the first two rotated components will be interpreted. The first two eigenfunctions plotted as positive and negative perturbations of the mean function are shown in Figure 6 with the first row corresponding to the rotation of eigenvectors (R3) and the second one to the rotation of loadings (R4 approach). The scores of the seventeen Spanish ACs on the first two rotated principal components of COVID-19 cases are displayed in Figure 7 for R3 (left) and R4 (right) rotation approaches, where the location of each AC is shown by the abbreviation of its name assigned in Table 3.

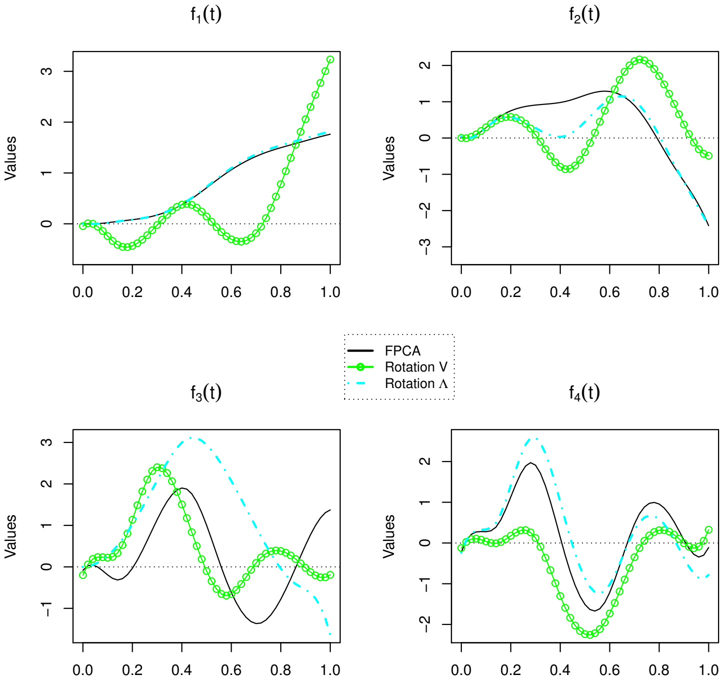


Figure 5: The first four principal component weight curves for COVID-19 data (eigenfunctions in solid black line) and the rotated eigenfunctions after applying the Varimax rotation criterion to the matrix of eigenvectors (R3 approach in dotted green line) and to the loadings (R4 approach in dashed cyan line).

Let us begin by interpreting the results given by R3 approach (rotation of eigenvectors). Now, the first eigenfunction is easier to interpret and represents those ACs that had an increase more or less constant until the 70% of the observed period where the number of cases shot up leaving the curve out of control. The three highest scores are assigned to La Rioja (RI), Madrid (MD) and Castilla la Mancha (CM), which were the communities with more problems controlling the infections and the largest negative scores to Canarias (CN), Murcia (MC) and Andalucía (AN), which were the communities that better controlled the infection curve. On the other hand, the behaviour of the second eigefunction represents those ACs which suffered an increase relatively rapid between the 40% and 70% of the period but they managed to have the curve under control from that moment.

Regarding R4 approach (rotation of loadings), the behaviour of the first and second eigenfunctions is very similar to the unrotated ones. That is, the first is associated with those ACs that did not control the curve because as the days passed, the number of cases increased very quickly. On the other hand, the second eigenfunction could be influenced by the ACs which controlled the number of cases since the time representing the 60% of the observed period. These conclusions are corroborated by Figures 5 and 6. Les us observe from Figure 7 (left) the high correlation between the first two rotated PCs scores provided by approach R3 that establishes two clearly differentiated groups between the autonomous communities: those ACs which managed to control moderately the curve of number of cases (third quadrant) and the ones that lost control of the cases by reaching numbers really concerning (first quadrant). On the other hand, thanks to the uncorrelation between the rotated PCs scores, approach R4 provides a much better clustering of AC. This can be seen in the biplot on the right in Figure 7 where each of these two groups is divided in other two so that four groups can be clearly distinguished.

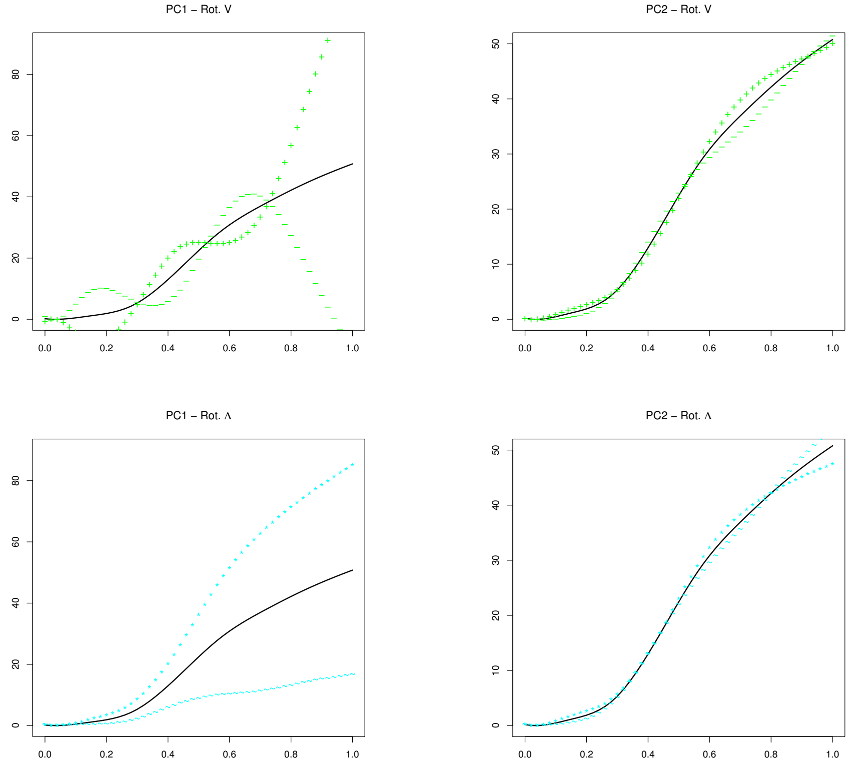


Figure 6: The mean curve of COVID-19 cases and the effects of adding (+) and subtracting (−) a suitable multiple of each PC weight curve (eigenfunction). The first row corresponds to the rotation of eigenvectors (R3) and the second one to the rotation of loadings (R4 approach).

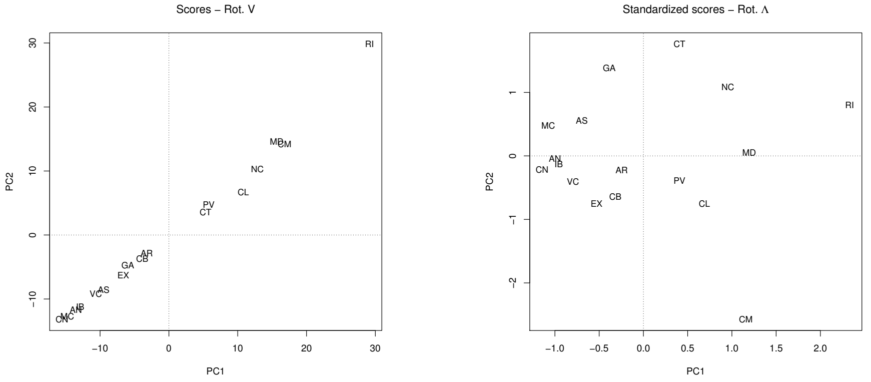


Figure 7: The scores of the seventeen Spanish autonomous communities on the first two rotated principal components of COVID-19 cases. The location of each AC is shown by the abbreviation of its name assigned in Table 3.

Table 3: Abbreviation of the seventeen Spanish autonomous communities.

Andalucía, AN	Castilla-La Mancha, CM	Madrid, MD
Aragón, AR	Castilla-León, CL	Murcia, MC
Asturias, AS	Cataluña, CT	Navarra, NC
Islas Baleares, IB	Comunidad Valenciana, VC	País Vasco, PV
Canarias, CN	Extremadura, EX	Rioja, RI
Cantabria, CB	Galicia, GA	

In fact, these conclusions agree with the results obtained after applying functional data clustering [15]. In particular, it has been considered the approach based on performing clustering using the basis expansion coefficients in terms of the basis of cubic B-splines aforementioned. Due to the fact that La Rioja (RI) could be an outlier, the K-medoids method, which is more robust than K-means, is applied next to Manhattan distance as similitude measure. Moreover, as the dataset is not too large, the algorithm called Partitioning Around Medoids is considered. In order to identify the optimum number of clusters, the reduction of intra-cluster total variance was evaluated for a range of values K (elbow method). It can be seen in the left panel of Figure 8 that the reduction seems to stabilize by starting at 4 cluster. Finally, the clustering results appear in the right panel of Figure 8 which is very similar to the biplot in the right panel of Figure 7. This is in accordance with multiple studies about the infections by

COVID-19 pandemic in Spain [13, 21, 27, 29], what corroborate the good interpretation and classification results provided by the new rotation approaches introduced in this paper.

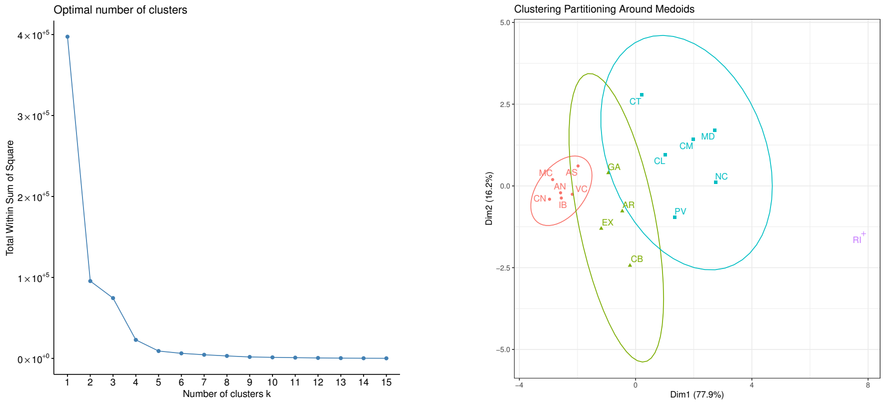


Figure 8: Scores of the number of cumulative informed cases by COVID-19 per 10,000 inhaibtans of seventeen autonomous community of Spain.

5 Discussion

FDA try to solve problems where the involved sample data are functions that vary over some continuum, usually time. One of the most important techniques in the field of FDA is Functional Principal Component Analysis, whose main purpose is to reduce the dimension of the problem and to explain the dependence structure of data in terms of a reduce set of uncorrelated variables called functional principal components. The interpretation of these components helps to understand the main characteristics and modes of variation of the underline stochastic process. Nevertheless, there are many situations in which this task is not easy. One is the case when the first PC represents a size effect that explains a very high percent of the total variability. The most common tool to solve this problem in PCA is Varimax rotation that redistributes the explained variance among all rotated components to make easier the interpretation. So far, there were only two approaches available in the literature to apply Varimax rotation in the FDA context, but neither of them is a direct rotation of eigenfunctions. The first one consists of rotating the values of the weight functions evaluated at the time points (R1), while the second one is based on rotating the weight function coefficients (R2). Both methods retain the orthogonality of the axis but the new scores will not be uncorrelated anymore. In this paper, two new approaches based on the equivalence between FPCA of basis expansion of the sample curves and PCA of a transformation of the matrix of basis coefficients are proposed: one is based on applying the Varimax criterion to principal components by rotating the matrix of eigenvectors (R3), and the other makes use of the Varimax criterion on the standardized principal components by rotating the matrix of loadings (R4). The first one guarantees the orthogonality of the rotated eigenfunctions and in the second one the rotated scores are still uncorrelated. Moreover, all of them are compared in an exhaustive simulation study. From this study it can be concluded that R3 provides the most accurate rotated eigenfunctions and is also more robust with respect to the number of discrete time observations of the sample curves. Finally, an application with the curves of infections by COVID-19 pandemic in Spain has been developed. Through the combination of these two new varimax approaches (R3 and R4), it has been possible to distinguish different behaviors in the evolution of infections in the Spanish autonomous communities during the first wave of the pandemic. These results are in agreement with other studies done in the country about this matter [13, 21, 27, 29]. These Varimax FPCA approaches are expected to be welcomed and highly employed in future researches in different areas of science thanks to their ability to facilitate the interpretation of the main patterns of variation in the data.

References

[1] A. M. Aguilera and M. C. Aguilera-Morillo. Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling*, 58(7-8):1568–1579, 2013.

[2] A. M. Aguilera and M. C. Aguilera-Morillo. Penalized PCA approaches for B-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219(14):7805–7819, 2013.

[3] A. M. Aguilera, R. Gutiérrez, F. A. Ocaña, and M. J. Valderrama. Computational approaches to estimation in the principal component analysis of a stochastic process. *Applied Stochastic Models and Data Analysis*, 11(4):279–299, 1995.

[4] A. M. Aguilera, R. Gutiérrez, and M. J. Valderrama. Approximation of estimators in the PCA of a stochastic proces using B-splines. *Communications in Statistics. Simulation and Computation*, 25(3):671–690, 1996.

[5] M. C. Aguilera-Morillo, A.M. Aguilera, F. Jiménez-Molinos, and J.B. Roldán. Stochastic modeling of random access memories reset transitions. *Mathematics and Computers in Simulation*, 159(1):197–209, 2019.

[6] P. Besse and J. O. Ramsay. Principal component analysis of sample functions. *Psychometrika*, 51(2):285–311, 1986.

[7] H. Cardot. Nonparametric estimation of the smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538, 2000.

[8] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–156, 1982.

[9] J. C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, 15:3–101, 1974.

[10] F. Ferraty and P. Vieu. *Nonparametric functional data analysis. Theory and practice.* Springer-Verlag, 2006.

[11] T. Górecki, M. Krzyśko, L. Waszak, and W. Wołyński. Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, 59:153–182, 2018.

[12] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society B*, 68(1):109–126, 2006.

[13] J. Henríquez, E. Gonzalo-Almorox, M. García-Goñi, and F. Paolucci. The first months of the covid-19 pandemic in spain. *Health Policy and Technology, in press*, 2020.

[14] L. Horvath and P. Kokoszka. *Inference for functional data with applications.* Springer-Verlag, 2012.

[15] J. Jacques and C. Preda. Functional data clustering: a survey. *Advanced Data Analysis and Classification*, 8:231–255, 2014.

[16] G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000.

[17] I.T. Jolliffe. *Principal Component Analysis (Second Edition).* Springer, 2002.

[18] C. Liu, S. Ray, G. Hooker, and M. Friedl. Functional factor analysis for periodic remote sensing data. *The Annals of Applied Statistics*, 6(2):601–624, 2012.

[19] J. Liu, J. Chen, and D. Wang. Wavelet functional principal component analysis for batch process monitoring. *Chemometrics and intelligence laboratory systems*, 196, 2020.

[20] H. G. Müller and J. L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005.

[21] P. Muñoz-Cacho1, J.L. Hernández, M. López-Hoyos, and V.M. Martínez-Taboada. Can climatic factors explain the differences in covid-19 incidence and severity across the spanish regions?: An ecological study. *Environmental Health, in press*, 2020.

[22] F. A. Ocaña, A. M. Aguilera, and M. Escabias. Computational considerations in functional principal component analysis. *Computational Statistics*, 22(3):449–465, 2007.

[23] F. A. Ocaña, A. M. Aguilera, and M. J. Valderrama. Functional Principal Components Analysis by Choice of Norm. *Journal of Multivariate Analysis*, 71(2):262–276, 1999.

[24] J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer-Verlag, 2009.

[25] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: Methods and case studies.* Springer-Verlag, 2002.

[26] J. O. Ramsay and B. W. Silverman. *Functional data analysis (Second Edition).* Springer-Verlag, 2005.

[27] L. Santamaría and J. Hortal. Covid-19 effective reproduction number dropped during spain's nationwide dropdown, then spiked at lower-incidence regions. *Science of the Total Environment*, 751:142257, 2021.

[28] B. W. Silverman. Smoothed functional principal component analysis by choice of norm. *Annal Statistics*, 24(1):1–24, 1996.

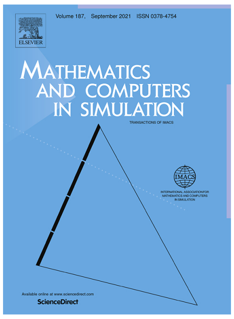
[29] C. Siqueira, Y. Leite de Freitas, M. Cancela, M. Carvalho, A. Oliveras-Fabregas, and D. Bezerra de Souza. The effect of lockdown on the outcomes of covid-19 in spain: An ecological study. *Plos One, in press*, 2020.

[30] A.J. Suárez and S. Ghosal. Bayesian estimation of principal components for functional data. *Bayesian Analysis*, 12(2):311–333, 2017.

[31] A. Van der Linde. Variational bayesian functional PCA. *Computational Statistics and Data Analysis*, 53(2):517–533, 2008.

A5 Homogeneity problem for basis expansion of functional data with applications to resistive memories

- Aguilera, Ana M.; Acal, Christian; Aguilera-Morillo, M. Carmen; Jimenez-Molinos, Francisco; Roldan, Juan B. (2021)
- Homogeneity problem for basis expansion of functional data with applications to resistive memories
- *Mathematics and Computers in Simulation*, vol. 186, pp. 41-51
- DOI: `https://doi.org/10.1016/j.matcom.2020.05.018`


Volume 187, September 2021 ISSN 0378-4754
ELSEVIER
MATHEMATICS AND COMPUTERS IN SIMULATION
TRANSACTIONS OF IMACS
IMACS
INTERNATIONAL ASSOCIATION FOR MATHEMATICS AND COMPUTERS IN SIMULATION
Available online at www.sciencedirect.com
ScienceDirect

Mathematics, Applied			
JCR Year	Impact Factor	Rank	Quartile
2019	1.620	68/261	Q2

Abstract

The homogeneity problem for testing if more than two different samples come from the same population is considered for the case of functional data. The methodological results are motivated by the study of homogeneity of electronic devices fabricated by different materials and active layer thicknesses. In the case of normality distribution of the stochastic processes associated with each sample, this problem is known as Functional ANOVA problem and is reduced to test the equality of the mean group functions (FANOVA). The problem is that the current/voltage curves associated with Resistive Random Access Memories (RRAM) are not generated by a Gaussian process so that a different approach is necessary for testing homogeneity. To solve this problem two different parametric and nonparametric approaches based on basis expansion of the sample curves are proposed. The first consists of testing multivariate homogeneity tests on a vector of basis coefficients of the sample curves. The second is based on dimension reduction by using functional principal component analysis of the sample curves (FPCA) and testing multivariate homogeneity on a vector of principal components scores. Different approximation numerical techniques are employed to adapt the experimental data for the statistical study. An extensive simulation study is developed for analyzing the performance of both approaches in the parametric and non-parametric cases. Finally, the proposed methodologies are applied on three samples of experimental reset curves measured in three different RRAM technologies.

1 Introduction

The methodological results in this paper linked to homogeneity tests for functional data are motivated by the study of variability in Resistive Random Access Memories. In this work, the devices under study were fabricated making use of different materials for the metal electrodes and dielectrics of different thicknesses. RRAMs are currently considered a serious contender for non-volatile memory applications. These devices operate under the principles of resistive switching (RS), i.e., their internal resistance is switched between different values by changing the nature and features of charge conduction within a dielectric layer. Many different developments are being considered in the research of these devices such as fabrication and characterization; also, the simulation and modeling facets are under study for this emerging technology [9, 20].

The use of devices based on RS for cryptographic applications is based on the inherent stochasticity of their operation. The device resistive state changes in many cases because of the creation (set) or destruction (reset) of a conductive filament that is formed by the random movement of ions in a dielectric. The result of this randomness is a sample of current-voltage curves corresponding to reset-set cycles with variability. So, the variability turns into different voltages and currents within the set and reset processes for each cycle. The analysis of the statistics of the RS operation is essential to understand the devices underlying physics [1, 18, 13]. It is necessary to theoretically investigate the stochastic characteristics of RRAMs (directly related to variability), from both the mathematical point of view and the compact modeling perspective. The variability characterization will be essential to develop the infrastructure for device and circuit design software tools.

As the experimental data set connected to each device is a group of current/-voltage curves associated with the reset/set cycles of the device, functional data analysis (FDA) methodologies could be the ideal tool to explain the associated variability. Nowadays, FDA is a leading research topic in statistics in which the methods developed for samples of vectors are becoming extended to the case of samples of curves. Besides, the interest in methodological developments as well as in applications to fields such as life science, chemometrics, environment, economy, electronics, among others, are growing continuously. A good review of the main FDA methods, interesting applications and computational algorithms with the free software R can be seen in the books [15, 16, 14]. The sample curves are usually observed at a finite set of discrete points so that the first step in FDA is usually the smoothing of each sample curve through its representation as a linear combination of basis functions. A comparison of different types of penalized smoothing with B-splines basis was performed in [3].

The basic tool in FDA is functional principal component analysis (FPCA) that reduces the dimension of the stochastic process generating the sample curves by providing a small set of uncorrelated scalar variables that represent the most important variation modes in the sample. Different penalized PCA approaches for B-spline expansions of smooth functional data were introduced in [2]. FPCA was recently applied to model the variability of the reset processes associated with RRAM devices [4]. In the FDA context, the problem in the present work consists of testing homogeneity for several independent samples of experimental data obtained from different RRAMs. The aim is to characterize the device variability by considering different metals as electrode materials and dielectrics of different thicknesses in the fabrication process. The homogeneity problem addressed in this contribution consists of deciding if several independent samples of curves have been generated by the same stochastic process (homogeneity), so that they have equal probability distributions. In order to solve this problem, different parametric and non-parametric approaches based on basis expansion of the sample curves are proposed here.

In the case of normality distribution of the stochastic processes associated with each sample, the homogeneity problem is known as the multi-sample problem or one-way ANOVA problem for functional data, and it is equivalent to equality of the mean functions among the different samples (FANOVA). A detailed description and comparison of tests for the one-way ANOVA problem for functional data can be seen in [10, 21]. Taking into account the basis expansion of the sample curves, the FANOVA is reduced to a multivariate ANOVA (MANOVA) with the vector of basis coefficients of the sample curves as dependent variable and the categorical variable representing the groups as independent variable. The problem is that the current/voltage curves associated with RRAMs are not generated by a Gaussian process so that a different approach is necessary for testing homogeneity. Multivariate non-parametric homogeneity tests [12] on the vector of basis coefficients are considered in this paper to solve the problem. Other important problem is that multivariate homogeneity tests do not perform well with high-dimensional vectors and the number of basis functions needed for an accurate approximation of the sample curves is usually high. In order to solve it, a new approach based on dimension reduction by using FPCA of the sample curves and testing homogeneity on the vector of the most explicative principal components scores is introduced.

Apart from this introduction, the manuscript scheme consists of a theoretical development of functional homogeneity test procedures adapted to the data measured for the devices under study (Section 2), a simulation study to evaluate the performance of the testing approaches in Section 3, an application with data from resistive memories and the corresponding discussion in Section 4, and finally, the main conclusions in Section 5.

2 Statistical homogeneity tests for basis expansion of functional data

Let $\{x_{ij}(t) : i = 1, \ldots, m; j = 1, \ldots, n_i; t \in T\}$ denote m independent samples (groups) of curves defined on a continuous interval T. Let us assume that they are realizations of i.i.d. stochastic processes (functional variables) $\{X_{ij}(t) : i = 1, \ldots, m; j = 1, \ldots, n_i t \in T\}$ with distribution $SP(\mu_i(t), \gamma_i(s,t)), \forall i = 1, ..., m$, with $\mu_i(t)$ being the mean function and $\gamma_i(s,t)$ the covariance function associated with each of the m stochastic processes. Let us also assume that all sample curves belong to the Hilbert space $\mathrm{L}^2[T]$ of the square integrable functions on T, with the natural inner product defined by

$$< f|g >= \int_T f(t)g(t)dt \quad \text{for all} \quad f, g \in L^2[T].$$

The homogeneity of the m samples of curves means that they have been generated by the same stochastic process $SP(\mu(t), \gamma(s,t))$ with the same probability distribution $\forall i = 1, ..., m$. This problem has been recently considered from different points of view. If the processes are Gaussian, then the problem is known as the multi-sample problem or one-way ANOVA problem for functional data (see the book [21] for a detailed study). A comprehensive comparison of tests for the one-way ANOVA problem for functional data was developed in [10]. More recently, an approach based on the concept of functional depth measures was introduced in [7]. In this paper we focus on basis expansion of functional data and propose two different type of approaches. One consists on testing multivariate homogeneity on the random vector of basis coefficients for the m groups, and the other is based on testing multivariate homogeneity on the associated functional principal components (p.c.'s).

Then, the starting point is to assume that the sample curves belong to a finite-dimension space spanned by a basis $\{\phi_1(t), \ldots, \phi_p(t)\}$, so that each stochastic process is represented by its vector of basis coefficients. Let us assume that

$$X_{ij}(t) = \sum_{k=1}^{p} a_{ijk}\phi_k(t) \quad i = 1, \ldots, m; j = 1, \ldots, n_i, \tag{1}$$

where a_{ijk} are scalar random variables with finite variance and p is sufficiently large to assure an accurate representation of each process. In vector form $X_{ij}(t) = \mathbf{a}_{ij}'\Phi(t)$ with $\mathbf{a}_{ij} = (a_{ij1}, ..., a_{ijp})'$ being the vectors of basis coefficients and $\Phi(t) = (\phi_1(t), ..., \phi_p(t))'$. On the one hand, the selection of the type and dimension of the basis (Fourier, B-splines, wavelets, polinomials, etc) is an important problem that must be solved by taking into account the sample curve characteristics. In the application in this paper a base of cubic splines is chosen because the analysed current/voltage curves are smooth enough. Other useful basis systems are Fourier functions for periodic data, piecewise constant functions for counting processes or wavelets bases for curves with strong local behavior. On the other hand, the basis coefficients are usually estimated by least squares (with or without penalization) from discrete-time noisy observations. A good review about different ways to proceed and how to do it with the software R can be studied in the books [16, 14].

2.1 Homogeneity testing on basis coefficients

The first type considered approach consists of performing a multivariate homogeneity test on the m samples of the basis coefficient vector $\{\mathbf{a}_{ij} : i = 1, \ldots, m; j = 1, \ldots, n_i\}$.

When the processes are Gaussian, the one-way ANOVA problem for functional data is equivalent to equality of the mean functions among the different samples provided that the covariance functions in the groups are equal (homoscedastic case) or different (heteroscedastic case). This problem can be formulated as the hypothesis test of equality of the unknown group mean functions of the m samples

$$H_0 : \mu_1(t) = \cdots = \mu_m(t), \forall t \in T, \tag{2}$$

against the alternative that its negation holds.

In the case of the one-way FANOVA problem (2) the functional data verify the following linear model:

$$X_{ij}(t) = \mu(t) + \alpha_i(t) + \epsilon_{ij}(t),\ i = 1, ..., m,\ j = 1, ..., n_i, \tag{3}$$

where $\mu(t)$ is the overall mean function, $\alpha_i(t)$ is the *i-th* main-effect function, and $\epsilon_{ij}(t)$ are the subject-effect functions (i.i.d. errors) with distribution $SP(0, \gamma(s,t))$ $\forall i = 1, \ldots, m; j = 1, \ldots, n_i$, and $\gamma(s,t)$ being the common covariance function in the homoscedastic case.

The main-effect functions are not identifiable so that in order to be estimated some constraint must be imposed. The most used constraint is $\sum_{i=1}^{m} \alpha_i(t) = 0$. Under this constraint you have that $\mu_i(t) = \mu(t) + \alpha_i(t)$. Then, by assuming the basis expansion in 1, the unbiased estimators of the functional parameters in model 3 are given by

- $\hat{\mu}(t) = \overline{x}(t) = \overline{\mathbf{a}}'\Phi(t)$,
- $\hat{\alpha}_i(t) = \bar{x}_i(t) - \bar{x}(t) = (\overline{\mathbf{a}}'_i - \overline{\mathbf{a}}')\,\Phi(t)$,
- $\hat{\epsilon}_{ij}(t) = x_{ij}(t) - \overline{x}_i(t) = \left(\mathbf{a}'_{ij} - \overline{\mathbf{a}}'_i\right)\Phi(t)$,

where $\bar{x}$ and $\bar{x}_i(t)$ are the usual unbiased estimators of the grand mean function and the group mean functions, respectively, and, $\bar{\mathbf{a}}$ and $\bar{\mathbf{a}}_i$ are the corresponding unbiased estimators of the grand mean vector and the group mean vector associated with the coefficient vectors $\mathbf{a}_{ij}$.

Taking into account the basis expansions of the sample curves the FANOVA testing problem is equivalent to the usual multivariate ANOVA test (MANOVA) for the matrix of basis coefficients $A = \left(a_{(ij)k}\right)_{n \times p}$, with $n = \sum_{i=1}^{m} n_i$. This is equivalent to test the equality of mean vectors for the basis coefficients in the m groups. This problem is solved by using one of the well known MANOVA tests: the Wilks's lambda, the Lawley-Hotelling's trace, the Pillai's trace, and the Roy's maximum root. In most cases, the exact null distributions of these four test criteria can not be computed, and approximate F-tests statistics are often used in computer programs. A detailed explanation on these tests can be seen in [17].

The main requirements for estimating a one-way MANOVA model are: 1) observations are randomly and independently sampled from the population; 2) the sample size in each group must be larger than the number of dependent variables; 3) dependent variables are multivariate normally distributed within each group; 4) homogeneity of variance-covariance matrices in the m groups; and 5) no multicollinearity.

When the m samples of vectors coefficients are not Gaussian, the F-type tests described earlier can not be applicable. Other approaches based on bootstrap versions of these tests methods may be considered [10]. In this paper nonparametric multivariate homogeneity tests based are considered to solve this problem. Specifically, the extensions of the univariate Kruskal Wallis's test and Moods's test [12, 6] that try to check whether the medians are equal in all groups, are applied in the simulation and the application developed in Section 3 and 4, respectively. The Moods's test is less sensitive with the outliers than the Kruskall Wallis's test but it is less powerful when the data are generated from some distributions as, for instance, the normal distribution.

Let us finally observe that unbalanced sample sizes can lead to unequal variances between samples that could affect to the statistical power and type I error rates of parametric (ANOVA type) tests [19]. In fact, equal sample sizes maximize statistical power. On the other hand, nonparametric rank-based tests could lead to paradox results due the non-centralities of the test statistics which may be non-zero for the traditional tests in unbalanced designs. A simple solution is the use of pseudo-ranks instead of ranks [5].

2.2 Homogeneity testing on functional principal components

This new approach for solving the homogeneity problem with functional data consists of reducing the infinite dimension of the stochastic procecess by using FPCA and then performing a multivariate homogeneity test on the vectors of the most explicative principal components scores.

FPCA provides the following orthogonal decomposition of the process (Karhunen-Loève expansion):

$$X_{ij}(t) = \mu(t) + \sum_{k=1}^{\infty} f_k(t)\xi_{ijk}, \tag{4}$$

where $\{f_k\}$ are the orthonormal eigenfunctions of the covariance operator associated with its decreasing sequence of non null eigenvalues $\{\lambda_k\}$, and $\{\xi_k\}$ are uncorrelated zero-mean random variables (principal components) defined by

$$\xi_{ijk} = \int_T f_k(t)(X_{ij}(t) - \mu(t))dt.$$

The *k-th* p.c. ξ_k has the maximum variance λ_k out of all the generalized linear combinations of the functional variable which are uncorrelated with ξ_l $(l = 1, .., k-1)$.

By truncating the expression (4), the process admits a principal component reconstruction in terms of the first q principal components so that the sum of their explained variances is as close as possible to one. Then, in vector form the functional variable $X(t)$ is approximated by $X_{ij}^q(t) - \mu(t) = \xi_{ij}'\mathbf{f}(t)$, with $\xi_{ij} = (\xi_{ij1}, ..., \xi_{ijq})'$ being the vectors of principal components scores and $\mathbf{f(t)} = (\mathbf{f}_1(t), ..., \mathbf{f}_q(t))'$.

In practice, and assuming the basis expansion of sample curves given in 1, the functional PCA is equivalent to multivariate PCA of matrix $A\Psi^{\frac{1}{2}}$ [11], with $\Psi^{\frac{1}{2}}$ being the squared root of the matrix of inner products between basis functions $\Psi = (\Psi_{ij})_{p\times p} = \int_T \phi_i(t)\,\phi_j(t)\,du$. Then, the principal component weight function $\hat{f}_k$ admits the basis expansion $\hat{f}_k(t) = \mathbf{b}_k'\Phi(t)$, so that, the vector $\mathbf{b}_k$ of basis coefficients is given by $\mathbf{b}_k = \Psi^{-\frac{1}{2}}\mathbf{u}_k$, where the vectors $\mathbf{u}_k$ are computed as the solutions to the eigenvalue problem $n^{-1}\Psi^{\frac{1}{2}}A'A\Psi^{\frac{1}{2}}u_k = \lambda_k u_k$, where $n^{-1}\Psi^{\frac{1}{2}}A'A\Psi^{\frac{1}{2}}$ is the sample covariance matrix of $A\Psi^{\frac{1}{2}}$.

Again, we propose two different ways to solve the problem of homogeneity of the vector of the first q principal components in the m groups. In the case of multivariate normality of the vector of principal components scores, a MANOVA testing procedure based on the F-type statistics is not advisable because the dependent variables are uncorrelated. In this case, we propose to perform univariate ANOVA on each p.c. score that has more power than MANOVA analysis. In order to control the Type I error when conducting these multiple ANOVA tests, the additive Bonferroni inequality will be applied so that the alpha level for each ANOVA test is given by the overall level divided by the number of tests. On the other hand, if normality is not verified, then non-parametric multivariate tests will be applied.

3 Simulation study

In this section, an extensive simulation study with artificial data is developed to check the performance of the two functional homogeneity approaches: one is based on testing homogeneity on the basis coefficients and the other on testing homogeneity on the principal components.

In this study, three groups have been considered (m=3) with the following three different models for the mean functions:

- $M1 : \mu_i(t) = 0.1|\sin(4\pi t)| \quad i = 1, 2, 3,$
- $M2 : \mu_i(t) = 0.05i|\sin(4\pi t)| \quad i = 1, 2, 3,$
- $M3 : \mu_i(t) = 0.025i|\sin(4\pi t)| \quad i = 1, 2, 3.$

Let us observe that $M1$ corresponds to situations where H_0 is true while $M2$ and $M3$ corresponds to situations where H_0 is false. In $M3$ the differences between the means are smaller so that the testing problem is more difficult.

In addition, two different type of error functions are added to simulate a sample of functional data in the interval [0,1] for each case according to the model in Equation 3. For the parametric approaches (Gaussian case), an approximation of the standard Wiener process given by its Karhunen-Loève expansion truncated in the qth term is used. This is a Gaussian process with covariance function given by $C(t,s) = \sigma^2 \min(t,s)$. The Karhunen-Loeve expansion of this process is given as follows in terms of the eigenvalues and eigenfunctions of its covariance operator: $\epsilon(t) = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_k f_k(t)$, where the p.c.'s ξ_k are independent Gaussian random variables with mean zero and variance one, the eigenvalues are given by $\lambda_k = \frac{\sigma^2}{(k-\frac{1}{2})^2\pi^2}$, with the associated eigenfunctions $f_k(t) = \sqrt{2}\sin\left(\left(k-\frac{1}{2}\right)\pi t\right)$. The truncation point in this study is $q = 20$, and five different values for the dispersion parameter ($\sigma = 0.02, \sigma = 0.05, \sigma = 0.10, \sigma = 0.20, \sigma = 0.40$) are considered. For the non-parametric approaches, the error functions are computed in the same form as the exponential, adequately centered, of $\epsilon(t)$ (log-normal distribution).

Then, i.i.d. samples, with three different sample sizes ($n_i = 15, n_i = 25, n_i = 35; i = 1, 2, 3$), are simulated at 51 equally spaced time points in the interval $[0, 1]$ for each one of the thirty considered functional models. Finally, 1000 Monte Carlo replications are developed for each one of the ninety considered cases (three mean models*two type of error *five dispersion parameters*three sample sizes). In order to obtain the basis coefficients for each sample curve from its discretized values in the interval [0,1], least squares approximation in terms of a basis of cubic B-splines of dimension 18 was used in all cases. All the computations were obtained with the packages *fda* [14] and *npmv* [6] of statistical software *R.* As indicator of the test performance, the observed acceptance proportions at a significance level 0.05 under every considered model were computed. Three different number of p.c.'s were considered for the principal component approach: the first three p.c.'s, the first five p.c.'s and the first eight p.c.'s that explain approximately a 95%, a 97%, and a 99% of the total variability, respectively. The results for the two testing parametric approaches with the F-type tests (Gaussian errors) appear in Table 1. MANOVA testing with the Pillai statistics was conducted for the basis coefficients approach and multiple univariate ANOVA for the principal component approach, using Bonferroni's inequality for preserving the overall significance level. On the other hand, the results for the non-parametric approaches (Log-normal errors) appear in Table 2. The multivariate extension of the Kruskall-Wallis univariate test was used to compute the p-values.

Next, a discussion of the simulation experiment is presented that can help to show the practical utility of the proposed methodology:

1. An important key point to keep in mind is the dispersion parameter σ. It seems that the testing performance depends strongly on the error dispersion, getting worse as σ increases in all cases. In fact, when $\sigma = 0.40$ the power of the tests is too small, especially in the case of the model M3 in which the differences between the group means are smaller. This must be taken into account for future analysis because previous simulation studies of this type (see [10]) do not consider a value of σ higher than 0.20.

2. Another interesting point has to do with sample sizes. For small values of

σ the sample size does not have an important effect in the power of the test. However, the sample size plays a fundamental role when σ increases, as the test converts into less conservative for the cases M2 and M3.

3. Regarding to the number of p.c.'s selected for the testing procedure, it can be seen in Tables 1 and 2 that the greater the number of p.c.'s, the better results the tests achieve. In fact, the tests don't behave well in the situations where the variability explained is lower than 99% and the term σ is large. So, it would be recommendable to consider a number of p.c.'s that guarantees around the 99% of the variability.

4. For the model M1 (H_0 is true), both the parametric and the non-parametric tests provide excellent results. The acceptance proportions are greater than 0.938 in all the cases.

5. In the case of model M2, the results obtained in the parametric case with the basis coefficients and with eight p.c.'s are really good, even when the dispersion is very high ($\sigma = 0.4$). Only some problems are detected when the sample size is small in this case. Nevertheless, the tests provide slightly better results for the basis coefficients approach. On the other hand, the outputs for M2 when we consider the non-parametric tests change a bit in comparison with the previous situation. Now, the basis coefficients model does not work very well when the sample is not large enough for $\sigma = 0.20$ and for $\sigma = 0.40$, with the acceptance proportion being 0.169 and 0.706, respectively. Instead, if we consider the approach with 8 p.c.'s the results are much better, only having controversy when $n_i = 15$ and $\sigma = 0.40$ just like in the parametric case.

6. The behavior of results with model M3 are very similar to the case of model M2, basis coefficients approach is slightly better in the parametric case but it occurs the opposite in the non-parametric case. In addition, the tables bring to light the lack of power of both tests (parametric and non-parametric) when the differences among group means are small and σ is large. We are rather concerned with the frequency of a correct decision in these situations.

To sum up, we can firstly conclude that the parametric tests are more powerful than the non-parametric ones and, for that reason, they must be considered when the conditions of validity are satisfied. Another important aspect to keep in mind is the reduction of the dimension provided by the principal component approach. This approach can be really interesting when the number of dependent variables (basis coefficients) is large and the problem is reduced to testing homogeneity on a small number of p.c's. Based on the results of this simulation study, it can be concluded that the principal component approach explaining a 99% of variability gives better results than the basis coefficients approach in the non-parametric case. Regarding to the parametric case, the fact of using the Bonferroni's inequality for correcting the significance level in the multiple ANOVA tests on the p.c.'s, could be the reason for a slight decrease in the power with respect to the basis coefficient approach in this study where 8 p.c.'s are needed to explain at least a 99% of the total variability. This problem disappears in practice when only two or three components are necessary so that the corrected level of significance is not so small and the acceptance proportion would increase.

4 Application results and discussion

In this paper, we will use experimental data measured at the Institute of Microelectronics of Barcelona (CNM-CSIC) where the devices were also fabricated. The devices are based on a metal-oxide-semiconductor stack [8]. The metal electrodes employed were Ni and Cu, the dielectrics (HfO_2) and Si-n^+ was employed as bottom electrode. In particular, the following devices were used: Device 1 (DV1): Ni/HfO_2 (20 nm thick)/Si-n^+, Device 2 (DV2) Ni/HfO_2(10 nm thick)/Si-n^+, Device 3 (DV3): Cu/HfO_2(20 nm thick)/Si-n^+. The I–V characteristics were measured using a HP-4155B semiconductor parameter analyzer. A negative voltage was employed although we used the absolute value for easiness in the numerical analysis. The functional homogeneity approaches presented here will be applied to decide about the existence of significant statistical differences between the three devices considered.

More precisely, RRAM operation is based on the stochastic nature of resistive switching processes; these, in the most cases, create and rupture conductive filaments that change drastically the resistance of the device. These processes are known as set and reset, respectively. Moreover, the resistance change gives rise to a sample of current-voltage curves corresponding to the reset-set cycles, where the mentioned variability is translated to different voltages and currents related to set and reset processes for each cycle. See the set and reset curves in Figure 1 of reference [4] and the variation of the set and reset voltages, where the current drastically increases or drops off.

In this study, we have information about 2782 reset curves corresponding to the device with the nickel electrode and the dielectric 20 nanometres thick (Device 1), 1742 reset curves for the device with nickel electrode and a dielectric 10 nanometres thick (Device 2) and 233 reset curves for devices with a copper electrode and a dielectric 20 nanometres thick (Device 3), denoted as $\{I_{ij}(v) : v \in [0, V_{ij-reset}]\}$ being $i = 1, 2, 3$ the type of device and $j = 1, ..., n_i$ the sample

Mean	n_i	Model	σ=0.02	σ=0.05	σ=0.10	σ=0.20	σ=0.40
M1	15	Basis coef.	0.949	0.956	0.951	0.957	0.944
		3 p.c.'s	0.938	0.954	0.968	0.959	0.949
		5 p.c.'s	0.948	0.944	0.941	0.945	0.958
		8 p.c.'s	0.958	0.955	0.946	0.946	0.942
	25	Basis coef.	0.952	0.953	0.953	0.964	0.951
		3 p.c.'s	0.952	0.942	0.953	0.948	0.959
		5 p.c.'s	0.954	0.962	0.945	0.948	0.950
		8 p.c.'s	0.962	0.947	0.951	0.954	0.956
	35	Basis coef.	0.949	0.953	0.944	0.948	0.946
		3 p.c.'s	0.951	0.956	0.951	0.952	0.959
		5 p.c.'s	0.949	0.953	0.954	0.953	0.948
		8 p.c.'s	0.951	0.955	0.960	0.943	0.957
M2	15	Basis coef.	0	0	0	0	0.170
		3 p.c.'s	0	0	0.001	0.448	0.852
		5 p.c.'s	0	0	0	0.043	0.681
		8 p.c.'s	0	0	0	0.003	0.298
	25	Basis coef.	0	0	0	0	0.005
		3 p.c.'s	0	0	0	0.265	0.789
		5 p.c.'s	0	0	0	0.006	0.595
		8 p.c.'s	0	0	0	0	0.042
	35	Basis coef.	0	0	0	0	0
		3 p.c.'s	0	0	0	0.157	0.746
		5 p.c.'s	0	0	0	0.001	0.552
		8 p.c.'s	0	0	0	0	0.007
M3	15	Basis coef.	0	0	0	0.181	0.783
		3 p.c.'s	0	0	0.465	0.879	0.937
		5 p.c.'s	0	0	0.043	0.724	0.914
		8 p.c.'s	0	0	0.004	0.307	0.752
	25	Basis coef.	0	0	0	0.005	0.537
		3 p.c.'s	0	0	0.286	0.802	0.925
		5 p.c.'s	0	0	0.009	0.607	0.879
		8 p.c.'s	0	0	0	0.044	0.627
	35	Basis coef.	0	0	0	0	0.330
		3 p.c.'s	0	0	0.178	0.724	0.903
		5 p.c.'s	0	0	0	0.485	0.880
		8 p.c.'s	0	0	0	0.004	0.466

Table 1: Observed acceptance proportions for each scenario at a significance level 0.05 in the case of Gaussian errors.

Mean	n_i	Model	σ=0.02	σ=0.05	σ=0.10	σ=0.20	σ=0.40
M1	15	Basis coef.	0.984	0.983	0.975	0.982	0.982
		3 p.c.'s	0.960	0.951	0.957	0.946	0.960
		5 p.c.'s	0.967	0.965	0.950	0.950	0.951
		8 p.c.'s	0.966	0.972	0.952	0.959	0.970
	25	Basis coef.	0.966	0.962	0.978	0.964	0.972
		3 p.c.'s	0.951	0.943	0.945	0.964	0.930
		5 p.c.'s	0.951	0.958	0.955	0.961	0.953
		8 p.c.'s	0.943	0.950	0.955	0.963	0.959
	35	Basis coef.	0.960	0.959	0.971	0.967	0.967
		3 p.c.'s	0.947	0.950	0.941	0.949	0.956
		5 p.c.'s	0.954	0.949	0.958	0.957	0.950
		8 p.c.'s	0.960	0.964	0.952	0.957	0.949
M2	15	Basis coef.	0.005	0.009	0.011	0.169	0.706
		3 p.c.'s	0	0	0.001	0.525	0.894
		5 p.c.'s	0	0	0	0.041	0.793
		8 p.c.'s	0	0	0	0	0.365
	25	Basis coef.	0	0	0	0	0.136
		3 p.c.'s	0	0	0	0.309	0.865
		5 p.c.'s	0	0	0	0	0.663
		8 p.c.'s	0	0	0	0	0.046
	35	Basis coef.	0	0	0	0	0.010
		3 p.c.'s	0	0	0	0.18	0.795
		5 p.c.'s	0	0	0	0	0.568
		8 p.c.'s	0	0	0	0	0.006
M3	15	Basis coef.	0.007	0.025	0.142	0.697	0.940
		3 p.c.'s	0	0.001	0.484	0.890	0.933
		5 p.c.'s	0	0	0.044	0.775	0.910
		8 p.c.'s	0	0	0.001	0.322	0.863
	25	Basis coef.	0	0	0	0.112	0.784
		3 p.c.'s	0	0	0.325	0.840	0.916
		5 p.c.'s	0	0	0	0.636	0.908
		8 p.c.'s	0	0	0	0.039	0.716
	35	Basis coef.	0	0	0	0.006	0.606
		3 p.c.'s	0	0	0.157	0.776	0.917
		5 p.c.'s	0	0	0	0.542	0.887
		8 p.c.'s	0	0	0	0.002	0.557

Table 2: Observed acceptance proportions for each scenario at a significance level 0.05 in the case of Log-normal errors

size of the group i. It would have been interesting to have at our disposal data related to RRAMs fabricated with a copper electrode and a dielectric 10 nanometres thick, but for reasons connected to the fabrication plans, it was not possible.

From mathematical viewpoint, and before applying FDA, the reset curves require some previous transformations because they are not defined on the same domain (reset voltages are different in each curve due to variability), and we only have discrete observations at a finite set of current values until the reset voltage is achieved in each curve. In order to solve these problems, [4] proposed a simple FDA approach to analyze these kind of curves prior to apply some specific statistical FDA techniques. Firstly, the initial domain $[0, V_{ij-reset}]$ was transformed in the interval $[0, 1]$ in a way that every registered sample curve $I^*_{ij}(u)(u \in [0, 1])$ has a new set of arguments given by transformation $u = v/V_{ij-reset}$. Secondly, taking into account that the curves are smooth enough, P-spline smoothing with B-spline bases was used to reconstruct all reset curves. The principal reasons why P-spline are usually considered a great accurate approximation of sample curves, are less numerical complexity and computational cost, and that the choice and position of knots is not determinant, so that it is sufficient to choose a relatively large number of equally spaced basis knots [2]. In this paper, for each reset curve of the three devices, it has been considered a cubic B-spline basis of dimension 20 with 17 equally spaced knots in the interval $[0, 1]$ and a penalty parameter $\lambda = 0.5$. In order to select the same smoothing parameter for all the sample paths a leave-one-out cross validation procedure was used.

Let us remember that the aim is to test if there are significant differences between RRAMs of the three different technologies under study. The first step is to test the equality of the three unknown mean functions by using the one-way FANOVA approach under the assumption that the reset curves of each group are generated by a Gaussian process with the same covariance operator. The estimation of the sample mean function in each group is displayed in Figure 1 (bottom-right) next to all the corresponding smoothed registered curves. Graphically, it seems that there are differences depending on the type of material and thickness.

In order to test the equality of the three unknown mean functions, MANOVA on the matrix of basis coefficients $A_{(4757\times 20)}$ could be applied (see Subsection 2.1) so that we have 20 dependent quantitative variables (the dimension of the B-spline basis) and one independent categorical variable (the three types of devices). It is well known that the main purpose of this technique is to compare the mean vectors of the three samples for significant differences. Equality of the mean vectors implies that the three single means are equal for each dependent variable. Before applying MANOVA, we must verify that the vectors of basis coefficients of each device technology have multivariate normal distribution with equal covariance matrices. However, these hypothesis are not fulfilled for the considered reset curves. As a matter of fact, the p-values associated with the Kullback's test or M-Box's test for the homogeneity of covariance matrices, and the p-value linked with the Kolmogorov-Smirnov's test for the univariate normality of each single basis coefficient are all < 0.001. This means that the assumptions of multivariate normality and homogeneity of covariance matrices are not true. Therefore, the second step consists of using a non-parametric test for the homogeneity of the vectors of basis coefficients in the three devices. In this application, due to the high presence of outliers, the extension of the univariate Mood's test which is based on spatial signs is employed (see results in Table 3). Taking into account that the associated p-value is less than 0.001, we can conclude that the reset voltage distribution is different according the kind of metal for the electrode and dielectric thickness used in the RRAM technologies.

Finally, we are going to test homogeneity on the functional principal components computed from the P-spline smoothing of the sample curves. The percentages of variance explained by the first four principal components are 99.639, 0.284, 0.046 and 0.020, respectively. Let us observe that only the first principal component explains more than 99% of the total variability of the process. Hence, by truncating K-L expansion 4 the reset process can be represented as $I^{*1}(u) = \overline{I}^{*}(u) + \xi_1^{*} f_1^{*}(u),\ u \in [0,1]$, where ξ_1^{*} is an scalar random variable called first principal component score and f_1^{*} is a function that represents the principal component weight curve. Thus, the problem of homogeneity is reduced to one-way ANOVA for the first principal component if this variable is normal distributed and with the same variance in the three devices. However, neither the normality nor the homogeneity of variance are accepted for these data so that the p-values associated with the corresponding tests (Kolmogorov-Smirnov's test for the univariate normality and Levene's Test for homogeneity of variance) are less than 0.001. Again, the ANOVA methodology can not be applied and so, non-parametric tests are used in order to test the differences among group means of the first p.c. Specifically, univariate Mood's median test is used for the general homogeneity hypothesis (see results in Table 3). It could be also interesting to test whether there are differences among pairs of devices in order to prove if the dielectric thickness or the electrode material by separated play some important role in the RRAMs operation. Wilcoxon's rank sum test is applied for the pairwise comparisons by means of the Benjamini's method for adjusting p-values. In both cases the associated p-values are less than < 0.001. On the other hand, the p-values provided by Wilcoxon's rank sum test for the pairwise comparisons are also smaller than 0.001. Based on these results we can conclude that the distribution of the first p.c. are significantly different for the

	Chi-squared	df	p-value
Basis coef.	5516.5	40	< 0.001
First p.c.	2538.2	2	< 0.001

Table 3: Chi-squared test statistic, the degrees of freedom of its approximated chi-squared distribution and the p-value for the Mood's median test

three considered devices. Therefore, it can be highlighted in what is referred to the reset curves of the technologies under consideration here that the type of metal employed for the electrodes and the dielectric thickness have a high influence on RRAMs operation and in the statistical information linked to their inherent variability.

5 Conclusions

The aim of this work is to decide if there are significant differences in the probability distribution that generates the reset processes associated with RRAMs fabricated making use of different materials for the electrodes and using dielectrics of different thicknesses. From the methodological point of view, this homogeneity problem consists of testing if different samples (groups) of curves come from the same population. Several FDA approaches have been proposed in literature when the stochastic processes associated with each sample are Gaussian. This problem is known as multi-sample problem or FANOVA and consists of testing the equality of the group mean functions. If the normality assumption is not true some bootstrap approaches were developed. In this paper, two different parametric and non-parametric homogeneity testing approaches are proposed by assuming a basis expansion of the sample curves. Both are reduced to testing multivariate homogeneity (parametric and non-parametric), the first one on a vector basis coefficients and the second one on a vector of principal component scores. The different proposals are motivated by the statistical study of the variability in the three samples of reset curves analyzed at the end of the paper. On the other hand, an extensive simulation study has been developed to check the practical performance of the testing approaches. In this study, the influence of sample size and variability of errors has been revealed, in addition to the improvement in the behavior of the tests with the principal component approach for the non-parametric case.

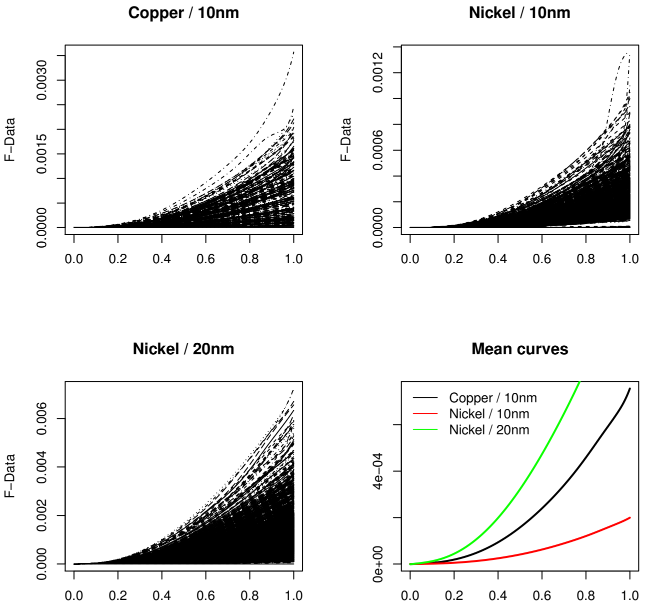


Figure 1: Sample group mean functions (bottom-right) and all the P-spline smoothed registered curves for each type of device.

References

[1] C. Acal, J.E. Ruiz-Castro, A.M. Aguilera, F. Jiménez-Molinos, and J.B. Roldán. Phase-type distributions for studying variability in resistive memories. *Journal of Computational and Applied Mathematics*, 345(1):23–32, 2019.

[2] A. M. Aguilera and M. C. Aguilera-Morillo. Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling*, 58(7-8):1568–1579, 2013.

[3] A. M. Aguilera and M. C. Aguilera-Morillo. Penalized PCA approaches for B-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219(14):7805–7819, 2013.

[4] M. C. Aguilera-Morillo, A.M. Aguilera, F. Jiménez-Molinos, and J.B. Roldán. Stochastic modeling of random access memories reset transitions. *Mathematics and Computers in Simulation*, 159(1):197–209, 2019.

[5] E. Brunner, E. Konietschke, A.C. Bathke, and M. Pauly. Ranks and pseudo-ranks: Paradoxical results of rank tests. *arXiv:1802.05650 [math.ST]*, 2018.

[6] A.R. Ellis, W.W. Burchett, S.W. Harrar, and A.C. Bathke. Nonparametric inference for multivariate data: the r package npmv. *Journal of Statistical Software*, 76(4):1–18, 2017.

[7] R. Flores, R. Lillo, and J. Romo. Homogeneity test for functional data. *Journal of Applied Statistics*, 45(5):868–883, 2018.

[8] M.B. González, J.M. Rafí, O. Beldarrain, M. Zabala, and F. Campabadal. Analysis of the switching variability in Ni/HfO$_2$-based RRAM devices. *Device and Materials Reliability, IEEE Transactions on*, 14(2):769–771, 2014.

[9] G. González-Cordero, J.B. Roldán, F. Jiménez-Molinos, J. Suñé, S. Long, and M. Liu. A new compact model for bipolar rrams based on truncated cone conductive filaments, a verilog-a approach. *Semiconductor Science and Technology,*, 31(11):115013, 2016.

[10] T. Gorecki and L. Smaga. Comparison of tests for the one-way anova problem for functional data. *Computational Statistics*, 30(4):987–1010, 2015.

[11] F. A. Ocaña, A. M. Aguilera, and M. Escabias. Computational considerations in functional principal component analysis. *Computational Statistics*, 22(3):449–465, 2007.

[12] H. Oja. *Multivariate Nonparametric Methods with R.* Springer Science & Business Media, 2010.

[13] E. Pérez, D. Maldonado, C. Acal, J.E. Ruiz-Castro, F.J. Alonso, A.M. Aguilera, F. Jiménez-Molinos, C. Wenger, and J.B. Roldán. Analysis of the statistics of device-to-device and cycle-to-cycle variability in tin/ti/al:hfo2/tin rrams. *Microelectronics Engineering*, 214(1):104–109, 2019.

[14] J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer-Verlag, 2009.

[15] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: Methods and case studies.* Springer-Verlag, 2002.

[16] J. O. Ramsay and B. W. Silverman. *Functional data analysis (Second Edition).* Springer-Verlag, 2005.

[17] A.C. Rencher and W.F. Christensen. *Methods of multivariate analysis (Third Edition).* Wiley, 2012.

[18] J.B. Roldán, F.J. Alonso, A.M. Aguilera, D. Maldonado, and M. Lanza. Time series statistical analysis: a powerful tool to evaluate the variability of resistive switching memories. *Journal of Applied Physics*, 125(1):174504, 2019.

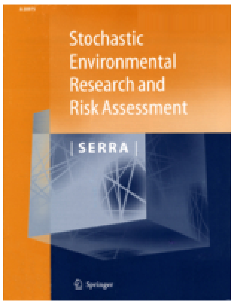
[19] S. Rusticus and C. Lovato. Impact of sample size and variability on the power and type one error rates of equivalence tests: A simulation study. *Practical Assessment, Research and Evaluation*, 19(11), 2014.

[20] M.A. Villena, J.B. Roldán, F. Jiménez-Molinos, E. Miranda, J. Suñé, and M. Lanza. Sim^{2}rram: A physical model for rram devices simulation. *Journal of Computational Electronics*, 16(4):1095–1120, 2017.

[21] J.T. Zhang. *Analysis of Variance for functional data.* CRC Press, 2014.

A6 Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis

- Acal, Christian; Aguilera, Ana M.; Sarra, Annalina; Evangelista, Adelia; Di Battista, Tonio; Palermi, Sergio (2021)
- Detecting changes in air pollution during the COVID-19 pandemic through Functional Data Analysis
- *Stochastic Environmental Research and Risk Assessment*, (under revision)



Statistics & Probability			
JCR Year	Impact Factor	Rank	Quartile
2019	2.351	21/124	Q1

Abstract

Faced with novel coronavirus outbreak, Italy as many other most hard-hit countries, in the Spring of 2020, adopted a lockdown strategy to contrast the spread of virus. Many studies have already documented that the COVID-19 control actions have resulted in improved air quality locally and around the world. Following these lines of research, in this paper, we analyze the impact of social distancing, travel limitations and restrictions placed upon economic activities on air quality changes in the urban territory of Chieti-Pescara (Central Italy), identified as an area of greatest criticality in terms of air pollution. Concentrations of NO_2, PM_{10}, $PM_{2.5}$ and benzene, measured in five monitoring stations, are used to evaluate air pollution changes in the area of interest. We track the air quality data over two specific periods: from 1^{st} February to 10^{th} March 2020 (before lockdown period) and from 11^{st} March 2020 to 20^{th} April 2020 (during lockdown period). The impact of lockdown on air quality is assessed by using functional data analysis methodologies. Our work makes an important contribution to the analysis of variance for functional data. Specifically, we propose a theoretical development for multivariate FANOVA for independent measures, based on multivariate functional principal components analysis of the sample curves and testing multivariate homogeneity on the vectors of the most explicative principal components scores. The functional analysis of variance has proven to be beneficial to monitoring the evolution of air quality before and during the lockdown tenure and to assessing the homogeneity of groups, individuated according to the location of measuring stations.

1 Introduction

After the discovery of the first case in Wuhan (China) in December 2019, the current outbreak of COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) has dramatically affected all the countries. On January 30 2020, the World Health Organization (WHO) declared worldwide public health emergency and in March 11 2020, due to widespread global infection, the WHO authorities categorised the new Coronavirus as pandemic [47]. To contain the virus and save lives, governments around the word have been taking a range of actions and measures, such as social and travel restrictions. More specifically, coronavirus pandemic has forced nations under partial or complete lockdowns, resulted in prohibition of unnecessary commercial activities in people's daily lives; prohibition of any types of gathering by residents; restrictions on private (vehicle) and public transportation. Different studies (see, among others, [18, 7, 40, 32, 28]) have already documented the effects of COVID-19 lockdown measures on many aspects of human activities. Certainly, COVID-19 has severe negative impact on the world activities as well as on local economy. In the major economics across the globe, lockdown will directly affect the Gross Domestic Product (GDP) of each country. Meanwhile, efforts to restrict transmission of the SARS-CoV-2 have had outstanding effects on the ecosystems which are being greatly recovered. In many cities where lockdown measures have been implemented, the decline in economic activities, the non-functioning of industries, the drop in road transport, have contributed to mitigate air pollution. In this respect, several researchers around the world reported that there is a considerable reduction of air pollution level across geographies. For instance, [8] investigated the changes in levels of air pollutants across USA during COVID-

19 pandemic. They reported a significant reduction on NO_2 (up to -25.5%) and an overall decline in $PM_{2.5}$, compared with pre-lockdown phase. [26] found a radically improvement for air quality indexes in 88 Indian cities only four days of beginning of the lockdown. [2] recorded a visible improvement in air quality parameters in some cities of India and China, selected on the basis of their availability of historical air pollution data, population density, monitoring station network, and the number of positive COVID-19 cases per million people. [48] studied the effects of quarantine policies on air pollutants concentrations in Quito, Ecuador. Using parametric methods, they detected a significant reduction of NO_2 and $PM_{2.5}$ since the introduction of lockdown measures. However, there was a noticeable growth in ozone levels. The similar trends of reducing air pollution and increasing air quality due to introduction of lockdowns were observed in several other countries as well, such as Malaysia [23], in Wuhan (China), Turin and Rome (Italy), Nice (France), Valencia [42].

In this study, we focus on investigating the possible effects of the lockdown due to the COVID-19 pandemic on air quality in the Pescara-Chieti urban area, Abruzzo (Italy), identified as an area of greatest criticality in terms of air pollution. Data of monitoring stations of the regional air quality network managed by the Regional Agency for the Environmental Protection (ARTA) of Abruzzo have been collected and examined. We compared data from the 1^{st} February to the 10^{th} March 2020, before the beginning of the main limitations on personal mobility, with data from the 11^{st} of March to the 20^{th} of April, during the adoption of lockdown restrictions. Measured concentrations of NO_2, PM_{10}, $PM_{2.5}$ and benzene were used to evaluate air pollution changes. Commonly, strategies used in monitoring air quality refer to descriptive statistics, box-plots, autocorrelation analysis and spatio-temporal models. Unfortunately, in the monitoring of environmental pollutants, the temporal observations of the different pollutants and for the different stations have not always been referred to the same instants of time. As a result, the implementation of classical statistical procedures might be problematic. Besides, for interpretative purposes, it is convenient to rely on statistical methods able to capture the speed and acceleration of pollutants variation over time. For these reasons, in our research, to overcome the weakness of classical statistical procedures and to effectively detect to what extent extreme changes in human behaviour after the quarantine policies adopted by the Italian Government have affected air quality, we followed an approach based on functional data analysis (FDA).

During the two last decades, it has emerged an important literature in this methodological framework. A comprehensive introduction to the foundations and applications of FDA can be found in [35, 36], whereas nonparametric functional methods are summarised in a monograph by [16]. As it is well known, FDA extends the classical multivariate techniques to data transformed in functions or curves, with the advantage of reducing thousands of observations to a few coefficients but conserving relevant information about the functional form. Recently, the use of FDA methods for environmental data has received attention (see, among others, [13, 17, 5, 46, 31, 39, 14, 21, 6, 18]). Also in this study, rather than simply considering the data as vectors to apply classical multivariate analysis methods, which may lead to a loss of useful information, we explicitly exploit the functional form of environmental data. The FDA makes it possible to work with the entire time spectrum of pollutants time series and detect small deviations from the normal behaviour of the data. Our goal in this paper is to ascertain whether the level of each pollutant has changed during the lockdown period. In other terms, we want to test the equality of mean functions related to each pollutant in two different periods of time: before and during lockdown days. The theoretical framework involves the use of FDA tools for repeated measures, and in particular, the analysis of variance. In the literature there are not many works related to this matter for the field of FDA. In this work, a basis expansion approach for the statistics proposed by [27] and [45] to test the equality of two mean functions is considered. On the other hand, in order to check the differences between the temporal evolution of all pollutants in terms of the location of measuring stations a novel approach for multivariate FANOVA for independent measures is introduced. This is based on multivariate functional principal components analysis of the sample curves of all pollutants and testing multivariate homogeneity or MANOVA (gaussian data) on the vectors of the most explicative principal components scores. This new methodology is the extension of the parametric and nonparametric approaches proposed by [3] for the univariate functional case.

The paper is organized as follows. Section 2 introduces the theoretical framework. Section 3 is dedicated to illustrate the study area, the monitoring stations where air quality data have been collected and some explorative analysis of pollutants used for this research. In Section 4, our method is applied to checking for differences between air quality data collected from the monitoring stations placed in the urban area of Chieti-Pescara. Finally, Section 5 concludes the paper.

2 Theoretical framework

Let $\boldsymbol{X}_{ijr} = (X_{ijr1}, ..., X_{ijrH}), i = 1, ..., g, j = 1, ..., n_i, r = 1, ..., R$ be a sample of vectors of curves. Note that g represents the number of independent groups, H is the number of observed response variables, R denotes the number of different periods of time (or conditions) where the response variable is observed (repeated measures) and $n = \sum_{i=1}^{g} n_i$ is the sample size. It is considered that these curves are realizations of a H-dimensional stochastic process $\boldsymbol{X} = (X_1, ..., X_H)$, whose components are second order and continuous in quadratic mean stochastic processes with sample paths belonging to the Hilbert space $\mathrm{L}^2[\mathcal{T}]$ of squared integrable functions on $\mathcal{T}$, with the natural inner product

$$< f, g >= \int_{\mathcal{T}} f(t)g(t)dt \ , \ \forall f, g \in \mathrm{L}^2[\mathcal{T}].$$

2.1 FANOVA for repeated measures

The goal is to test the equality of mean functions associated with the observation of a functional variable in two different conditions or periods of time for the same subjects. For instance, the problem laid out in this paper about the evolution of the quality of the air before and during the lockdown. That is, whether the level of each pollutant has changed during the lockdown. The theoretical framework involves the use of tools for repeated measures, and in particular, the analysis of variance. In the literature there are not many works related to this matter for the field of FDA. [27] introduced the first testing procedure for this problem by keeping in mind the between group variability. They proposed three different approaches in order to approximate the null distribution. The first technique consisted of applying a bootstrap parametric method through resampling some involved Gaussian processes. The second and third methods were based on non-parametric approaches via bootstrap and permutation tests. Later, [44] proposed another perspective focused on the Box-Type approximation. In that study, the four methods were compared, turning out to be the Box-Type approximation the most efficient option from the computational viewpoint. In relation to size control and power all of them gave similar results, and its behaviour with finite samples was very satisfactory. However, both works agree that for very small sample size, the bootstrap tests are lightly nonconservative. [45] adapted two new statistics from the classical paired t-test to the functional data framework. This new approach is more powerful than the testing procedures aforementioned because takes the within group variability into account as well. The distributions of the statistics were also approximated by parametric methods based on asymptotic distributions as well as non-parametric bootstrap and permutation approaches. The simulation study proved that the asymptotic and Box-Type tests are not recommended because of their liberality. [45] suggested the permutation tests, although the non-parametric bootstrap methods also worked correctly. Nevertheless, it was emphasized that there are evidences that the procedures proposed tend to be nonconservative for small sample size. On the other hand, the false discovery rate for functional data has been recently introduced in [30] for the continuous statistical testing of the null hypothesis along the functional data domain, which can be seen as an extreme case of the multiple comparisons problem.

In what follows, a single functional variable is considered because they are going to be dealt separately. In this context, it is assumed that the sample functions can be represented as $X_{jr}(t)$ with $t \in \mathcal{T} = [a,b]$, $j = 1, ..., n$ and $r = 1, ..., R$, such that $E[X_{jr}(t)] = \mu_r(t)$. Only two different conditions or periods of time are evaluated in the current work ($R = 2$). Besides, each trajectory can be expressed as $X_{jr}(t) = \mu_r(t) + e_{jr}(t)$ where $e_{jr}(t)$ are independent random functions centered in mean. In this kind of problem the pursued goal is to test the hypothesis

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

[27] proposed the following statistics in order to solve the statistical hypothesis testing

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

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

Due to $\mathcal{C}_n$ only takes the between group variability, [45] proposed the following two statistics in order to consider both the between and within group variabilities

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

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

with $\hat{K}(t,t) = \frac{\sum_{j=1}^n \left[(X_{j1}(t) - \overline{X}_1(t)) - (X_{j2}(t) - \overline{X}_2(t))\right]^2}{n-1}$.

One of the biggest problems in practice is that curves are observed at a finite set of times because it is impossible to observe a set of functions continuously in time. Thus, the first step would be to reconstruct the functional form of the curves. [16] proposed to use non-parametric techniques for this purpose, meanwhile [35, 36] suggested an approach based on basis expansions of the sample curves. This last strategy consists of assuming that curves belong to a finite-dimension space spanned by a basis $\{\phi_1(t), ..., \phi_p(t)\}$, so that they can be expressed as

$$X_{jr}(t) = \sum_{k=1}^{p} a_{jrk} \phi_k(t) = \mathbf{a}'_{jr} \boldsymbol{\phi}(t) \ , \ j = 1, ..., n; r = 1, 2,$$

where a_{jrk} represent the basis coefficients of the reconstruction for the corresponding sample curve with $\mathbf{a}_{jr} = (a_{jr1}, ..., a_{jrp})'$ and $\boldsymbol{\phi}(t) = (\phi_1(t), ..., \phi_p(t))'$.

Note that p must be sufficiently large to guarantee an accurate approximation of the original curve. Besides, it is necessary to choose properly the dimension and the type of basis by keeping in mind the nature of the curves. There are numerous basis systems but the most employed ones are Fourier functions (for periodic data), B-spline (for non-periodic and smooth data) and wavelets (for curves with strong local behaviour). Finally, sample trajectories can be observed with error or without error. For the first case, least squares approximation is usually used in order to estimate the basis coefficients, whereas for the second case some interpolation method could be applied. For more details about these methodologies, [36] carried out an exhaustive study and [34] implemented them with the software R.

In this paper, $\mathcal{C}_n$, $\mathcal{D}_n$ and $\mathcal{E}_n$ are computed by considering the basis expansion approach. In fact, it is direct to prove that

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

and

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

with $\overline{\mathbf{d}} = (\overline{d}_1, ..., \overline{d}_p)' = \overline{\mathbf{a}}_1 - \overline{\mathbf{a}}_2 = (\overline{a}_{11}, ..., \overline{a}_{1p})' - (\overline{a}_{21}, ..., \overline{a}_{2p})'$, where $\overline{a}_{rk} = n^{-1}\sum_{j=1}^{n} a_{jrk}$ $r = 1, 2;$ $k = 1, ..., p$. In addition, $\hat{\Sigma}_r$ is the sample covariance matriz of the matrix A_r of basis coefficients in the group r, whose elements are $A_r = (a_{jrk})$, and $\hat{\Sigma}_{12}$ is the sample cross-covariance matrix between A_1 and A_2. Note for major clarity that $\overline{X}_r = n^{-1}\sum_{j=1}^{n} \mathbf{a}_{jr}'\boldsymbol{\phi}(t) = \overline{\mathbf{a}}_r'\boldsymbol{\phi}(t)$.

2.2 Multivariate FANOVA for independent measures

Now the idea in this kind of analysis is a little bit different than the case of repeated measures. The aim is to test the equality of the mean functions coming from independent groups. For example, the evolution of level of benzene in the air in two different regions. If there is only a response variable (e.g. level of benzene), the problem is known as univariate FANOVA. Likewise, another fundamental aspect in these studies is the number of factors that determine the different groups. If there exists only one factor (e.g. regions) the problem is called one-way FANOVA. There are several existing methods for testing the one-way FANOVA problem [15, 10, 49, 50]. A robust simultaneous confidence band for the difference of mean functions of two independent populations was introduced in [25]. On the other hand, [19] made a detailed comparison of tests for the one-way FANOVA problem with approaches based on a basis expansion of curves. These tests were inspired by the idea of the B-Spline method of [41]. In this line, [3] suggested a novel approach by using Functional Principal Component Analysis (FPCA). This method consists of testing multivariate homogeneity on a vector of principal components scores. However, although there are available many works for the univariate functional case, the natural extension for the multivariate case (more than one functional response variable) had not been studied deeply. Permutation tests based on a basis function representation and tests based on random projections are studied in [20]. Here, a novel approach based on multivariate FPCA is introduced for dealing with the multivariate FANOVA problem. This new methodology can be seen as the extension of the parametric and nonparametric approaches proposed by [3].

Let us consider a set of curves $X_{ijh}(t)$ with $i = 1, ..., g$, $j = 1, ..., n_i$ and $h = 1, ..., H$ are a set of curves. Then, the information for each subject is a vector of curves denoted by $\boldsymbol{X}_{ij}(t) = (X_{ij1}(t), ..., X_{ijH}(t))'$. Besides, it is assumed that $\boldsymbol{X}_{ij}(t)$ are i.i.d. multivariate functional variables with mean vector $\boldsymbol{\mu}_i = (\mu_{i1}(t), ..., \mu_{iH}(t))'$ and matrix covariance function $\mathbf{C}$ such that $\mathbf{C}(t, s) = (C_{h,h'}(t, s))$, $t, s \in \mathcal{T}$ and $h, h' = 1, ..., H$. Note that if $h = h'$, then $C_{h,h}$ is the covariance function and otherwise, that is $h \neq h'$, $C_{h,h'}$ represents the cross-covariance function. Now, the aim is to test

$$H_0 : \boldsymbol{\mu}_1(t) = ... = \boldsymbol{\mu}_g(t) \ \forall t \in [a, b], \tag{1}$$

against the alternative that its negation holds.

In the field of FDA, it is very common to deal with high dimension data. These type of data are defined as data associated to a great number of highly correlated variables where the sample size is too much small. For this reason, one of the most important techniques in FDA is FPCA. This tool reduces the dimension of the problem and explains the main characteristics and modes of variation of the curves in terms of a reduce set of uncorrelated variables called functional principal components (PC's). Basis theory on FPCA was first introduced by [12] and asymptotic properties were studied in [11]. Penalized estimation approaches for univariate FPCA were later developed in [43] and [4]. Recently, a new varimax rotation for FPCA has been performed in [1]. [35] presented a detailed study of the basis expansion estimation for univariate FPCA and discussed its extension to the case of bivariate functional data. It is immediate to adapt this theory when more than two response variables are considered. PC's are obtained as generalized linear combinations with maximum variance. Formally, the m-th principal component scores are determined by

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

where $\boldsymbol{\mu}(t) = (\mu_1(t), ..., \mu_H(t))$ is the overall mean function and $\boldsymbol{f}_m(t) = (f_{m1}(t), ..., f_{mH}(t))'$ are the vector of weight functions (or loadings) that maximizes the variance subject to $\sum_{h=1}^{H} \int_{\mathcal{T}} f_{mh}(t) f_{m'h}(t) dt = 1$ if $m = m'$ and 0 otherwise. These functions are obtained as the solutions to the eigenequation system

$$C\boldsymbol{f}_m = \lambda_m \boldsymbol{f}_m,$$

with C being the covariance operator and the sequence $\{\lambda_m\}_{m \geq 1}$ of positive real eigenvalues decreasing to zero indicating the amount of variance attributable to each component. The aforementioned system can be written in detail as follows:

$$\begin{array}{c}\int_{\mathcal{T}} C_{11}(s,t) f_{m1}(t)dt + ... + \int_{\mathcal{T}} C_{1H}(s,t) f_{mH}(t)dt = \lambda_m f_{m1}(s) \\ \int_{\mathcal{T}} C_{21}(s,t) f_{m1}(t)dt + ... + \int_{\mathcal{T}} C_{2H}(s,t) f_{mH}(t)dt = \lambda_m f_{m2}(s) \\ \vdots \\ \int_{\mathcal{T}} C_{H1}(s,t) f_{m1}(t)dt + ... + \int_{\mathcal{T}} C_{HH}(s,t) f_{mH}(t)dt = \lambda_m f_{mH}(s).\end{array}$$

Highlight that each PC is a zero-mean random variable with maximum variance and uncorrelated with the remainder of PC's. Hence, in the multidimensional context and similar to the univariate setting, the vectorial process admits the following orthogonal decomposition known as Karhunen-Loève expansion

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

This decomposition can be truncated so that the sample curves can be optimally approximated (in the least squares sense) in terms of the first q PC's, $\boldsymbol{X}_{ij}^q(t) = \boldsymbol{\mu}(t) + \sum_{m=1}^{q} \xi_{ijm} \boldsymbol{f}_m(t)$. The parameter q is normally chosen so that the explained cumulative variability is as close as possible to one. With this approach, the dimension of the problem is considerably reduced.

Multivariate FPCA with basis expansions was first introduced by [22] and later summarized in [38]. The main ideas are briefly explained hereafter. If the basis expansion is considered, $\boldsymbol{X}_{ij}(t)$ can be expressed as

$$\boldsymbol{X}_{ij}(t) = \boldsymbol{\Phi}(t)\mathbf{a}'_{ij},$$

where the basis coefficients are gathered as $\mathbf{a}_{ij} = (a_{ij11}, ..., a_{ij1p_1}, a_{ij21}, ..., a_{ij2p_2}, ..., a_{ijH1}, ..., a_{ijHp_H})$ with p_h being the number of basis functions for the h-th response variable and

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

In general $\boldsymbol{X}(t) = \boldsymbol{A\Phi'}(t)$, where $\boldsymbol{A}$ is the resultant matrix after joining by row all $\mathbf{a}_{ij}$. Thus, whether the mean vector is subtracted to each row of $\boldsymbol{X}(t)$, the spectral decomposition of the covariance operator C becomes

$$\boldsymbol{\Phi}(s)\Sigma_A \boldsymbol{W}\boldsymbol{b}'_m = \lambda_m \boldsymbol{\Phi}(s)\boldsymbol{b}'_m,$$

with Σ_A being the covariance matrix of $\boldsymbol{A}$, $\boldsymbol{b}_m$ being a row-vector that contains the basis coefficients of $\boldsymbol{f}_m(t) = \boldsymbol{\Phi}(t)\boldsymbol{b}'_m$ and $\boldsymbol{W} = \int_{\mathcal{T}} \boldsymbol{\Phi}(t)'\boldsymbol{\Phi}(t)dt$ being the matrix of inner products between basis functions with dimension $\sum_{h=1}^{H} p_h \times \sum_{h=1}^{H} p_h$. Since the showed spectral decomposition is true for all s, the expression can be reduced as $\Sigma_A \boldsymbol{W}\boldsymbol{b}'_m = \lambda_m \boldsymbol{b}'_m$. Now, by considering $\boldsymbol{u}_m = \boldsymbol{b}_m \boldsymbol{W}^{1/2}$, the multivariate FPCA is equivalent to the multivariate PCA of the matrix $\boldsymbol{A}\boldsymbol{W}^{1/2}$, whose covariance matrix can be diagonalised as follows

$$\boldsymbol{W}^{1/2'}\Sigma_A \boldsymbol{W}^{1/2}\boldsymbol{u}'_m = \lambda_m \boldsymbol{u}'_m.$$

Therefore, the PC's are given by

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

[37] implemented in the software R [33] the package called '*funHDDC*' that provides the principal component scores for the multivariate case.

Once the multivariate PC scores are computed, two different ways for solving the multivariate testing problem (1) are proposed in this paper, both are based on testing homogeneity on the vector of the first q principal components scores in the g groups. The first consists of performing univariate ANOVA on each principal component by correcting the level of significance for the normality case. It is well-known that whether the multivariate normality is suitable, the uncorrelatedness implies independence and then, it does not make sense to consider a multivariate approach. Otherwise, when the multivariate normality is not satisfied, the option is to apply non-parametric multivariate tests such as the extensions of the univariate Kruskal Wallis's test and Moods's test. In addition, it is recommended to use the permutation version of these tests when the sample size is small [29].

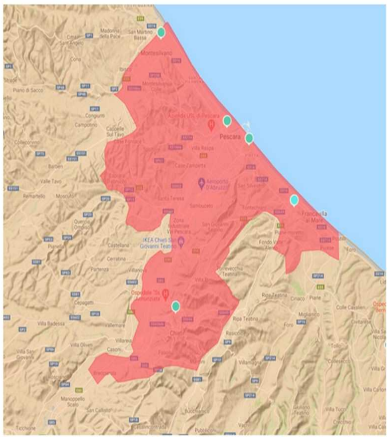


Figure 1: Abruzzo and Chieti-Pescara metropolitan area

3 Air quality data and studied period

This section gives details about the study area with respect to its geographic characteristics and the monitoring stations used to collect air quality data. In addition, the dataset employed to derive insights into research problem has been analysed with descriptive statistics and visualization tools. For high quality graphics, we used the *'openair'* R package [9].

3.1 Description of study region

In this work, we closely examine the air quality of the metropolitan area of Chieti-Pescara, situated in the Abruzzo region, along the Adriatic coast of central Italy. The Chieti-Pescara metropolitan area (Fig.1) is a territory, identified according to a functional criterion, formed by six municipalities, Pescara, Montesilvano, Chieti, Francavilla al Mare, San Giovanni Teatrino, Spoltore, covering a total area of 159.33 km^2 , and accounting for around 281,101 inhabitants at 31/12/2019.

The configuration of Chieti-Pescara urban area is limited by the sea, in the North-East, and by hilly reliefs in the South-West. The central city is formed by the two provincial capitals: Chieti, not in a central position for the municipalities of the province, and the city of Pescara, which are extremely close to each other (approximately 12 Km). Pescara city, located on the centre of a metropolitan area (on the coast), is the administrative and commercial heart of Abruzzo and in a few decades, it has become the most populated city of the region, with 120,000 inhabitants. It developed on a flat territory, with a surface of 33.62 km^2, whose urban area develops around the terminal stretch of the homonymous river and a restricted coastal area.

The Chieti-Pescara conurbation is characterised by a system of infrastructures, which is one of the strongholds of the Abruzzo: significant and industrial sites are located around this pole. However, the progressive growth of the industrial activity, the increased road travel, the urban expansion, make the metropolitan area the locus of growing environmental concerns, for the rising levels of energy and resource consumption, greenhouse gas emissions and air quality pollution. For these reasons, the conurbation of Chieti-Pescara has been identified as a mitigation area by the "Plan for the Protection of Air Quality", drafted by the Abruzzo Region, in accordance with current Italian legislation (Legislative Decree 155/2010). That document has highlighted the need to reduce the impact on the population of the air pollution levels that characterize this urban territory. Also, in the Plan, there is the inventory of the main sources of polluting emissions (updated to 2012), which in this area largely sees the contribution of non-industrial combustion plants (mainly domestic heating plants) as regards particulate emissions (78.5% of the total for PM_{10} , 88.8% for $PM_{2.5}$ and 88.4% for benzene), while for nitrogen oxides road traffic prevails (49.7%), with industrial combustion plants in second place with 23.6 %.

3.2 Data

This study tracks four pollutants over two specific periods: from 1^{st} February to 10^{th} March 2020 (before lockdown period) and from 11^{st} March 2020 to 20^{th} April 2020 (strict lockdown period). The analyses include measures of NO_2, PM_{10}, $PM_{2.5}$ and benzene (C_6H_6) obtained from the automatic reporting platform, run by Regional Agency for the Environmental Protection of Abruzzo Region (ARTA). These variables are measured in micrograms per cubic meter (μg/m^3) and information are obtained from five monitoring sites. The spatial location of all five monitoring stations is shown by blue points in Figure 1. The air monitoring stations of Pescara (Via Firenze) and Montesilvano are designed as *Urban Traffic type* (UT) and are located where the pollution level is most influenced by traffic emissions from neighboring roads with medium-high traffic intensity; conversely, air quality data collected from the monitoring stations of Pescara (Teatro d'Annunzio), Chieti and Francavilla are deemed *Urban Background measuring stations* (UB), located where the pollution level is not influenced mostly by emissions from specific sources and are representative of

Table 1: Net and % variation of pollutants concentration levels in the urban area of Chieti-Pescara

	UT		**UB**		
Net variation	**fi**	**mo**	**th**	**ch**	**fr**
NO_2	-13.9	-14.7	-21.2	-10.3	-7.6
PM_{10}	5.1	3.7	5.7	4.3	7.3
$PM_{2.5}$	2.9	2.2	3.1	4.4	4.1
Benzene	-0.31	-0.15	0.22	0.18	0.04
% variation					
NO_2	-57.9	-58.7	-65.2	-54.8	-49.1
PM_{10}	20.5	16.8	22.0	19.4	40.8
$PM_{2.5}$	19.0	15.6	19.8	26.7	34.4
Benzene	-32.57	-27.56	40.06	19.63	4.27

Acronyms of monitoring stations:

fi=Via Firenze; mo=Montesilvano; th=Teatro d'Annunzio; ch=Chieti; fr=Francavilla al Mare

the population average exposure. Hourly measurements of pollutants have been collected from February to April 2020.

3.3 Descriptive statistics and graphical analysis

After the implementation of strict lockdown measures starting from 11^{st} March 2020, air pollution of the urban area of Chieti-Pescara has witnessed a substantial improvement. Table 1 highlights the net and percentage variations of each pollutant, in each monitoring site, before and during the lockdown. It can be noticed that NO_2 has shown the most significant declining trend. In particular, the concentrations of this pollutant were approximately 50% lower compared to the previous average. On the other hand, we recorded an increase of PM_{10} and $PM_{2.5}$ concentrations during the lockdown weeks, whereas benzene levels dropped in the traffic measuring stations and increased in the background monitoring sites. A trend analysis of 24-hour daily average data for the four pollutants was also considered for the above stated periods in all monitoring stations to better understand the impact in the levels of pollutants accumulation amid the lockdown period.

Figure 2 allows to capture the changes in concentrations of four pollutants for the pre-lockdown and during-lockdown period. The reduction of NO_2 during the lockdown is clearly visible and marked in all monitoring sites and is due to the collapse of vehicular flows, even if differences emissions in magnitude exist depending on the stations. The behaviour of atmospheric particular matter (PM_{10} and $PM_{2.5}$) seems to be rather independent from the measures adopted during the COVID-19 nation-wide lockdown: background and traffic stations undergo an increase during the period of lockdown directives. This implies that the monitoring sites might be under the effect of multiple non-transportation related emission sources. In particular, the noticeable increase of $PM_{2.5}$ concentration level during the strict lockdown phase is mainly ascribable to the increase of domestic solid fuel burning. Besides, a pertinent amount of PM_{10} and $PM_{2.5}$ has a meteorological origin. Particular matter concentrations may also fluctuate with inter-seasonal dissimilarity in meteorological conditions. Thus, the higher PM_{10} and to a lesser extent $PM_{2.5}$ episodes, occurred at the end of March (29-31 March), were recorded during a massive advection of dust from the Central Asia (Aral Sea). Concerning the benzene concentrations, as stated earlier, it appears that this pollutant exhibits very different behaviours in background stations compared to traffic ones. In the latter, there is a decline, albeit contained, due to the reduction of vehicular flows during the lockdown, while in the former there is stability or even slight increases, probably due to domestic heating systems and the role of weather variables. To get a comprehensive understanding of how lockdown policies have affected air pollution, we also look at the weekly concentrations of each pollutant at background and traffic stations before and during the restriction periods. From Figures 3 and 4, it is evident that on Sunday, the traffic during the lockdown phase is virtually zero, therefore the concentrations of NO_2, pollutant specifically linked to vehicle emissions, reduce more than in the other days of the week. It is worth noting that "Teatro d'Annunzio" measuring station seems to be affected by traffic emissions in an anomalous way being it a background site. Also the inspection of weekly concentrations reveals that the impact of restrictions measures on PM_{10} and $PM_{2.5}$ is the most complex of the four pollutants studied: we are not able to detect consistent patterns with vehicular flows even if the small sample size, only five weeks and half, could affect the empirical findings. More in detail, among the background monitoring stations, that of "Teatro d'Annunzio" results more subject to the natural component of PM_{10}, probably due to marine aerosol: this station is about 200 m from the sea, with no buildings in the way. Regarding the weekly benzene concentrations, the comparison between the two traffic monitoring sites indicates that "Via Firenze" is more subject to the emissions arising from combustion (mainly domestic heating) compared to those due to traffic road. Conversely, the traffic station of "Montesilvano" being on the edge of the urban area is especially exposed to road traffic emissions.

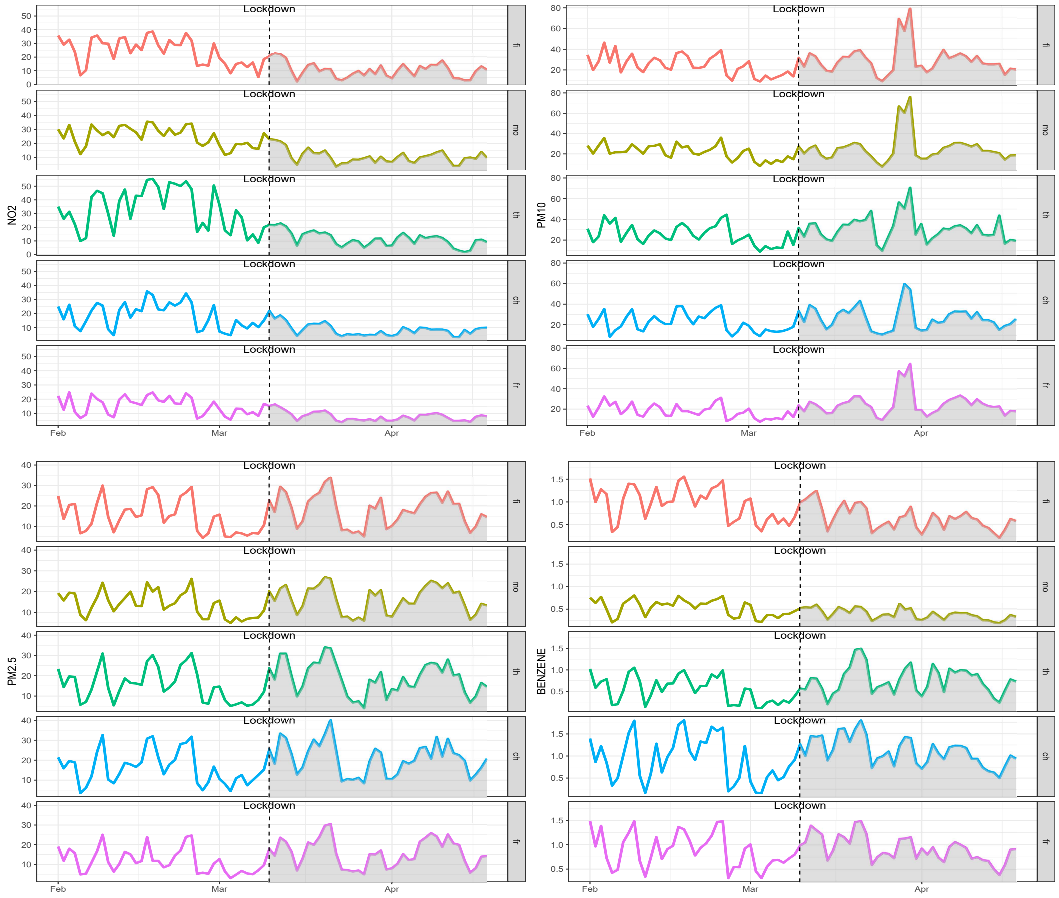


Figure 2: Daily variation of pollutants for stations before and during lockdown occurred on 10th March

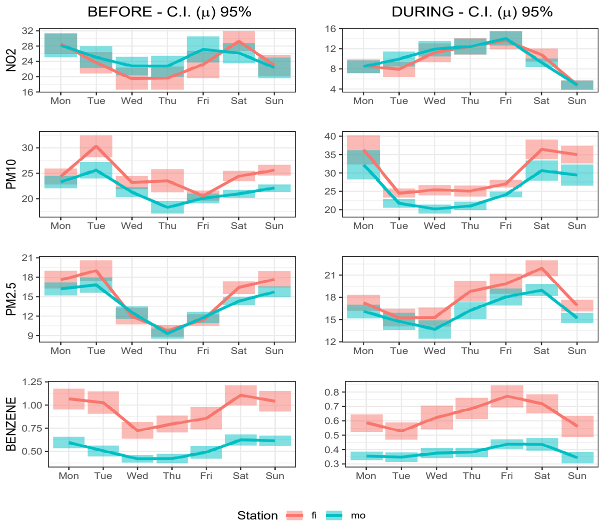


Figure 3: Weekly concentrations of each pollutant at traffic stations before (left panel) and during (right panel) lockdown

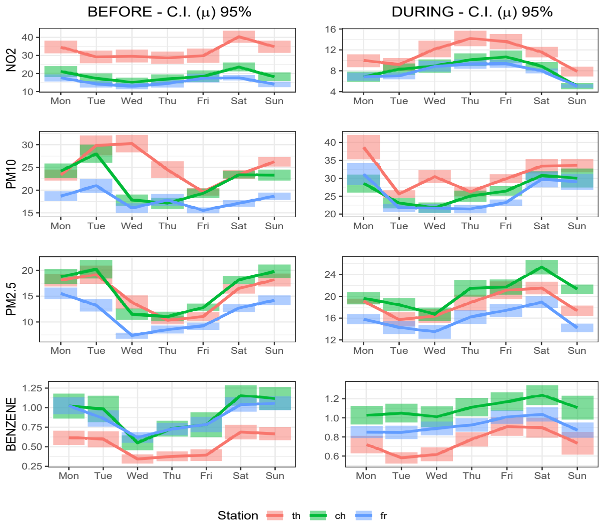


Figure 4: Weekly concentrations of each pollutant at background stations before (left panel) and during (right panel) lockdown

4 Results

We now illustrate the use of testing procedures previously described to ascertain whether the level of each pollutant has changed during the lockdown period. As pointed out in Section 3, to reveal the impact of restriction measures due to the COVID-19 on the air quality, the obtained environmental datasets were divided in two time frames, of the same length (39 days): i) pre-lockdown (February 1, 2020-March 10, 2020) and ii) during-lockdown (March 11, 2020-April 18, 2020). From a theoretical viewpoint, we have longitudinal functional data corresponding with the observation of the same functional variables in two different periods of time.

4.1 Functional reconstruction of pollutant curves

As a first step of our data analysis, we reconstructed the functional form of curves from the initial points that come from the discrete values measured in the study. To convert the discretely observed data to smooth functions, the reconstruction of curves is made by using a cubic B-spline smoothing. The B-spline functions are one of the most prominent spline basis, used for non-periodic functions, which is proven to be numerically stable and flexible [36]. Initially, in tailoring a basis system to fit our data, we used 7 basis functions. This option is conservative: it allows to capture the trend of curves but not their local behaviour. To recover the underlying functions of the observed data, we were increasing the number of basis functions up to 20. This choice preserves important information about the real form of the curves. Figure 5 illustrates the shape the data would take after smoothing them into these basis systems. It is clear that the increase of the number of basis functions produces smaller differences between the smoothed sample curves and the observed data (see Figure 6). Hereinafter, an approximation of each sample curve in terms of a basis of cubic B-splines of dimension 20 is considered.

4.2 FANOVA for repeated measures results

Before moving on more complex studies, we carried out a univariate analysis to evaluate the behaviour of each pollutant before and during lockdown. To statistically confirm the effect of lockdown on the mean of each pollutant, we first implemented the FANOVA for repeated measures, as defined in Section 2.1. Specifically, we apply the statistics $\mathcal{D}_n$ and $\mathcal{E}_n$ which are the best to control the between and within group variability that there are behind the repeated measures design. In order to construct the tests based on these statistics, a permutation method is used to approximate their null distributions. This technique

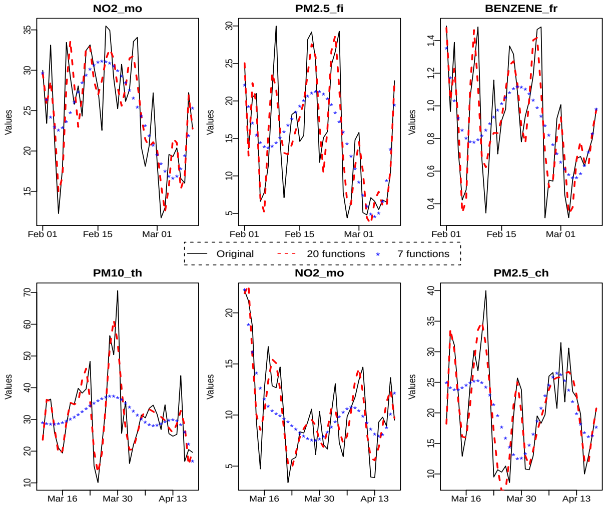


Figure 5: Functional approximation, using 7 and 20 basis functions, of some pollutants for stations before lockdown (upper panel) and during lockdown (lower panel)

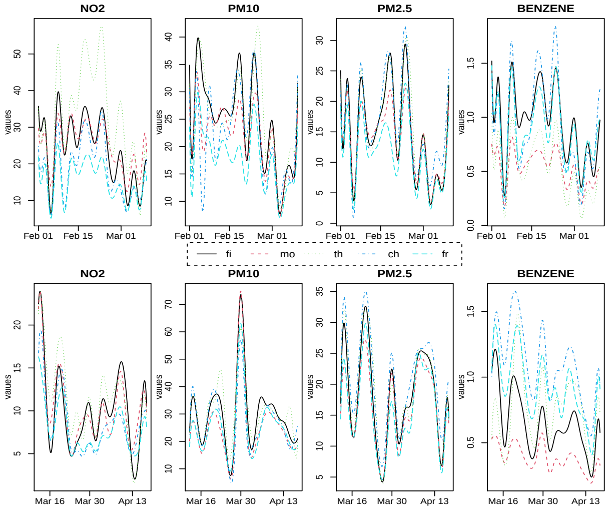


Figure 6: Functional approximation, using 20 basis functions, of pollutants for stations before lockdown (upper panel) and during lockdown (lower panel)

Table 2: FANOVA for repeated measures on the test statistics $\mathcal{D}_n$ and $\mathcal{E}_n$

p-value	$\mathcal{D}_n$	$\mathcal{E}_n$
NO_2	0.034	0.035
PM_{10}	0.000	0.034
$PM_{2.5}$	0.028	0.030
Benzene	0.049	0.070

consists of a random permutation of each sample unit. Let us denote the original data by $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ where $\boldsymbol{X_j} = (X_{j,1}, X_{j,2})$ $(j = 1, \ldots, n)$, and the resampling vectors by $\boldsymbol{X}_* = (\boldsymbol{X_1}*, \ldots, \boldsymbol{X_n}*)$ with $\boldsymbol{X_j}* = (X_{j,1}*, X_{j,2}*)$ being a random permutation of the sample unit X_j. This process is repeated Δ times, with Δ a number sufficiently large, so that $\mathcal{D}_n^\delta*$ and $\mathcal{E}_n^\delta*$ are calculated for each replication, being $\delta = 1, \ldots, \Delta$. Later, p-values are obtained as the proportion of times that $\mathcal{D}_n^\delta*$ and $\mathcal{E}_n^\delta*$ overcome $\mathcal{D}_n$ and $\mathcal{E}_n$, respectively. Here, the p-values were obtained from 2000 replications. The results of the proposed testing procedures are shown in Table 2. The p-values of all tests are less than the significance level $\alpha = 0.05$ for NO_2, PM_{10} and $PM_{2.5}$. For benzene, $\mathcal{E}_n$ shows no differences between both periods, but it is very close to the limit region. Therefore, and taking into account the sample size, we have evidence to reject the null hypothesis for benzene and we state that there are also differences in the mean curves of this pollutant in the pre and during lockdown phases.

4.3 Multivariare FANOVA results for independent measures

Once the impact of the lockdown has been studied, a further step of our data analysis has involved the assessment of equality of mean functions of individual groups. In our context, the groups have been individuated according to the location of the monitoring sites. In more detail, our interest lies in investigating if the mean function of all the pollutants measured in the background stations is equal to that of the urban traffic ones. The multivariate analysis of variance is carried out both before and during lockdown tenure to detect differences attributable to the government restrictions. This comparison has been evaluated firstly globally, considering all the pollutants together, and then for each variable separately. In Table 3 the results for multivariate and univariate FANOVA based on FPCA are displayed. On this matter, four principal components are chosen for both cases (multivariate and univariate analysis), since more than a 99% of total variability is explained with four components in all situations. Besides,

Table 3: Multivariate FANOVA for independent measures

p-value	**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

Acronyms:

BL=Before Lockdown; DL=During Lockdown

due to the fact that the normality is in question and the sample size is really small, the extension of the univariate Kruskal-Wallis's test with the permutation version is conducted by means of *'MNM'* R package.

Looking at the results of pre-lockdown phase, we found that the groups are different from each other and the main discrimination is ascribable to the PM_{10} concentrations. Furthermore, it seems that there could be indications of significance as well regarding the benzene because the p-value is 0.186 and by increasing the sample size we could reject the homogeneity in this pollutant. Conversely, the multivariate test is not able to distinguish the two groups in the lockdown period. In fact, the p-value for the multivariate test is equal to 0.30. However, when we carry out the univariate tests, we record significant differences between the groups in relation to the benzene. This appreciation is also corroborated by the visual inspections of Figures 7 and 8. A possible explanation for these results can be found in the simulation study performed by [3] where it was shown that these approaches tend to be very conservative for small sample sizes.

5 Conclusions

Recent studies suggested that lockdown measures, adopted by the most hard-hit countries around the world in the Spring of 2020 to prevent the spread of COVID-19, have had a positive significant impact on air quality. In this paper, a novel approach for multivariate functional analysis of variance for independent measures is presented, as a methodology for a more effective understanding of the impact of lockdown on four critical air pollutants, measured in five monitoring sites in the urban area of Chieti-Pescara (Central Italy). Being a powerful

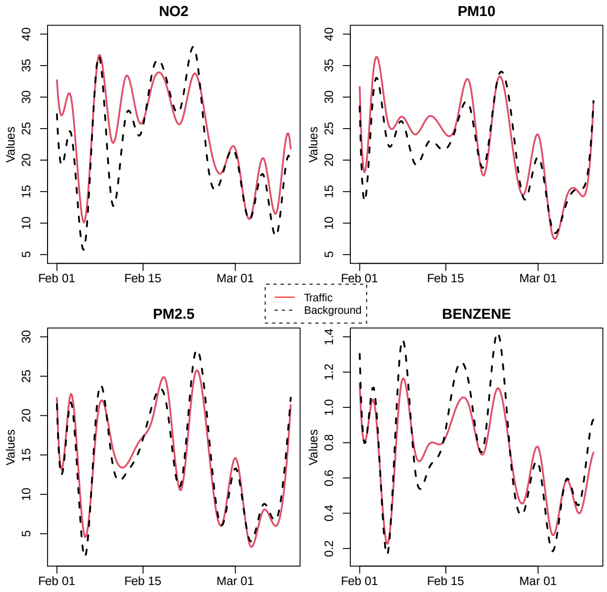


Figure 7: Mean function per pollutant of each group before lockdown

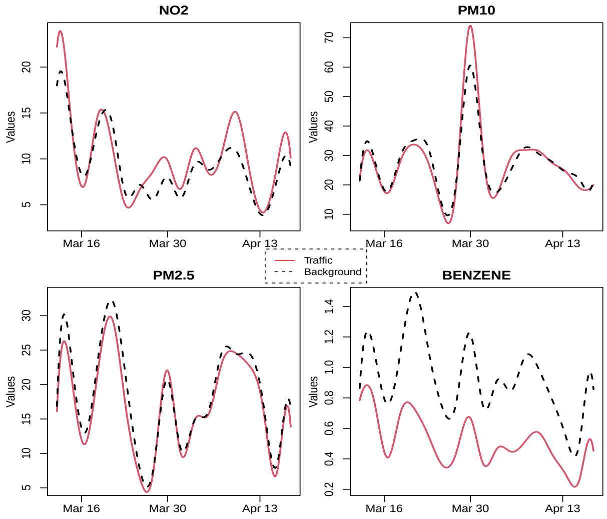


Figure 8: Mean function per pollutant of each group during lockdown

approach to modelling temporal observations, which is complementary to the usual time series techniques, the FDA allowed us to reconstruct the temporal profiles of the studied pollutants for the lockdown and unlock phases in each measuring station. We have found significant reduction in NO_2 levels during the lockdown period albeit some differences in magnitude are recorded according to the monitoring station. These results are in line with the findings of other published studies on this topic [26, 8, 18, 24] in which significant NO_2 reductions for different locations have been determined. Unlike the NO_2 pollutant, for particular matter, that is for PM_{10} and $PM_{2.5}$, the monitoring stations experienced an increase during the quarantine weeks. Besides, less clear was the impact of lockdown on benzene levels: the concentrations of this pollutant were smaller in the traffic stations while an increasing trend was observed in the background measuring sites. Equally important was to determine if these differences were statistically significant. In this respect, the functional analysis of variance has proven to be beneficial to monitoring the evolution of air quality before and during the lockdown tenure and to assessing the equality of mean functions of individual groups, individuated according to the location of measuring sites. The considered FANOVA approaches based on basis expansion of sample curves, dimension reduction by using FPCA of pollutants curves and testing homogeneity on the vector of the most explicative principal component scores have made this analysis feasible providing contrasted evidence to reject the null hypothesis of equality in the mean functions of all pollutants, both in the time frame considered and the localization of monitoring stations.

In general, the FDA framework has provided a valid understanding and knowledge of the temporal behaviour of air pollutants in a kind of controlled experiment such that offered by the lockdown. The COVID-19 restrictions reduced the anthropogenic emissions and created an “unprecedented scenario” in which the source of road traffic has been drastically dropped out. We believe that the results of this study are of interest for environmental protection agencies involved in developing policies to achieve air quality improvements, encouraging them to establish mechanism to reduce pollution emissions.

References

[1] C. Acal, A. M. Aguilera, and M. Escabias. New Modeling Approaches Based on Varimax Rotation of Functional Principal Components. *MATHEMATICS-BASEL*, 8(11):2085, 2020.

[2] A. Agarwal, A. Kaushik, S. Kumar, and R.K Mishra. Comparative study on air quality status in Indian and Chinese cities before and during the COVID-19 lockdown period. *Air Qual Atmos Health*, 13:1167–11178, 2020.

[3] A. M. Aguilera, C. Acal, M. C. Aguilera-Morillo, F. Jimènez-Molinos, and J. B. Roldàn. Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Math Comput Simul*, 186:41–54, 2021.

[4] A. M. Aguilera and M. C. Aguilera-Morillo. Penalized PCA approaches for B-spline expansions of smooth functional data. *Appl Math Comput*, 219(14):7805–7819, 2013.

[5] Aguilera, A. M. and Escabias, M. and Valderrama, M. J. Forecasting binary longitudinal data by a functional PC-ARIMA model. *Comput Stat Data An*, 52(6):3187–3197, 2008.

[6] M.C. Aguilera-Morillo, A.M. Aguilera, and M. Durban. Prediction of Functional Data with Spatial Dependence: a Penalized Approach. *Stoch Env Res Risk A*, 31:7–22, 2017.

[7] M. F. Bashir, B. Ma, Bilal, B. Komal, M. A. Bashir, D. Tan, and M. Bashir. Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci Total Environ*, 728:138835, 2020.

[8] J. D. Berman and K. Ebisu. Changes in U.S. air pollution during the COVID-19 pandemic. *Sci Total Environ*, 739:139864, 2020.

[9] D. C. Carslaw and K. Ropkins. openair - An R package for air quality data analysis. *Environ Model Softw*, 27-28:52–61, 2012.

[10] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Comput Stat Data Anal*, 47(1):111–122, 2004.

[11] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J Multivariate Anal*, 12(1):136–156, 1982.

[12] J. C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Ann Insee*, 15:3–101, 1974.

[13] M. Escabias, A. M. Aguilera, and M. J. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics*, 16(1):95–107, 2005.

[14] M. Escabias, M. J. Valderrama, A. M. Aguilera, M. E. Satofimia, and M. C. Aguilera-Morillo. Stepwise selection of functional covariates in forecasting peak levels of olive pollen. *Stoch Env Res Risk A*, 27(2):367–376, 2013.

[15] J. Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997.

[16] F. Ferraty and P. Vieu. *Nonparametric functional data analysis. Theory and practice*. Springer Verlag, New York, 2006.

[17] H. O. Gao and D. A. Niemeier. Using functional data analysis of diurnal ozone and NOx cycles to inform transportation emissions control. *Transp Res D Transp Environ*, 13(4):221–238, 2008.

[18] S. Gautam and U. Trivedi. Global implications of bio aresol in pandemic. *Environ Dev Sustain*, 22:3861–3865, 2020.

[19] T. Gorecki and L. Smaga. Comparison of tests for the one-way anova problem for functional data. *Comput Stat*, 30(4):987–1010, 2015.

[20] T. Gorecki and L. Smaga. Multivariate analysis of variance for functional data. *J Appl Stat*, 44(12):2172–2189, 2017.

[21] S. Hörmann, L. Kidziǹski, and M. Hallin. Dynamic functional principal components. *J R Stat Soc B*, 77(2):319–348, 2015.

[22] J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Comput Stat Data Anal*, 71:92–106, 2014.

[23] K. D. Kanniah, N. A. Kamarul Zaman, D. G. Kaskaoutis, and M. T. Latif. COVID-19's impact on the atmospheric environment in the Southeast Asia region. *Sci Total Environ*, 736:139658, 2020.

[24] A. Kerimray, N. Baimatova, O.P. Ibragimova, B. Bukenov, P. Pavel, and F. Karaca. Assessing air quality changes in large cities during COVID-19 lockdowns: The impacts of traffic-free urban conditions in Almaty, Kazakhstan. *Sci Total Environ*, 730:139179, 2020.

[25] I.R. Lima, G. Cao, and N. Billor. Robust simultaneous inference for the mean function of functional data. *TEST*, 28:785–803, 2019.

[26] S. Mahato, S. Pal, and K.G. Ghosh. Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. *Sci Total Environ*, 730:139086, 2020.

[27] P. Martinez-Camblor and N. Corral. Repeated measures analysis for functional data. *Comput Stat Data Anal*, 55(12):3244–3256, 2011.

[28] J. Martorell-Marugán, J. A. Villatoro-García, A. García-Moreno, R. López-Domínguez, F. Requena, J. J. Merelo, M. Lacasaña, J. de Dios Luna, J. J. Díaz-Mochón, J. A. Lorente, and P. Carmona-Sáez. DatAC:A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. *Sci Total Environ*, 750:141424, 2021.

[29] H. Oja. *Multivariate Nonparametric Methods with R.* Springer Science & Business Media, New York, 2010.

[30] N.L. Olsen, A. Pini, and S. Vantini. False discovery rate for functional data. *TEST (in press)*, 2021.

[31] A. Park, S. Guillas, and I. Petropavlovskikh. Trends in stratospheric ozone profiles using functional mixed models. *Atmos Chem Phys*, 13(22):11473–11501, 2013.

[32] U. K. Pata. How is COVID-19 affecting environmental pollution in US cities? Evidence from asymmetric Fourier causality test. *Air Qual Atmos Health*, 13:1149–1155, 2020.

[33] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[34] J.O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer-Verlag, New York, 2009.

[35] J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag, New York, 2002.

[36] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis, second ed.* Springer-Verlag, New York, 2005.

[37] A. Schmutz, J. Jacques, and C. Bouveyron. *funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces*, 2019. R package version 2.3.0.

[38] A. Schmutz, J. Jacques, Bouveyron C., L. Chéze, and P. Martin. Clustering multivariate functional data in group-specific functional subspaces. *Comput Stat*, 35:1101–1131, 2020.

[39] N. Shaadan, S.M. Deni, and A.A. Jemain. Assessing and Comparing PM10 Pollutant Behaviour using Functional Data Approach. *Sains Malays*, 13(22):11473–11501, 2013.

[40] K. Shehzad, M. Sarfraz, and S. G. M. Shah. The impact of COVID-19 as a necessary evil on air pollution in India during the lockdown. *Environ Pollut*, 266:115080, 2020.

[41] Q. Shen and J. Faraway. An F test for linear models with functional responses. *Stat Sin*, 14(4):1239–1257, 2004.

[42] P. Sicard, A. De Marco, E. Agathokleous, Z. Feng, X. Xu, E. Paoletti, J. J. D. Rodriguez, and V. Calatayud. Amplified ozone pollution in cities during the COVID-19 lockdown. *Sci Total Environ*, 735:139542, 2020.

[43] B. W. Silverman. Smoothed Functional Principal Component Analysis by Choice of Norm. *Ann Stat*, 24(1):1–24, 1996.

[44] L. Smaga. Repeated measures analysis for functional data using Box-type approximation with applications. *REVSTAT*, 17(4):523–549, 2019.

[45] L. Smaga. A note on repeated measures analysis for functional data. *AStA Adv Stat Anal*, 104(1):117–139, 2020.

[46] M.J. Valderrama, F.A. Ocaña, A.M. Aguilera, and F.M. Ocaña Peinado. Forecasting pollen concentration by a two-step functional model. *Biometrics*, 66, 2010.

[47] WHO 2020. Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-corona virus-2019.

[48] M. A. Zabrano-Monserrate and M. A. Ruano. Has air quality improved in Ecuador during the COVID-19 pandemic? A parametric analysis. *Air Qual Atmos Health*, 13:929–938, 2020.

[49] J.T. Zhang. *Analysis of Variance for Functional Data.* CRC Press, 2014.

[50] J.T Zhang, M.Y. Cheng, C.J. Tseng, and H.T. Wu. A new test for one-way ANOVA with functional data and application to ischemic heart screening. *Comput Stat Data An*, 132:3–17, 2019.

A7 COVID-19 data imputation by multiple function on function principal component regression

- Acal, Christian; Escabias, Manuel; Aguilera, Ana M.; Valderrama, Mariano (2021)
- COVID-19 data imputation by multiple function on function principal component regression
- *Mathematics*, in press

mathematics
an Open Access Journal by MDPI

Mathematics			
JCR Year	Impact Factor	Rank	Quartile
2019	1.747	28/235	Q1

Abstract

The aim of this paper is the imputation of missing data of COVID-19 hospitalized and intensive care curves in several Spanish regions. Taking into account that the curves of cases, deceases and recovered people are completely observed, a function-on-function regression model is proposed to estimate the missing values of the functional responses associated with hospitalized and intensive care curves. The estimation of the functional coefficient model in terms of principal components regression with the completely observed data provides a prediction equation for the imputation of the unobserved data for the response. An application with data from the first wave of COVID-19 in Spain is developed after properly homogenizing, registering and smoothing the data in a common interval so that the observed curves become comparable. Finally, Canonical Correlation Analysis on the functional principal components is performed to interpret the relationship between hospital occupancy rate and illness response variables.

1 Introduction

The virus SARS-CoV-2 has been the main global concern ever since its start at the end of 2019 in China. Its rapid propagation has put on alert all areas of society, not only the field of medicine. Nevertheless, a year and half later from the beginning of the pandemic, the virus incidence does not seem to decrease and the number of deaths continues its upward trend throughout the world. To get some idea of extremely negative impact of the pandemic, Coronavirus Disease (COVID-19) has caused a total of 2.780.266 deaths over the planet as of 28/03/2021 according to the real-time database developed by Johns Hopkins University [19]. At the same time, another crucial topic derived from the illness is the economic crisis that devastates all countries. For instance, the unemployment rate is up 5.1% in last three months of 2020 in UK according to official data.

In order to combat this terrible situation, there is a great need to understand the development of the pandemic. Knowing its behaviour will enable correct decision making to mitigate the spread of the virus and to recover people's daily life as soon as possible. On this matter, the science community is focusing all its efforts on developing new techniques capable of modelling and predicting the evolution of the COVID-19. The main variables of interest that gauge how the epidemiological situation stands in a country are the number of positive, recovered and deceased cases. Another important indicator is the number of people who are hospitalized or in intensive care units. From a mathematical viewpoint, many authors have already attempted to tackle these variables from different statistical perspectives. A new Bayesian indicator is introduced in [10] to forecast the beginning of a new wave. In [37], semi-empirical models based on the logistic map are considered in order to predict the variables in different phases of the pandemic in Spain. Likewise, [2] apply SIR models to analyse the trend of the disease over the world and more specifically, in India. These variables are also addressed from the time series design by considering quasi-Poisson regression and two piece scale mixture normal distribution when there is lack of symmetry in the error's distribution in [52, 35], respectively. Additionally,

[56] make an exhaustive comparison of five deep learning methods to forecast the number of new cases and recovered cases in Italy, Spain, France, China, USA and Australia. Regarding the role of the environmental conditions in the evolution of the illness, [44] study if the number of cases in China is connected with the daily average temperature and relative humidity through a generalized additive model. On this point, [11] show how the choice of the spatio-temporal model may affect the relation between the spread of the virus and certain environmental conditions. Information theory metrics are also used to understand how time series associated with the pandemic are interconnected or causally related each other [55]. In addition, how the incubation period distribution could vary by age and gender is investigated in [42]. On the other hand, a new family of distributions is introduced in [36] to model daily cases and deaths in Egypt and Saudi Arabia.

Taking into account the nature of the variables of interest, an approach based on Functional Data Analysis (FDA) is proposed in the current paper for data imputation. FDA is a modern branch of the statistics that aims to analyse the information coming from curves or functions that evolve over time, space or other continuous argument. Under this definition, it is clear that the number of COVID-19 positive, recovered, deceased, hospitalized and intensive care people come from the observation of functional variables. FDA is usually applied in many areas of knowledge as Biosciences, Environment, Economy, Chemometrics and Electronics, among others. A detailed review about the most important FDA methodologies, applications and computational aspects can be seen in books [48, 47, 46, 26, 30]. In this regard, some works focused on revealing complex patterns of COVID-19 illness from a FDA viewpoint have been developed. Functional Principal Component Analysis (FPCA) and functional time series approaches based on dynamic FPCA are applied in [51] for explaining variability and predicting COVID-19 confirmed and death cases in the United States. On the other hand, a new Varimax rotation approach for FPCA is introduced in [1] to better interpret the main modes of variability in COVID-19 confirmed cases in the first wave in Spain. Time-varying FDA methods for modeling the cumulative COVID-19 curves of cases by pooling data across countries are applied in [12]. A multivariate FDA approach has also been considered for spatio-temporal prediction of COVID-19 mortality counts in Spain [53].

All statistical models require complete and high quality data to be able to provide accurate predictions, but, unfortunately, neither of these aspects are normally fulfilled during a pandemic. In the first wave of COVID-19 in Spain, a change in the way of recording data in some Autonomous Communities produced incomplete data in hospitalized and intensive care curves. In this paper, a functional linear regression model is proposed for the imputation of these missing curves so that complete data are available to be able to estimate the predictive models with guarantees. Although there are many works related to the imputation in multivariate data [34, 27], there is a lot to be done in the functional framework yet. A novel approach for multiple imputation based on functional mixed effects models was proposed by [29] in a longitudinal data context. Different solutions for scalar-on-function regression with missing observations in the response are considered in [25, 32, 33, 15, 24]. Besides, an extension to multiple functional regression imputation that handles both scalar and functional response variables related to EEG data is proposed in [14]. Likewise, different FDA imputation methods under sparse and irregular functional data settings are performed in [49]. The extension of the function-on-function linear regression (FFLR) model [5, 54, 45] to the case of multiple functional predictors is proposed in this paper for estimating the curves of hospitalized and intensive care people (functional responses) from the curves of confirmed, deceased and recovered cases (functional predictors).

In addition to this introduction, the manuscript scheme consists of a description of the data where the process of homogenization, registration and smoothing of the sample curves is detailed in order to make them to be comparable (Section 2). The theoretical framework on multiple function-on-function linear regression and the imputation procedure based on principal components regression appear in Section 3. An application on COVID-19 data in the Spanish Autonomous Communities during the first wave of the pandemic is developed in Section 4. Finally, Section 5 contains a discussion about the results obtained throughout this paper.

2 Data homogenization, registration and smoothing

Spain is organized administratively in autonomous communities (ACs) or territorial governments that have transferred health affairs. This territorial organization consists of 17 ACs plus two autonomous cities (Ceuta and Melilla) located on the African continent and which have been excluded from the analyses presented here because they have no exclusive competences in the organization of health care assumed by the Spanish government. The 17 ACs are, in alphabetical order: Andalucía, Aragón, Asturias, Islas Baleares, Islas Canarias, Cantabria, Castilla La Mancha, Castilla León, Cataluña, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La Rioja and Valencia. The population is highly variable between different ACs. While Madrid, Catalunya and Andalucía have more than six, seven and eight million inhabitants, respectively, (6,663,394, 7,675,217 and 8,414,240), La Rioja has approximately three hundred thousand inhabitants (316,798).

The first wave of the Covid-19 pandemic in Spain occurred between February 2nd and April 27th, 2020. In those early days of the pandemic, Spanish authorities published daily and accumulated data of the evolution of the pandemic in Spain based on the information communicated by the different ACs. Specifically, the data published daily correspond to the following variables: number of confirmed (positive) cases, hospitalized people, people in intensive care units (ICUs), recovered people and deceased persons. The observed data for some of the ACs can be seen in Figure 1. The problem that arose, and gave rise to this work, is that some ACs (Castilla La Mancha, Castilla León, Madrid and Galicia) modified the recording of the data associated with people in ICU and hospitalized people from a specific day (see Figure 2). The mathematical action against COVID-19 of the Spanish Mathematics Committee called for the development of a meta-predictor (collaborative prediction) based on the predictions from different models/algorithms, contributed by interested researchers, which builds optimized combinations of them, disaggregated by ACs. Therefore, the imputation of the missing hospitalized and ICU data is fundamental to building forecasting models to be able to provide optimal predictions of the evolution of the pandemic through these variables. In order to solve this problem, a functional regression model is proposed in this paper to estimate the expected form of the missing accumulated data of ICU admissions and hospitalizations from the observed accumulated data of cases, deaths and recoveries.

From now on, the time evolution of COVID-19 cases, deceases, recoveries, hospitalizations and ICU admission will be considered as functional variables that will be denoted as $X_1(t)$, $X_2(t)$, $X_3(t)$, $Y_1(t)$ and $Y_2(t)$, respectively (the $X-$variables will be considered as predictors and the $Y-$variables as responses in the functional regression models). The observed data are the number of daily cumulative informed values of these five functional variables for the seventeen ACs in Spain from 20/02/2020 to 27/04/2020. Then, a random sample of curves $\{(x_{ij}(t), y_{ik}(t)) : i = 1\ldots, 17; j = 1, 2, 3; k = 1, 2\}$ observed daily are available.

Before carrying out the functional analysis of the data it becomes necessary a data registration given the absence of uniformity in the publication of the observations. This means that in the same functional variable the first day with available data and the number of discrete observations in each AC are different. For example, in Andalucía the first recorded data of hospitalized persons was on March, 10th in which 32 hospitalized people were registered and for this variable there were 49 discrete observations in this AC. On the other hand, in Madrid the first recorded data of hospitalized persons was on March, 12th in which 1304 hospitalized people were registered, and the number of discrete observations in this AC was 47. However, the curves of positive cases recorded 62 and 63 discrete

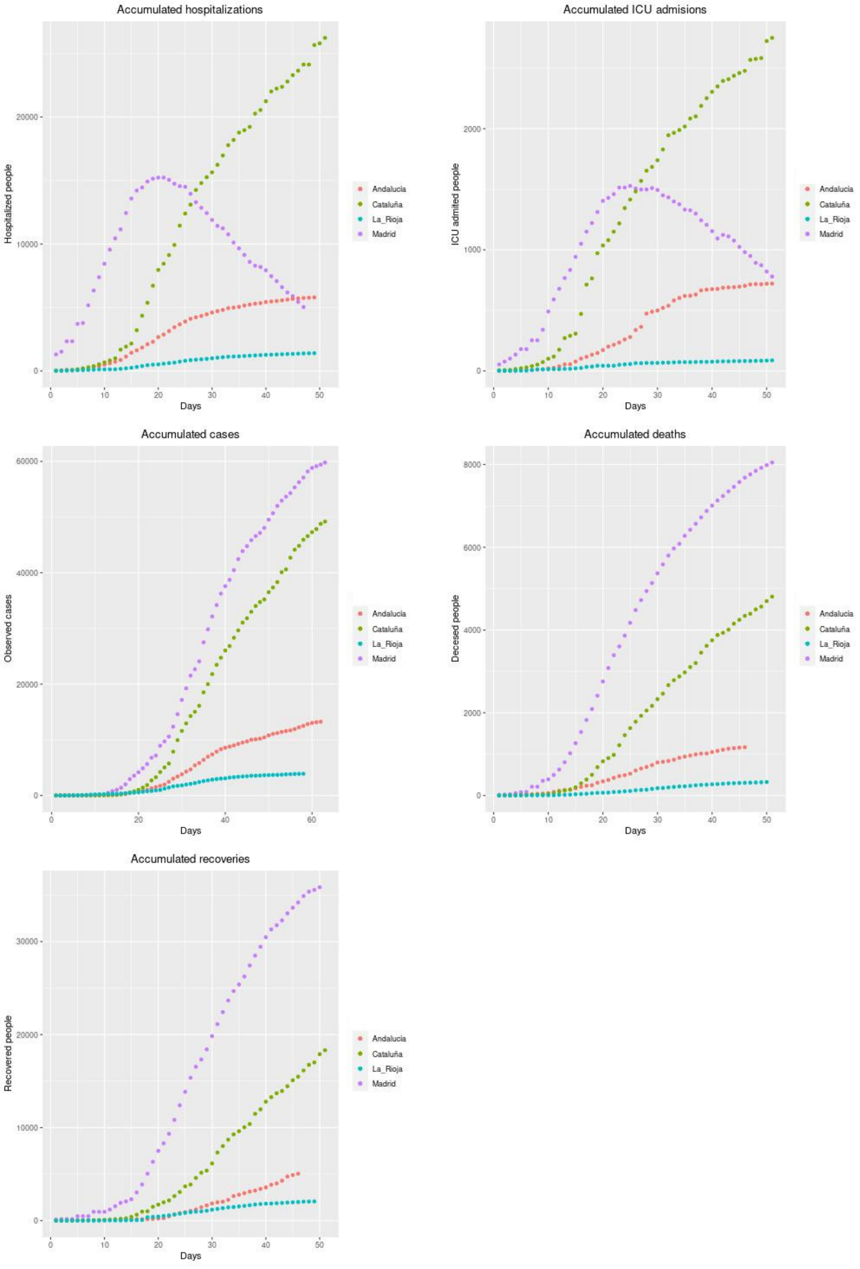


Figure 1: Discrete daily observations of accumulated positive cases, deaths, hospitalizations, ICU admissions and recoveries in Madrid, Andalucía, Cataluña and La Rioja.

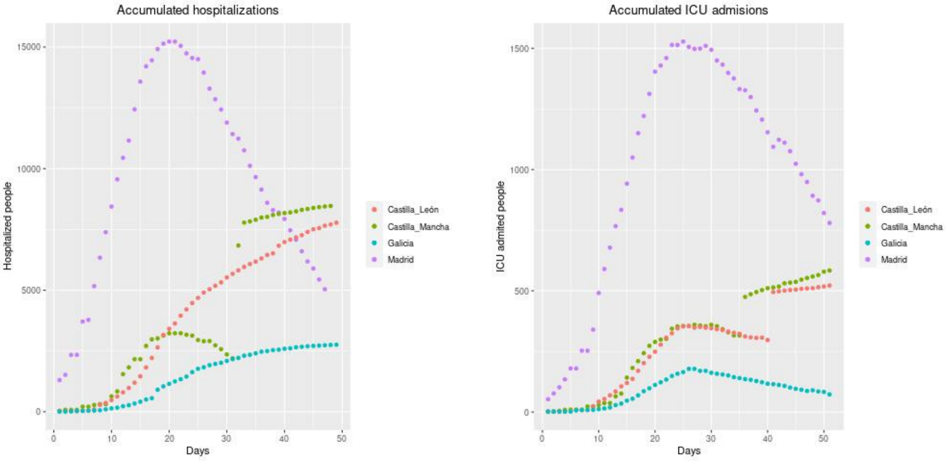


Figure 2: Discrete daily observations of accumulated hospitalizations and ICU admissions in Madrid, Castilla La Mancha, Castilla León and Galicia.

observations in Andalucía and Madrid, respectively. In addition, the population size of each AC could influence the adjustments of the proposed models, as larger numbers of cases have been observed in larger communities, and the different size of the population of the different ACs makes impossible to compare data between them. In order to avoid both problems, the number of cases per 10000 inhabitants is considered and the first observation for each curve corresponds to the day that exceeds for the first time the maximum of the first reported values, discarding the previous ones.

After data homogenization, the period of observation and the number of discrete observations of each functional variable for each AC continue being different. An important constraint of FDA methods is that all sample curves of a functional variable must be observed in the same domain. Classic solutions to this problem are based on registration of all curves in a common interval (see [48]). In this paper we propose to register all curves in the interval $[0,1]$ by applying the FDA methodologies to the synchronized curves defined by

$$x^*_{ij}(u) = x_{ij}(T_{ij-start} + u(T - T_{ij-start})),$$

$$y^*_{ik}(u) = y_{ik}(T_{ik-start} + u(T - T_{ik-start})) \ \forall u \in [0,1],$$

where $[T_{ij-start}, T]$ and $[T_{ik-start}, T]$ are the observed domains for the $i-$th predictor and the $k-$th response curves, respectively $(i = 1, \ldots, 17; j = 1, 2, 3; k =$

$1, 2)$. From now on, and by abuse of notation that helps simplify the exposition, x_{ij} and y_{ik} will represent the registered curves.

2.1 From discrete daily observations to curves

Although functional data are sets of curves, their true functional form is unknown and the recorded data are observations of each curve at a finite collection of time points. Then, the first step in FDA is to reconstruct the functional form of the curves from the observed discrete data.

There are different approaches to the processing of functional data among which we can highlight the classic ones based on basis representation of the curves [48] and the ones based on local-polynomial regression [26]. In this paper, basis expansions of the curves are considered by assuming that each of the functional variables $(X_1, X_2, X_3; Y_1, Y_2)$ generating the sample curves, are smooth stochastic processes with trajectories in the space $L^2([0, 1])$ of squared integrable functions in the interval $[0, 1]$. In what follows, the basis expansion approach is illustrated for a random sample of a functional variable defined on a general interval T. In our data set, this procedure must be performed on each of the five considered functional variables for which the type and dimension of the basis could be different.

Consider a random of sample curves $\{x_i(t) : i = 1, \ldots, n;\ t \in T\}$ from a functional variable X with values in $L^2(T)$, and let us assume that noisy observations $\mathtt{x}_{ik}$ are available for each curve at a set of time knots $t_{i1}, t_{i2}, \ldots, t_{im_i} \in T$, that is

$$\mathtt{x}_{ik} = x_i(t_{ik}) + \epsilon_{ik} \quad i = 1, \ldots, n;\ k = 1, \ldots, m_i.$$

Let us also suppose that the sample curves belong to a finite-dimensional space generated by a basis of functions $\{\phi_1(t), \ldots, \phi_p(t)\}$. Therefore, each curve of the functional data set admits a basis representation in the form

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t),\ i = 1, \ldots, n. \tag{1}$$

The functional form of each curve is then determined by the vector of its basis coefficients $a_i = (a_{i1}, \ldots, a_{ip})'$, that can be estimated in different ways, with least squares approximation being the most common method that provides the following estimation: $\hat{a}_i = (\Phi_i'\Phi_i)^{-1}\Phi_i'x_i$, where $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p},\ j = 1, \ldots, p,\ k = 1, \ldots, m_i$.

The type of basis must be selected according to the characteristics of the curves in the functional data set. The most common basis are B-splines and trigonometric functions (see for example [48]). The former generates spaces of spline functions, piecewise polynomials functions that are smoothly joined and have good local behaviour. The latter provides suitable spaces for periodic functions. Many other bases have been used in practice, such as bases of wavelets which are more appropriate for curves with discontinuities and sharp spikes. An application of wavelet approximation from sample curves of lupus and stress level was developed in [5]. A robust estimation of the mean function, together with a simultaneous confidence band, based on polynomial spline estimation is developed in [31].

In this paper, a basis of cubic B-splines of dimension ten with equally spaced knots has been used to approximate the five samples of curves of COVID-19 from their daily discrete data. Least squares approximation was performed on each curve for estimating the basis coefficients. The cubic regression splines of all curves considered here can be seen in Figure 3.

3 Functional linear regression imputation with missing values in the response

Motivated by the imputation of the missing curves of COVID-19 hospitalized and intensive care people, a functional linear regression model with functional response and several functional predictors is proposed in this paper. The general formulation of this multiple function-on-function linear regression (MFFLR) model and its estimation in terms of functional principal components regression are summarized in this section.

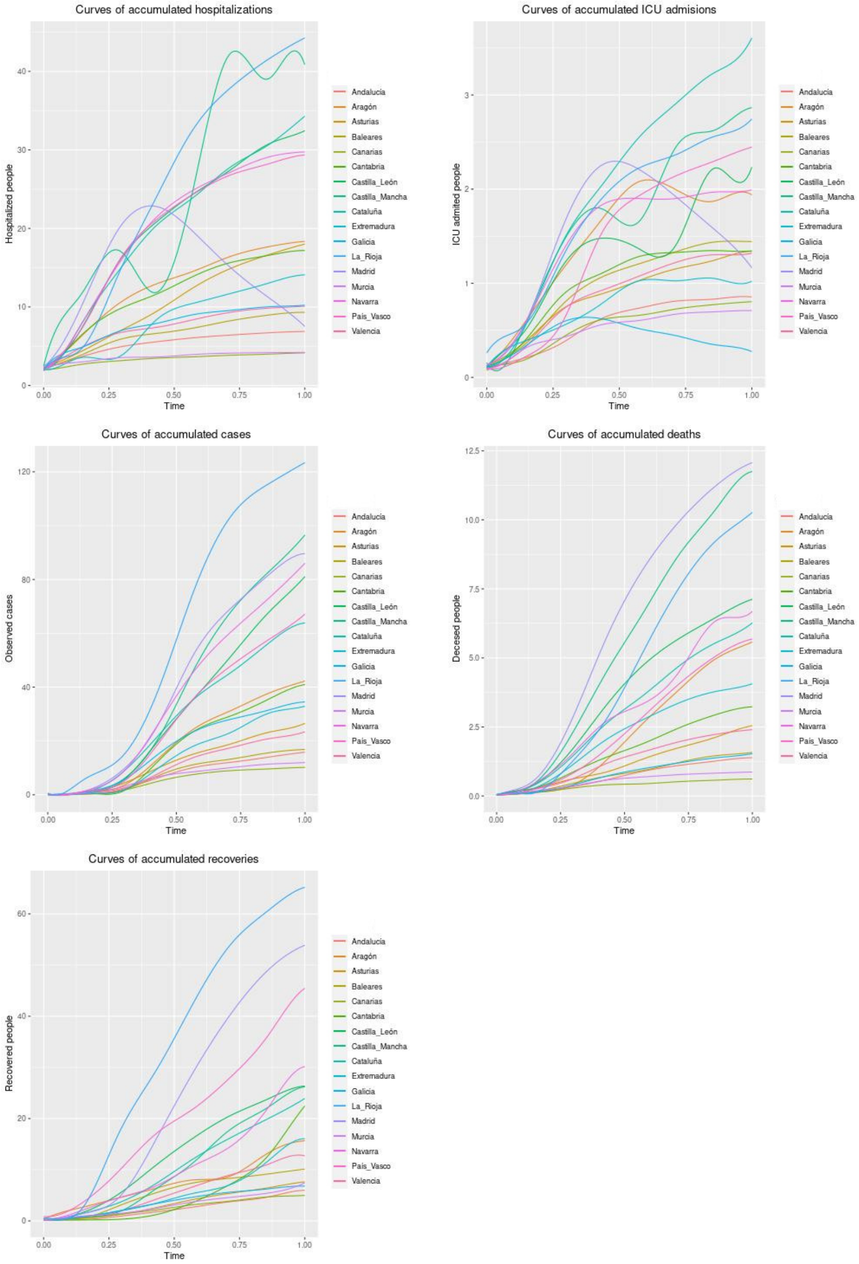
3.1 Multiple function-on-function linear model

The multiple function-on-function linear model allows to estimate a functional response Y from a vector of J functional predictor variables denoted by $X = (X_1, \ldots, X_J)'$. Let us consider a random sample from (X, Y) denoted by $\{(x_i, y_i) : i = 1, \ldots, n\}$ with $x_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})'$, and let us assume that all functional variables take values on the Hilbert space $L^2(T)$ of the squared integrable functions on the interval T, with the usual inner product defined by $< f, g >= \int_T f(t)g(t)dt$, $\forall t \in T$.

Then, the functional linear model is formulated as

$$y_i(t) = \alpha(t) + \sum_{j=1}^{J} \int_T x_{ij}(s)\beta_j(s,t)ds + \varepsilon_i(t),\ i = 1, \ldots, n, \tag{2}$$

where $\alpha(t)$ is the intercept function, $\beta_j(s,t)$ are the J coefficient functions and $\epsilon_i(t)$ are independent functional errors.


Curves of accumulated hospitalizations

Curves of accumulated ICU admisions

Curves of accumulated cases

Curves of accumulated deaths

Curves of accumulated recoveries

Figure 3: Curves of accumulated positive cases, deaths, recoveries, hospitalizations and ICU admissions.

Model 2 can be written in matrix form as

$$y_i(t) = \alpha(t) + \int_T x_i(s)'\beta(s,t)ds + \varepsilon_i(t),\ i = 1, \ldots, n,$$

where $x_i(s) = (x_{i1}(s), x_{i2}(s), \ldots, x_{iJ}(s))'$ and $\beta(s,t) = (\beta_1(s,t), \beta_2(s,t), \ldots, \beta_J(s,t))'$.

This expression considers that all functional variables are defined in the same interval T, but this is not a restriction and the model can be easily generalized for different domains in each of the functional variables. The estimation of this model is an ill-posed problem that is usually solved by least squares penalized approaches and basis expansion of functional parameters and/or sample curves [48]. Some of the basis expansion approaches reduce the model to a multivariate linear model for the matrix of response basis coefficients in terms of the matrix of predictors basis coefficients. The main problem is that this multivariate model is affected of high multicollinearity which causes an inaccurate estimation of the parameters. Despite the good predictive ability of the model, this fact makes its interpretation more difficult. The most studied solutions avoid the need for cross-validation to estimate the penalty parameter by reducing the problem to linear regression on uncorrelated predictor variables. Approaches based on functional PCA [13, 20, 38, 9, 22, 4] and functional Partial Least Squares (PLS) [43, 21, 6, 17, 23, 3] have been widely studied in literature in the context of different functional regression models.

In this paper, a principal components regression approach is considered. It can be seen as an extension of the principal components prediction models developed in [7, 8] for predicting a functional variable in a future interval of time from its evolution in the past. In the so called PCP models, the functional response and the functional predictor correspond to the same functional variable but observed in different time periods. In the present approach, truncated principal component decompositions of the functional response and the functional predictors turn the functional linear model in a multivariate linear model in terms of a reduced set of response and predictor principal components.

3.2 Functional principal component regression

Let us consider the principal component decompositions of both, the response and the predictor functional variables, given by

$$x_{ij}(t) = \overline{x}_j(t) + \sum_{l=1}^{n-1} \xi_{il}^{x_j} f_l^{x_j}(t), \quad y_i(t) = \overline{y}(t) + \sum_{l=1}^{n-1} \xi_{il}^{y} f_l^{y}(t), \tag{3}$$

where the principal components scores are given by

$$\begin{aligned}
\xi_{il}^{x_j} &=< x_{ij} - \bar{x}_j, f_l^{x_j} >= \textstyle\int_T (x_{ij}(t) - \bar{x}_j(t)) f_l^{x_j}(t) dt, \\
\xi_{il}^{y} &=< y_i - \bar{y}, f_l^{y} >= \textstyle\int_T (y_i(t) - \bar{y}(t)) f_l^{y}(t) dt,
\end{aligned} \tag{4}$$

with the weight functions $f_l^{x_j}$ and f_l^y being the eigenfunctions of the sample covariance operators of $x_{ij}(t)$ and $y_i(t)$, respectively. The principal components scores are centered uncorrelated scalar variables with maximum variance given by the eigenvalues associated with their weight functions: $Var(\xi_{il}^{x_j}) = \lambda_l^{x_j},\ Var(\xi_{il}^{y}) = \lambda_l^{y}$.

Theoretical and asymptotic properties of FPCA for Hilbert-valued random functions were studied in [18, 16, 41, 28, 50]. In the case of a basis expansion for each functional variable (see Equation 1), each functional PCA is equivalent to multivariate PCA of the matrix $A\Psi^{1/2}$, with $A = (a_{ij})$ being the $n \times p$ matrix of basis coefficients and Ψ being the $p \times p$ matrix of inner products between basis functions, $\Psi = (\Psi_{ij}) =< \phi_i, \phi_j >,\ i, j = 1, ..., p$. The vector of basis coefficients of the the $l-$th PC weight function $f_l(t)$ is given by $b_l = \Psi^{-1/2} v_l$, where v_l is the $l-$th eigenvector of the sample covariance matrix of $A\Psi^{1/2}$ (see [40] for a detailed study).

The principal component decompositions given in Equation 3 turn the MFFLR Model 2 into a linear regression model for each PC of the functional response Y on all PCs of the functional predictors

$$\xi_{ik}^{y} = \sum_{j=1}^{J} \sum_{l=1}^{n-1} b_{kl}^{x_j} \xi_{il}^{x_j} + \varepsilon_{ik}, \quad i = 1, \ldots, n;\ k = 1, \ldots, n-1, \tag{5}$$

with the functional coefficients given by $\beta_j(s,t) = \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} b_{kl}^{x_j} f_k^{x_j}(s) f_l^{y}(t)$.

By truncating each principal component decomposition the following principal component multiple function-on-function linear regression (PC-MFFLR) model for the functional response is obtained:

$$\hat{y}_i(s) = \bar{y}(s) + \sum_{k=1}^{K} \hat{\xi}_{ik}^{y} f_k^{y}(s) = \bar{y}(s) + \sum_{k=1}^{K} \left(\sum_{j=1}^{J} \sum_{l \in L_{kj}} \hat{b}_{kl}^{x_j} \xi_{il}^{x_j} \right) f_k^{y}(s), \tag{6}$$

with $\hat{b}_{kl}^{x_j}$ being the linear least squared estimation of the regression coefficients b_{kl}.

Different selection model approaches have been developed to select the optimum PCs of each predictor variable (subsets L_{kj}) to be considered in Model 6 when it comes to estimating the first J PCs of the response variable. It is well known that PCs are ordered according to their explained variability and that the most explanatory components of the predictor variable might not be the most correlated with the response variable. In the case of the simple function-on-function linear model with only one predictor, a procedure that selects pairs of response/predictor PCs based on both, explained variability and correlation, was developed in [5]. A supervised version of FPCA that estimates the PCs by considering the correlation of the functional predictor and response variable was developed for the scalar-on-function regression model [39]. Usual selection models procedures based on stepwise and best subset regression combined with cross-validation can be adapted to this functional regression context.

3.3 Imputation of missing response curves

Let us consider that all the predictor variables X_j are completely observed and only the response variable Y has missing values. Let us assume without loss of generality that in the sample the first n values of the response are observed and the last m values are missing. That means that there are n complete observed curves for all variables and m incomplete observations that are missing values for the response.

In order to estimate the missing response curves, the parameters b_{kl} in Model 5 are estimated with the complete n sample curves of response and predictors. Then, the missing response curves $\{y_i^{miss}(s) : i = n+1, \ldots, n+m\}$ are estimated by computing the principal component scores of predictors $\{\xi_{il}^{x_j} : i = n+1, \ldots, n+m, l = 1, \ldots, n-1\}$ given by the Expression 4, and substituting them in the Equation 6. Then, the estimated PC-MFFLR model can be used to predict new values of the response Y on a test sample and to provide accurate interpretation of the relationship between the predictor and the response variables.

If the objective is to predict the response variable in a future interval, a regression model of type 6 could be estimated for predicting the response variable $Y(s)$ in the future interval of amplitude k denoted by $[T, T+k]$, in terms of the predictor variables $(X_1(t), \ldots, X_J(t), Y(t))$ in the past interval of time $[0, T]$. In the case of the COVID-19 data, the parameter k must be selected taking into account the average number of days it takes for a person to develop severe symptoms and need to be admitted to the hospital.

4 Covid-19 application results

Let us remember that the main aim of this paper is the imputation of hospitalized and intensive care curves for those ACs with missing data. To do it, multiple function-on-function linear regression approaches are developed here.

In addition, a canonical correlation analysis (CCA) is performed to interpret the relationship between variables related with hospital occupation (hospitalized and intensive care people) and illness response (positive, deceased and recovered people). Computational results were obtained with the free software R ('fda' and 'yacca' R-packages for FPCA and CCA, respectively).

4.1 Data imputation

The imputation problem is solved by applying a multiple function-on-function linear regression for each of the responses $Y_1(t)$ (hospitalized) and $Y_2(t)$ (intensive care) from the three functional predictors $X_1(t)$ (sick), $X_2(t)$ (deceased) and $X_3(t)$ (recovered). Both functional regression models are estimated from the data of the thirteen ACs with complete data (training sample). Then, the predictions for the four ACs with missing data (Castilla La Mancha, Castilla León, Galicia and Madrid) are used for data imputation.

The first step is the estimation of the functional PCs for each of the five functional predictors. As a result, the first PC explained almost all variability of the five predictors $(99.32\%, 98.73\%, 97.97\%, 98.59\%, 96.37\%$ for X_1, X_2, X_3, Y_1, Y_2, respectively). Figures 4 and 5 show the weight functions associated to each first PC, and the perturbations of the sample mean curves obtained by adding and subtracting a multiple of them. In order to obtain weight functions and PC scores much easier to interpret, two new functional Varimax rotation approaches were introduced in [1] with application to COVID-19 confirmed people.

After obtaining these functional principal components analysis, we consider a training sample composed by all the ACs except Castilla La Mancha, Castilla León, Galicia and Madrid, which will be considered as the prediction sample.

Taking into account that the first component of $X_1(t)$, $X_2(t)$ and $X_3(t)$ were revealed highly and significantly correlated with the first components of $Y_1(t)$ and $Y_2(t)$, meanwhile the other cross-correlations between PCs were not significant, the function-on function linear regression models were reduced to following linear models for the first PC of the response in terms of the first PC of each of the predictors:

$$\hat{\xi}_{i1}^{y_k} = \gamma_0 + \xi_{i1}^{x_1}\gamma_1^{y_k} + \xi_{i1}^{x_2}\gamma_2^{y_k} + \xi_{i1}^{x_3}\gamma_3^{y_k} + \varepsilon_i^{y_k}, \quad k = 1, 2;\ i = 1, \ldots, 17.$$

These models allow to accurately estimate the first component of $Y_1(t)$ and $Y_2(t)$ from the first components of $X_1(t)$, $X_2(t)$ and $X_3(t)$ with a determination coefficient of $R^2 = 0.9249$ and $R^2 = 0.7443$, respectively. Finally, the Karhunen-Loève expansion in terms of the predictor principal components, provides the following prediction equation for $Y_1(t)$ and $Y_2(t)$:

$$\hat{y}_{ik}(t) = \overline{y}_k(t) + \widehat{\xi}_{i1}^{y_k} f_1^{y_k}(t),\ k = 1, 2;\ i = 1, \ldots, 17. \tag{7}$$

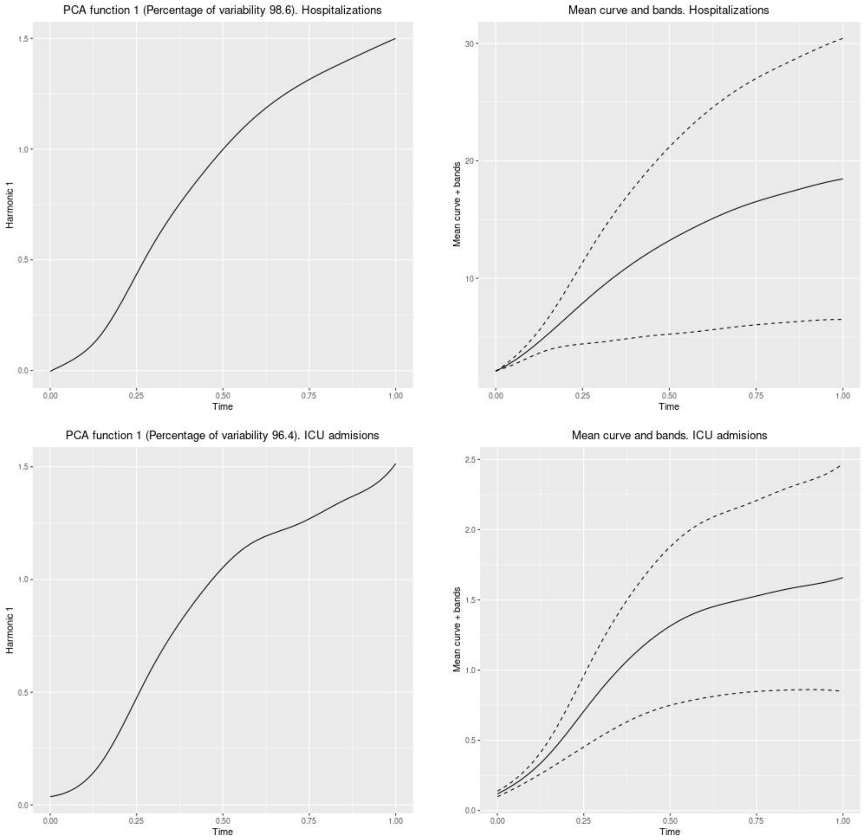


Figure 4: First weight PC functions, sample mean curves and the perturbations for each functional response: $\bar{y}_k \pm 2\sqrt{\lambda_1^{y_k}} f_1^{y_k}; \quad k = 1, 2.$

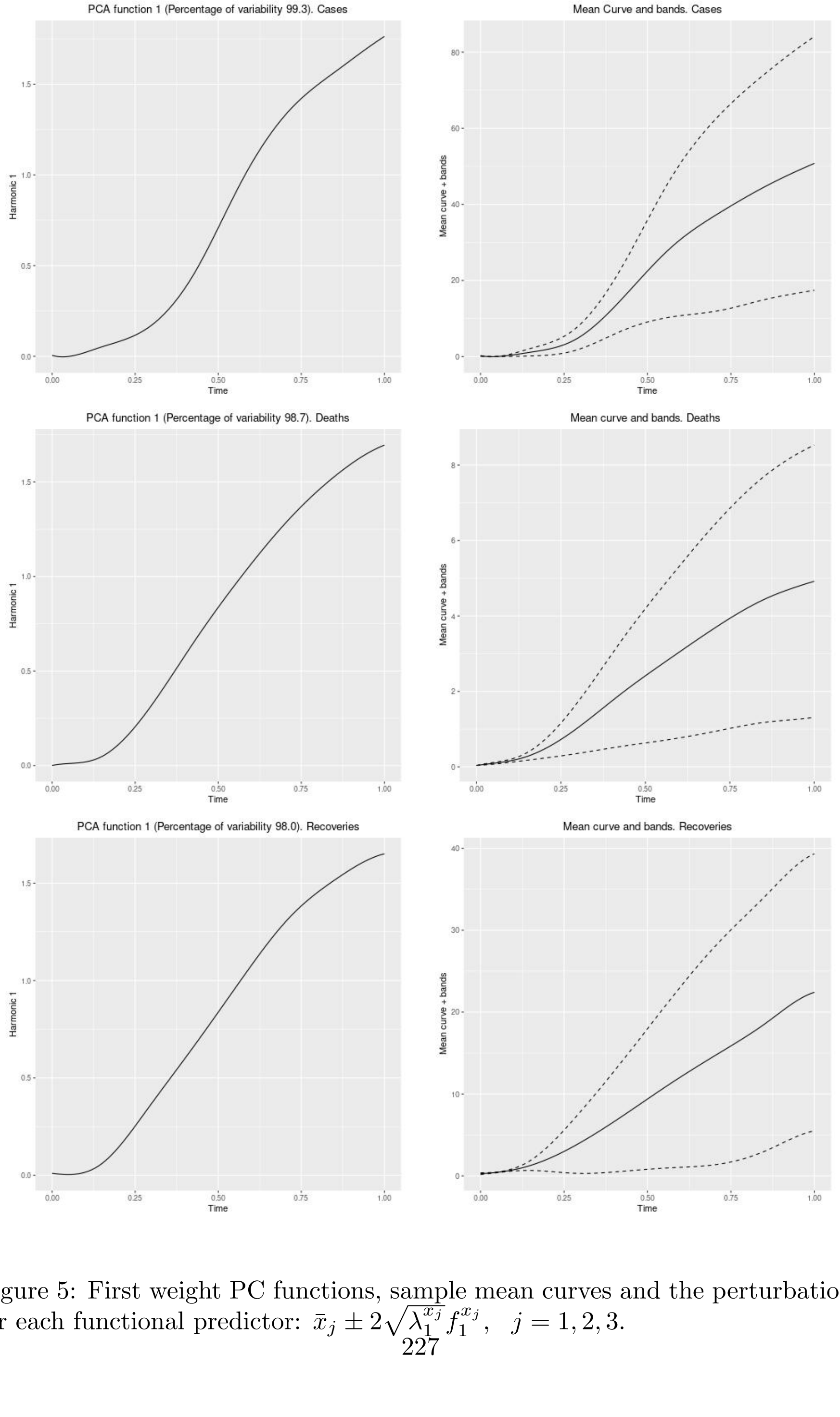


Figure 5: First weight PC functions, sample mean curves and the perturbations for each functional predictor: $\bar{x}_j \pm 2\sqrt{\lambda_1^{x_j}} f_1^{x_j}, \quad j = 1, 2, 3.$

AC	$RMSE(y_{i1})$	$RMSE(y_{i2})$
Andalucía	0.77577948	0.13645363
Aragón	1.51075014	0.16559390
Asturias	3.05564828	0.05135305
Islas Baleares	0.47397162	0.30168996
Islas Canarias	1.17563681	0.04168788
Cantabria	0.91993038	0.06896749
Catalunya	3.45656121	0.62176527
Valencia	0.92591220	0.04144271
Extremadura	3.30961380	0.59974926
Murcia	1.62014752	0.11596922
Navarra	1.26109742	0.20248242
País vasco	4.46884752	0.30765768
La Rioja	3.37692798	0.22644853

Table 1: Root mean squared prediction errors for Hospitalizations (y_{i1}) and ICU admissions (y_{i2}) curves in the different training ACs.

In order to evaluate the prediction ability of these models, the square root of the mean squared errors between observed and predicted curves are calculated by the expression

$$RMSE(y_{ik}) = \left(\int_0^1 (y_{ik}(t) - \widehat{y}_{ik}(t))^2 dt \right)^{\frac{1}{2}} \quad k = 1, 2; \; i = 1, \ldots, 13.$$

These results can be seen in Table 1 where it can be observed that the predictions for ICU admission curves are more accurate. Some of the observed and estimated training curves can be seen in Figures 6 and 7 next to confidence bands for the predicted curves. These confidence bands are obtained by pointwise confidence intervals, computed for each fixed time point t_p as follows:

$$\widehat{y}_{ik}(t_p) \pm 2 \times \widehat{SE}(\widehat{y}_{ik}(t_p)), \; k = 1, 2,$$

where $\widehat{SE}(\widehat{y}_{ik}(t_p)) = \widehat{SE}(\widehat{\xi}_{ik}^{y_k}) f_1^{y_k}(t_p), \; k = 1, 2$ with $\widehat{SE}(\widehat{\xi}_{ik}^{y_k})$ being the standard error of the PC prediction given by the corresponding multiple linear regression fit.

Finally, the expected curves provided by the regression models in Equation 7 for hospitalizations and ICU admissions in the badly recorded ACs, next to their confidence bands and observed curves, are drawn in Figures 8 and 9.

The prediction of the missing curves by using the PC-MFFLR considered models, provides a pointwise estimation of hospitalizations and ICU admissions

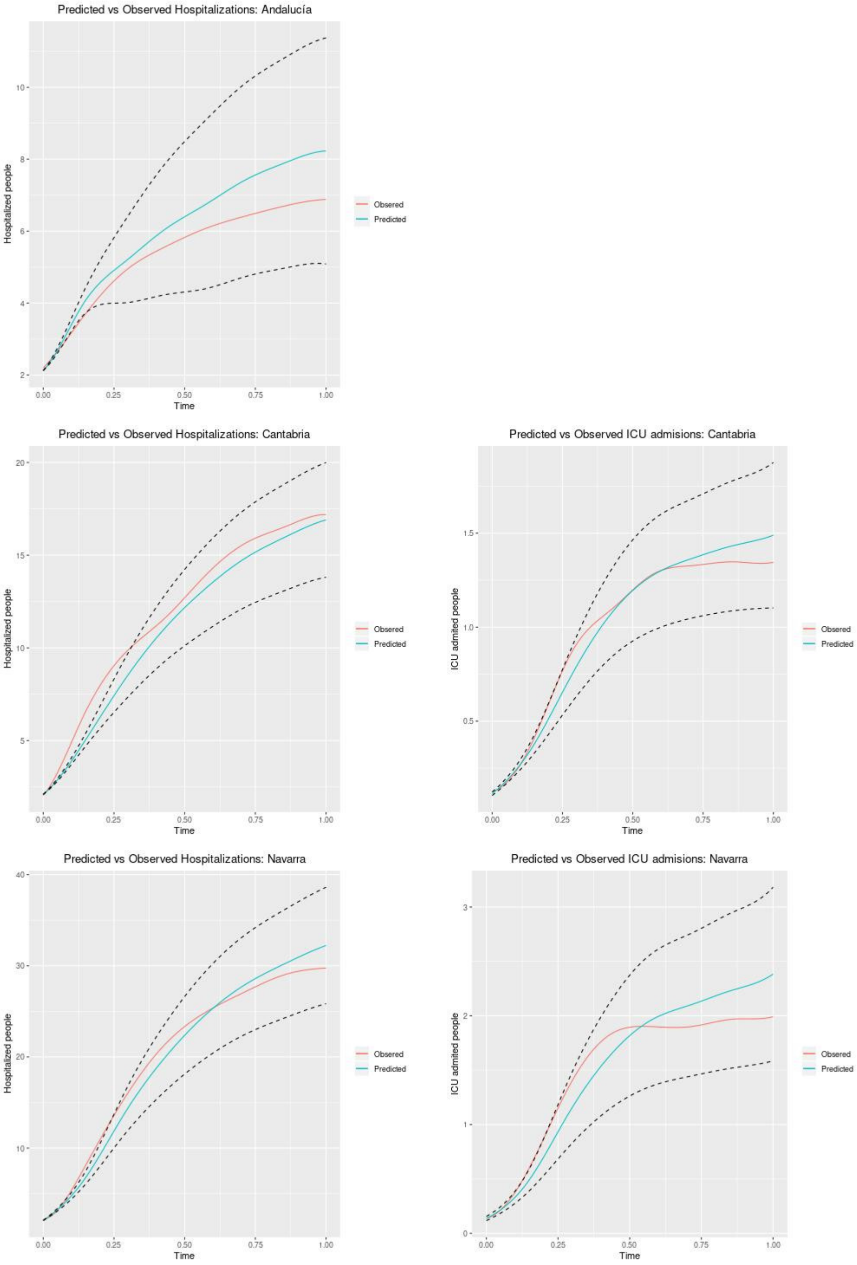


Figure 6: Observed and predicted curves (with pointwise confidence bands) of hospitalizations and ICU admissions in some of the training ACs.

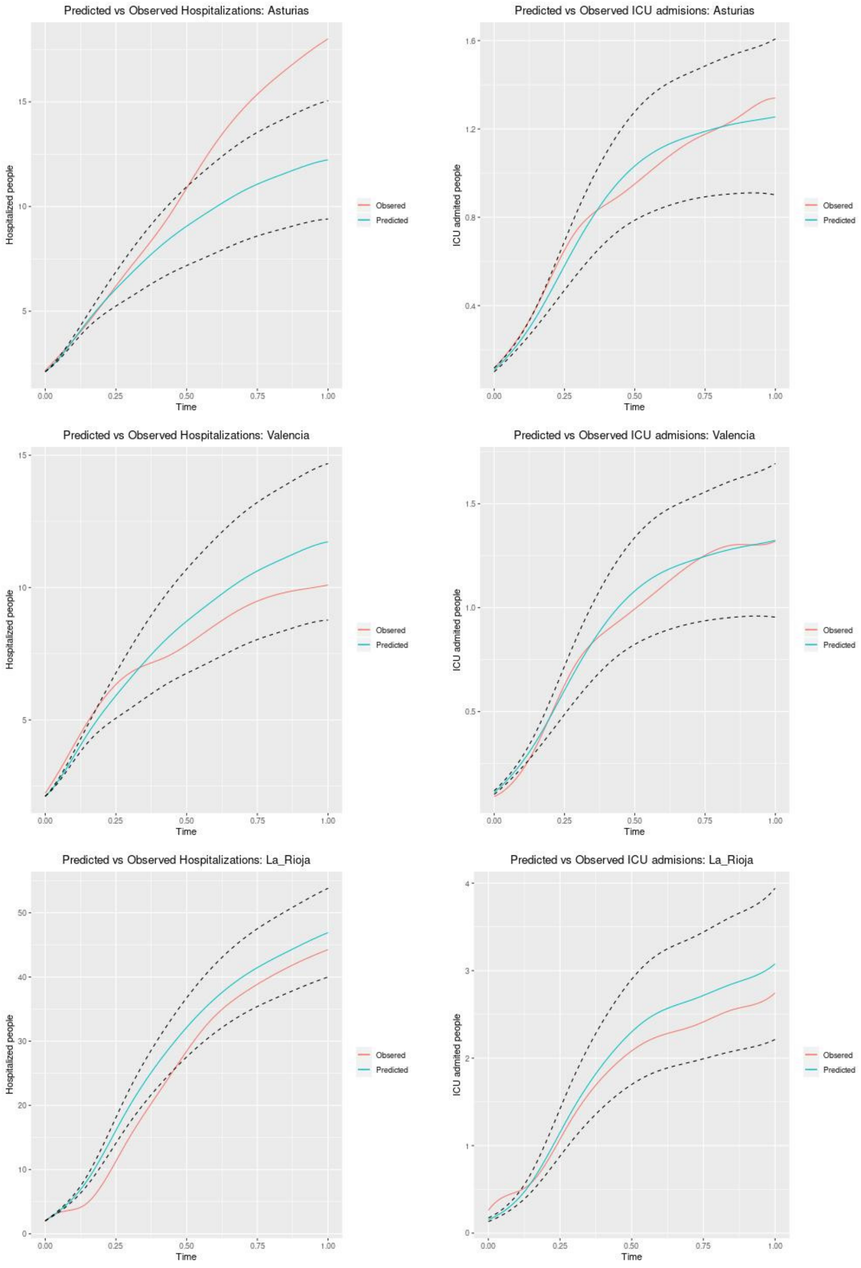


Figure 7: Observed and predicted curves (with pointwise confidence bands) of hospitalizations and ICU admissions in some of the training ACs.

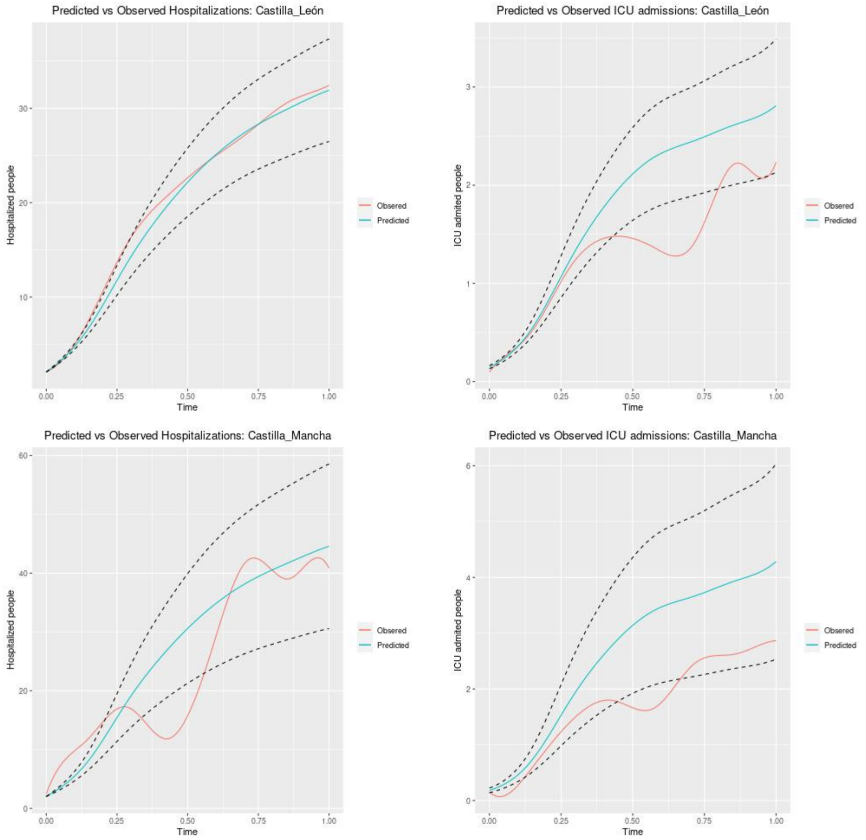


Figure 8: Observed and predicted curves (with pointwise confidence bands) of hospitalizations and ICU admissions in Castilla La Mancha and Castilla León.

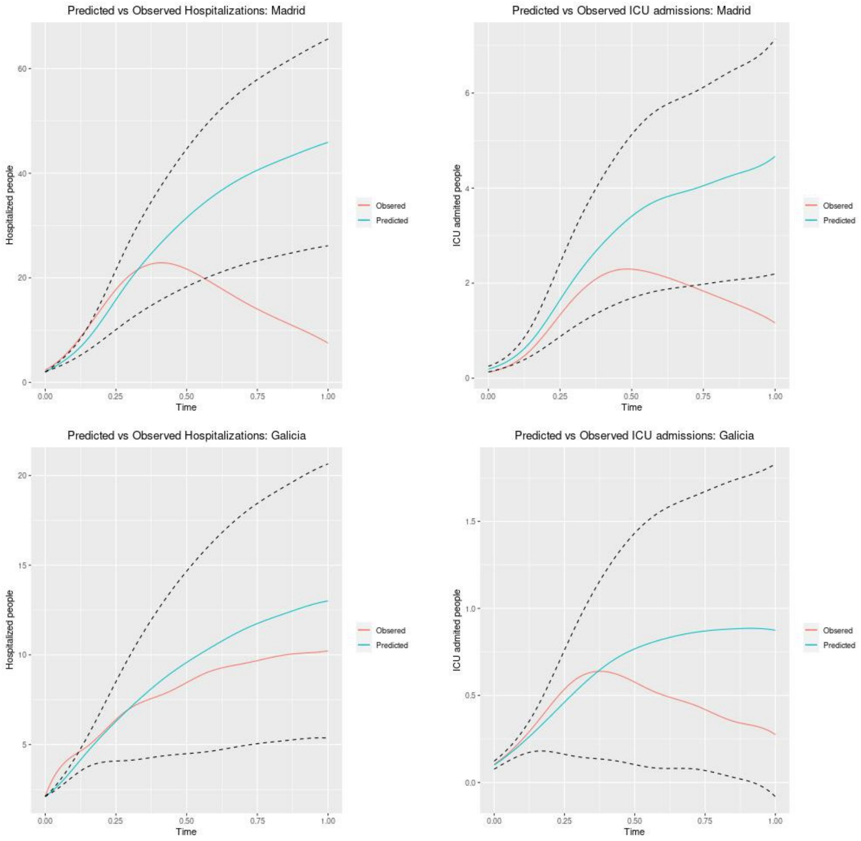


Figure 9: Observed and predicted curves (with pointwise confidence bands) of hospitalizations and ICU admissions in Madrid and Galicia.

that corrects the inaccurate reported data. These pointwise predictions next to their anomalous values for the first and last days of the first wave of COVID-19 in Castilla La Mancha, Castilla León, Madrid and Galicia can be seen in Table 2. The obtained predictions can be considered as an imputation of the real behaviour of these curves in the observation period if the mode of data communication would not have changed. Thus, in Castilla La Mancha on April 27, 8464 people were reported as hospitalized and the imputation provides by the model is 9062 cases; in Castilla León 7777 versus 7658; in Galicia 2758 versus 3511; and in Madrid 5039 versus 30587. For ICU admissions the differences between the registered and imputed cases are again evident. It can be seen that in Castilla La Mancha on 27 April 584 people were registered as admitted in ICU and the model gives an estimation of 871; in Castilla León 522 versus 674; in Galicia 73 versus 236; and in Madrid 780 versus 3111.

4.2 Canonical Correlation Analysis

Once the missing data have been imputed and complete curves are available for the 17 ACs, the relationship between the variables related to the number of people admitted in hospitals (hospitalized and ICU people) and the ones affected by the disease (sick, deceased and recovered) can be studied. Canonical Correlation Analysis (CCA) on the two sets of first principal components associated with these functional variables, is applied to explore this relationship without necessarily distinguishing between independent and dependent variables. The analysis makes sense because the correlations between PCs in the two groups are very high what suggests that the variables are not linearly independent.

In agreement with the above, the first principal component of each functional variable is selected to carry out the analysis. Thus, the dataset consists of a sample of the seventeen Spanish ACs in an attempt to determine which factors influence in the hospital occupancy rate. The two groups of variables are, on the one hand, *Hospital occupancy rate* (HOR) formed by the first PC of hospitalized people and of ICU people ($\hat{\xi}_1^{y_1}, \hat{\xi}_1^{y_2}$), and on the other hand, *Illness response* (IR) comprised by the first PC of positive people, of deceased people and of recovered people ($\hat{\xi}_1^{x_1}, \hat{\xi}_1^{x_2}, \hat{\xi}_1^{x_3}$). The estimates of the squared canonical correlations between the two canonical variables for each pair appear in Table 3, next to the outcomes associated with the Barlett's test for testing the null hypothesis that the two canonical variate pairs are uncorrelated. As a result, it can be concluded that both canonical pairs are significantly correlated and dependent each on other (there is relationship between the two sets of variables).

Note that the squared canonical correlations represent, for each pair, the percentage of variance in one canonical variate explained by the variation in the other one, but say nothing about the extent to which the canonical vari-

	Hospitalizations							
	Castilla La Mancha		Castilla León		Galicia		Madrid	
Time	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred
1	635	413	476	495	557	570	1518	1351
2	838	565	629	630	906	676	2337	1779
3	1547	735	798	784	1043	809	2337	2247
4	1826	932	977	961	1147	961	3710	2772
5	2162	1164	1197	1163	1250	1120	3778	3371
6	2162	1436	1457	1394	1338	1276	5168	4059
7	2707	1758	1823	1656	1447	1424	6338	4853
8	2977	2124	2214	1948	1630	1564	7388	5768
9	3018	2520	2648	2259	1767	1698	8441	6794
···	···	···	···	···	···	···	···	···
···	8173	8200	7080	6970	2609	3171	8191	28159
···	8199	8317	7174	7064	2652	3222	7930	28482
···	8243	8430	7264	7155	2674	3270	7464	28802
···	8304	8542	7397	7246	2694	3316	7077	29120
···	8342	8654	7506	7336	2707	3362	6601	29434
···	8385	8763	7555	7424	2722	3407	6183	29740
···	8417	8868	7653	7508	2735	3449	5892	30037
···	8444	8969	7703	7586	2746	3484	5441	30320
···	8464	9062	7777	7658	2758	3511	5039	30587
	ICU admissions							
	Castilla La Mancha		Castilla León		Galicia		Madrid	
Time	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred
1	23	37	24	35	29	27	77	127
2	23	45	43	44	35	35	102	152
3	29	56	54	54	47	44	135	184
4	37	70	69	67	55	53	180	224
5	37	88	85	83	69	63	180	273
6	65	110	106	102	86	74	253	332
7	76	136	120	124	98	85	253	404
8	142	167	137	150	112	96	340	488
9	182	202	170	178	123	107	491	587
···	···	···	···	···	···	···	···	···
···	531	784	501	615	108	237	1111	2826
···	534	793	503	622	101	237	1076	2853
···	537	801	505	628	96	238	1024	2878
···	546	809	508	634	92	239	981	2903
···	553	817	510	639	87	239	949	2930
···	559	826	511	645	90	239	892	2962
···	565	838	515	653	85	239	873	3001
···	579	852	518	662	83	238	821	3050
···	584	871	522	674	73	236	780	3111

Table 2: Pointwise imputation of hospitalizations and ICU admissions for the first and last days of the first COVID-19 wave in Castilla La Mancha, Castilla León, Madrid and Galicia.

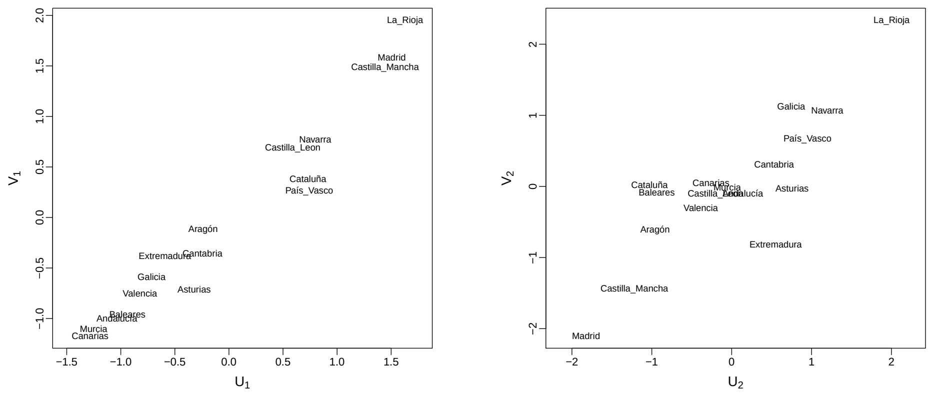


Figure 10: Scatterplot for the first (left) and second (right) canonical variate pairs.

ates themselves account for variation in the original variables. Then, around 95.4% of the variation in the first canonical variate for HOR (U_1) is described by the variation in the first canonical variate for IR (V_1), and almost 71% of the variation in U_2 is explained by V_2. This fact suggests that both canonical correlations are important. Figure 10 displays how the values of the canonical variates are spread over the plane. The linear relation in each pair is clearly visible in these scatterplots. Likewise, it is possible to draw conclusions about which ACs behave similarly during the first wave. The results are in concordance with multiple studies about the COVID-19 pandemic in Spain (see for example [1]).

Canonical Corr.	Squared Canonical Corr.	Stat	df	p-value
0.9765693	0.9536876	55.82721	6	<0.001
0.8398669	0.7053764	15.88673	2	<0.001

Table 3: Estimates of the canonical correlations next to χ^2 values associated with Bartlett's omnibus statistic, degrees of freedom and p-values for each canonical variate pair.

Additionally, the estimated canonical coefficients (loadings) for the HOR and RI variables are in Table 4 and Table 5, respectively. The magnitudes of these coefficients give the contributions of the individual variables to the corresponding canonical variable. Hence, the canonical variables are determined as follows:

$$
\begin{aligned}
U_1 &= -0.1767528 \times \xi_1^{y_2} + 0.1214623 \times \xi_1^{y_1} \\
U_2 &= -3.2637387 \times \xi_1^{y_2} + 0.2556912 \times \xi_1^{y_1} \\
V_1 &= 0.0336045 \times \xi_1^{x_1} + 0.1877252 \times \xi_1^{x_2} - 0.0044276 \times \xi_1^{x_3} \\
V_2 &= 0.1094736 \times \xi_1^{x_1} - 1.0135127 \times \xi_1^{x_2} + 0.0047968 \times \xi_1^{x_3}
\end{aligned}
$$

	U_1	U_2
ICU	-0.1767528	-3.2637387
Hospitalized	0.1214623	0.2556912

Table 4: Canonical coefficients for HOR variables.

	V_1	V_2
Cases	0.0336045	0.1094736
Deceased	0.1877252	-1.0135127
Recovered	-0.0044276	0.0047968

Table 5: Canonical coefficients for IR variables.

Once the raw canonical coefficients have been estimated, the following step is to interpret each canonical component. For that purpose, the squared correlations between the variables in each group and the canonical components are computed in Table 6 and 7 for HOR and IR groups, respectively. These parameters indicate the fraction of HOR and IR variance associated with each of their components separately. Let us observe that for the second canonical variables (U_2, V_2) none of the correlations are large so that this pair provides very little information about the variables. Regarding the first canonical variate pair (U_1, V_1), all the correlations with the variables are uniformly high. This means that U_1 and V_1 are an overall measure of HOR and IR variables, respectively, with U_1 being highly correlated with hospitalizations and V_1 more correlated with positive cases and deceased people.

These outcomes expose that the level of saturation in the hospitals are determined especially by the number of hospitalized people, meanwhile response to pandemic is governed by the number of positive cases and deaths. Despite the fact that the number of people in UCI and the number of recovered people play also an important role over the canonical variates, their contribution is smaller.

	U_1	U_2	V_1	V_2
ICU	0.8158880	0.184111953	0.7781023	0.129868216
Hospitalized	0.9970756	0.002924353	0.9508986	0.002062769

Table 6: Squared correlations between the HOR variables and the canonical variables.

	V_1	V_2	U_1	U_2
Cases	0.9660682	0.03366429	0.9213272	0.02374600
Deceased	0.9269440	0.07237154	0.8840149	0.05104917
Recovered	0.7485881	0.04012734	0.7139192	0.02830487

Table 7: Squared correlations between the IR variables and the canonical variables.

	U_1	U_2	V_1	V_2
HOR	-	-	0.86450046	0.06596549
RI	0.83975376	0.03436668	-	-

Table 8: Total fraction of HOR (IR) variance accounted by IR (HOR) variables, through each canonical variate in first row (second row).

Finally, a canonical redundancy analysis is performed in order to study the percentage of variance of one group of variables that is accounted by the other (in the usual least squares sense). The results of this analysis can be seen in Table 8 and the correlations between each set of variables and the opposite group of canonical variates in Tables 6 and 7. Table 8 shows that both components of the first canonical pair are a good overall predictor of the opposite set of variables, since the explained proportions of variance for HOR and IR are 0.864 and 0.839, respectively. Nevertheless, despite the correlation for the second pair was significant, these variables does not account for a great amount of variability. This statement is corroborated by the squared correlations displayed in Table 6 and 7. These measures indicate that the first canonical variate of IR group has an outstanding predictive power for the number of hospitalized (95.09%) and a considerable influence for the number of people in ICU (77.81%) as well. Similar interpretations are reached for the first canonical variable of HOR, which is a superb predictor of the number of cases and deaths (92.13% and 88.40%, respectively), and to lesser extent, of the number of recuperated

(71.39%). The second canonical variables add virtually nothing given that the fraction of variance in each variable set attributable to the other group through the respective canonical variates barely overcome the 10% of the total variability.

5 Conclusions

The current economic and sanitary crisis provoked by the virus SARS-CoV-2 is concentrating all of the planet's attention since the World Health Organization declared the worldwide emergency state in the middle of March 2020. In order to control the propagation of the virus, the scientific community is immersed in the development of statistical models that enable the governments to control the behaviour of the pandemic and to mitigate the devastating effects of COVID-19 illness. Thus, it is essential to build powerful models to be able to guarantee accurate predictions. Taking into account the nature of the variables of interest (for instance, number of positive cases, deceases, recovered, hospitalised and people in intensive care units), a wide variety of models have been tackled by considering Functional Data Analysis methodologies. Nevertheless, good performance of these models depends on the quality of the data, which is not always as good as one might expect, especially in periods of pandemic where the data are usually incomplete. On this matter, an extension of function-on-function linear regression is proposed for the imputation of missing values in the response, where the functional coefficients are estimated by means of principal components regression. The motivation for this work is to forecast the curves of hospitalized and intensive care people (functional responses) from the curves of positive cases, deaths and recoveries (functional predictors) for several Spanish Autonomous Communities that changed the way of recording data related to hospital occupancy rate. The imputation of these curves is made once the linear model is estimated with a training sample composed by the remainder of communities that did not modify their way of data registering. The performance of the model is outstanding for the training sample, since the observed and predicted curves are very similar for both functional responses. Regarding the prediction sample, the obtained forecasts can be considered as an imputation of what should have been the real behaviour of these curves in the observation period if the mode of data communication would not have changed. It can be observed that the model captures the trend of the curves up to the change. Additionally, once the missing data were imputed, a canonical correlation analysis was carried out in order to study the possible relationship between the two groups of variables: hospital occupancy rate (number of hospitalized people and ICU admissions) and illness response (number of positive cases, deaths and recovered people). The first principal component score of each variable was selected to make the canonical analysis, since only the first principal component explains almost all the variability of the five functional variables. After an exhaustive analysis, both sets of variables have shown to be highly correlated with each other and moreover, each of the first canonical variables is a good overall predictor of the opposite group of variables. At this point, the variables with more predictive power are the number of hospitalizations, positives and deceases. To sum up, the present document introduces a new mechanism for the imputation of missing at random functional response curves and shows the relationship among interesting functional variables associated with the COVID-19 pandemic.

References

[1] C. Acal, A. M. Aguilera, and M. Escabias. New modeling approaches based on varimax rotation of functional principal components. *Mathematics*, 8(11):2085, 2020.

[2] P. Agarwal and K. Jhajharia. Data analysis and modeling of covid-19. *Journal of Statistics and Management Systems*, 24(1):1–16, 2021.

[3] A. M. Aguilera, C. Acal, M. C. Aguilera-Morillo, F. Jiménez-Molinos, and J. B. Roldán. Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Mathematics and Computers in Simulation*, 186:41–51, 2021.

[4] A. M. Aguilera, M. C. Aguilera-Morillo, and C. Preda. Penalized versions of functional pls regression. *Chemometrics and Intelligent Laboratory Systems*, 154:80–92, 2016.

[5] A. M. Aguilera, M. Escabias, F. A. Ocaña, and M. J. Valderrama. Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress. *Methodology and Computing in Applied Probability*, 17(4):1015–1028, 2015.

[6] A. M. Aguilera, M. Escabias, C. Preda, and G. Saporta. Using basis expansion for estimating functional PLS regression. Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104(2):289–305, 2010.

[7] A. M. Aguilera, F. A. Ocaña, and M. J. Valderrama. An approximated principal component prediction model for continuous-time stochastic processes. *Applied Stochastic Models and Data Analysis*, 13(2):61–72, 1997.

[8] A. M. Aguilera, F. A. Ocaña, and M. J. Valderrama. Forecasting with unequally spaced data by a functional principal component approach. *Test*, 8(1):233–254, 1999.

[9] M. C. Aguilera-Morillo, A. M. Aguilera, M. Escabias, and M. J. Valderrama. Penalized spline approaches for functional logit regression. *TEST*, 22(2):251–277, 2013.

[10] A. Berihuete, M. Sanchez-Sanchez, and A. Suarez-Llorens. A bayesian model of covid-19 cases based on the gompertz curve. *Mathematics*, 9(3):228, 2021.

[11] A. Briz-Redon. The impact of modelling choices on modelling outcomes: a spatio-temporal study of the association between covid-19 spread and environmental conditions in catalonia (spain). *Stochastic Environmental Research and Risk Assessment*, 2021.

[12] C. Carroll, S. Bhattacharjee, Y. Chen, P. Dubey, J. Fan, A. Gajardo, X. Zhou, H. G. Müller, and J. L. Wang. Time dynamics of covid-19. *Scientific reports*, 10:21040, 2020.

[13] J. M. Chiou, H. G. Müller, and J. L. Wang. Functional response models. *Statistica Sinica*, 14(3):659–677, 2004.

[14] A. Ciarleglio, E. Petkova, and O. Harel. Multiple imputation in functional regression with applications to eeg data in a depression study, 2020.

[15] C. Crambes and Y. Henchiri. Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, 201:103–119, 2019.

[16] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–156, 1982.

[17] A. Delaigle and P. Hall. Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, 40(1):322–352, 2012.

[18] J. C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, 15:3–101, 1974.

[19] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[20] M. Escabias, A. M. Aguilera, and M. J. Valderrama. Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4):365–384, 2004.

[21] M. Escabias, A. M. Aguilera, and M. J. Valderrama. Functional PLS logit regression model. *Computational Statistics and Data Analysis*, 51(10):4891–4902, 2007.

[22] M. Escabias, A.M. Aguilera, and M.C Aguilera-Morillo. Functional pca and base-line logit models. *Journal of Classification*, 31.

[23] M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Functional principal component regression and functional partial least squares regression: An overview and a comparative study. *International Statistical Review*, 85:61–83, 2017.

[24] M. Febrero-Bande, P. Galeano, and W. González-Manteiga. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. *Computational Statistics and Data Analysis*, 131:91–103, 2019.

[25] F. Ferraty, M. Sued, and P. Vieu. Mean estimation with data missing at random for functional covariables. *Statistics*, 47(4):688–706, 2013.

[26] F. Ferraty and P. Vieu. *Nonparametric functional data analysis. Theory and practice*. Springer-Verlag, 2006.

[27] J. W. Graham. *Missing data: Analysis and design.* Springer Science & Business Media, 2012.

[28] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society B*, 68(1):109–126, 2006.

[29] Y. He, R. Yucel, and T. E. Raghunathan. A functional multiple imputation approach to incomplete longitudinal data. *Statistics in medicine*, 30(10):1137–1156, 2011.

[30] L. Horvath and P. Kokoszka. *Inference for functional data with applications.* Springer-Verlag, 2012.

[31] I.R. Lima, G. Cao, and N. Billor. Robust simultaneous inference for the mean function of functional data. *TEST*, 28:785–803, 2019.

[32] N. Ling, L. Liang, and P. Vieu. Nonparametric regression estimation for functional stationary ergodic data with missing at random. *Journal of Statistical Planning and Inference*, 162:75–87, 2015.

[33] N. Ling, Y. Liu, and P. Vieu. Conditional mode estimation for functional stationary ergodic data with responses missing at random. *Statistics*, 50:991–1013, 2016.

[34] R. J. Little and D. B Rubin. *Statistical analysis with missing data (Third Edition).* John Wiley & Sons, 2019.

[35] M. Maleki, M. R. Mahmoudi, M. H. Heydari, and K. H. Pho. Modeling and forecasting the spread and death rate of coronavirus (covid-19) in the world using time series models. *Chaos, Solitons & Fractals*, 140:110151, 2020.

[36] M.M. Mansour, M.A. Farsi, S.M. Mohamed, and M.A. Elrazik. Modeling the covid-19 pandemic dynamics in egypt and saudi arabia. *Mathematics*, 9.

[37] J. C. Mora, S. Pérez, and A. Dvorzhak. Application of a semi-empirical dynamic model to forecast the propagation of the covid-19 epidemics in spain. *Forecasting*, 2(4):452–469, 2020.

[38] H. G. Müller and U. Stadtmüller. Generalized functional linear models. *Annals of Statistics*, 33(2):774–805, 2005.

[39] Y. Nie, L. Wang, B. Liu, and J. Cao. Supervised functional principal component analysis. *Statistics and Computing*, 28:713–723, 2018.

[40] F. A. Ocaña, A. M. Aguilera, and M. Escabias. Computational considerations in functional principal component analysis. *Computational Statistics*, 22(3):449–465, 2007.

[41] F. A. Ocaña, A. M. Aguilera, and M. J. Valderrama. Functional Principal Components Analysis by Choice of Norm. *Journal of Multivariate Analysis*, 71(2):262–276, 1999.

[42] D. Pak, K. Langohr, J. Ning, J. Cortés-Martínez, G. Gómez-Melis, and Y. Shen. Modeling the coronavirus disease 2019 incubation period: Impact on quarantine policy. *Mathematics*, 8.

[43] C. Preda and G. Saporta. PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48(1):149–158, 2005.

[44] H. Qi, S. Xiao, R. Shi, M. P. Ward, Y. Chen, W. Tu, Q. Su, W. Wang, X. Wang, and Z. Zhang. Covid-19 transmission in mainland china is associated with temperature and humidity: A time-series analysis. *Science of The Total Environment*, 728:138778, 2020.

[45] X. Qi and R. Luo. Function-on-function regression with thousands of predictive curves. *Journal of Multivariate Analysis*, 163:51–66, 2018.

[46] J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB.* Springer-Verlag, 2009.

[47] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: Methods and case studies*. Springer-Verlag, 2002.

[48] J. O. Ramsay and B. W. Silverman. *Functional data analysis (Second Edition)*. Springer-Verlag, 2005.

[49] A. R. Rao and M. Reimherr. Modern multiple imputation with functional data. *Stat*, 10(1):e331, 2021.

[50] J. E. Ruiz-Castro, C. Acal, A. M. Aguilera, M. C. Aguilera-Morillo, and J. B. Roldán. Linear-phase-type probability modelling of functional pca with applications to resistive memories. *Mathematics and Computers in Simulation*, 186:71–79, 2021.

[51] C. Tang, T. Wang, and P. Zhang. Functional data analysis: An application to covid-19 data in the united states, 2020.

[52] A. Tobias. Evaluation of the lockdowns for the sars-cov-2 epidemic in italy and spain after one month follow up. *Science of The Total Environment*, 725:138539, 2020.

[53] A. Torres-Signes, M. P. Frías, and M. D. Ruiz-Medina. Covid-19 mortality analysis from soft-data multivariate curve regression and machine learning, 2021.

[54] M.J. Valderrama, F.A. Ocaña, A.M. Aguilera, and F.M. Ocaña-Peinado. Forecasting pollen concentration by a two-step functional model. *Biometrics*, 66.

[55] M. Zanin and D. Papo. Assessing functional propagation patterns in covid-19. *Chaos, Solitons & Fractals*, 138:109993, 2020.

[56] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun. Deep learning methods for forecasting covid-19 time-series data: A comparative study. *Chaos, Solitons & Fractals*, 140:110121, 2020.