Trabajo Fin de Máster

Modelos de Regresión Funcional Aplicados a Datos del Covid-19

Luis Fidel Herrera de los Santos

Máster en Técnicas Estadísticas

Curso 2020-2021

Luis Fidel Herrera de los Santos

Propuesta de Trabajo Fin de Máster

Título en galego: Modelos de Regresión Funcional Aplicados a Datos Covid-19
Título en español: Modelos de Regresión Funcional Aplicados a Datos del Covid-19
English title: Functional Regression Models Applied to Covid-19 Data
Modalidad: Modalidad A
Autor/a: Luis Fidel Herrera de los Santos, Universidade da Coruña
Director/a: Javier Roca Padiñas, Universidade de Vigo; Manuel Oviedo de la Fuente, Universidade da Coruña
Tutor/a: Tutor/a 1, Empresa 1; Tutor/a 2, Empresa 2
Breve resumen del trabajo: EL objetivo de este trabajo es evaluar y comparar los diversos modelos de regresión funcional e incorporar una versión iterativa que contemple estructuras de dependencias temporal de los errores. Para los fines, se propone un algoritmo de simulación con varios escenarios y un caso de aplicación a datos del Covid-19.
Recomendaciones:
Otras observaciones:

Luis Fidel Herrera de los Santos

Don/doña Javier Roca Padiñas, Profesor Asociado de la Universidade de Vigo, don/doña Manuel Oviedo de la Fuente, Profesor suplente de doctorado de la Universidade da Coruña, informan que el Trabajo Fin de Máster titulado

Modelos de Regresión Funcional Aplicados a Datos del Covid-19

fue realizado bajo su dirección por don/doña Luis Fidel Herrera de los Santos para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En [Coruña], a xx de [Septiembre] de 2021.

El/la director/a:

Don/doña Javier Roca Padiñas

El/la director/a:

Don/doña Manuel Oviedo de la Fuente

El/la tutor/a:

Don/doña Tutor/a 1

El/la tutor/a:

Don/doña Tutor/a 2

El/la autor/a:

Don/doña Luis Fidel Herrera de los Santos

Luis Fidel Herrera de los Santos

Agradecimientos

Agradecer a Dios su voluntad, a mi madre por su apoyo incondicional, a la MESCyT y al MINERD por la oportunidad y apoyo constante, a mis colegas y amigos por la motivación y a mis compañeros de maestría por compartir conmigo alegrías y penas.

A las personas que me recibieron sin conocerme y compartí el mismo techo, por decir que si a mi necesidad de alojamiento y estar pendiente de que siempre estuviera a gusto. ¡Gracias por todo el calor humano!

A todos los profesores por la oportunidad de transmitir sus conocimientos con un alto nivel y humildad.

Finalmente, a mis directores de TFM Javier por ser consciente y aterrizado, por darme las bases de las estadística computacional y de manera especial a Manuel por ser paciente, su dedicación, preocupación y enseñanzas durante todo este tiempo. ¡De verdad que muchísimas gracias!

Luis Fidel Herrera de los Santos

Índice general

Resumen xvii

1. Introducción 1

2. Análisis exploratorio de datos funcionales 3
2.1. Conceptos de datos funcionales . . . 3
2.1.1. Estadísticos descriptivos . . . 5
2.2. Espacio funcional . . . 5
2.3. Representación en base . . . 8
2.4. Suavizado . . . 13
2.5. Uso de las derivadas . . . 16

3. Modelos de regresión funcional (FRM) 19
3.1. Regresión lineal funcional (FLR) . . . 20
3.2. Regresión No paramétrica funcional (FRNP) . . . 20
3.3. Regresión aditiva funcional (FAR) . . . 21
3.4. Regresión aditiva generalizada funcional (FGAM) . . . 21
3.5. Modelos de regresión funcional con errores dependientes . . . 22

4. Simulación 31

5. Aplicación de FRM para predecir la incidencia acumulada de contagio del Covid-19 41
5.1. Metodología . . . 42
5.1.1. Definición de las variables . . . 42
5.2. Explotación del conjunto de datos . . . 43
5.2.1. Derivadas de las curvas de incidencia . . . 48
5.3. Predicciones de incidencia acumulada de contagio del Covid-19. . . . 48
5.3.1. Error cuadrático medio de las predicciones. . . . 49

6. Conclusión 61
6.1. Investigaciones futuras . . . 62

Apéndices 65

A. Resultados de la simulación 65
A.1. Tablas . . . 65
A.1.1. Número de bases estimado . . . 65
A.1.2. MSE del parámetro β . . . 66
A.1.3. MSE del parámetro ϕ . . . 66
A.1.4. MSPE de los modelos . . . 67
A.2. Resumen de los modelos . . . 70

A.2.1. Modelos con ε generado por proceso ARMA(1,1) y $\delta = 1$ 70
A.2.2. Modelos con ε generado por proceso ARMA(1,1) y $\delta = 0{,}70$ 76
A.2.3. Modelos con ε generado por proceso AR(1) y $\delta = 1$ 82
A.3. Gráficos 89
A.3.1. Comparativo de los β estimados 89
A.3.2. Ajuste de modelo 91
A.3.3. Predicciones 93

B. Resultados de la aplicación **95**
B.1. Resultados de los modelos FRM para estimar la incidencia acumulada de contagio del Covid-19. 95
B.1.1. Salida de modelos para la comunidad de Cataluña. 95
B.1.2. Salida de modelos para la comunidad de Galicia. 100
B.2. Tablas de error a $H = 1, ..., 7$. 110
B.3. Estimación de la estructuras de dependencia de los errores 111

Bibliografía **113**

Índice de figuras

2.1. Temperaturas medias (en grados Celsius) diarias de 73 estaciones meteorológicas españolas. Período 1980-2009. *"data aemet".* 4
2.2. Media de temperaturas medias(en grados Celsius) diarias de 73 estaciones meteorológicas españolas. Período 1980-2009. *data "aemet".* 7
2.3. Temperaturas (en grados Celsius) diarias de una estación meteorológica española y su representación en base de B-spline ($K_x = 11$). Período 1980-2009. *data "aemet".* 9
2.4. Promedio de temperaturas (en grados Celsius) diarias de una estación meteorológica española y su representación en base de componentes principales ($K_x = 3$). Período 1980-2009. *data "aemet".* 11
2.5. Comportamientos de las componentes principales ($K_x = 3$) para representar la *data "aemet".* 12
2.6. Comparativo de la representación en base y el suavizado GCV para las temperaturas medias diarias (en grados Celsius) de una estación meteorológica española. Período 1980-2009. *data "aemet".* 15
2.7. En la figura (1): las temperaturas medias (en grados Celsius) diarias de las estaciones meteorológica española. En la figura (2): el suavizado GCV con base de B-spline para las temperaturas medias diarias (en grados Celsius) de las estaciones meteorológica española. Período 1980-2009. *data "aemet".* 15

4.1. Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. ARMA(1,1). $\delta = 1$. . . 34
4.2. Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. ARMA(1,1). $\delta = 0{,}7$. . 34
4.3. Predicciones de los modelos a horizonte $h = 10$. ARMA(1,1). $\delta = 1$ 36
4.4. Predicciones de los modelos a horizonte $h = 10$. ARMA(1,1). $\delta = 0{,}7$. 37

5.1. Curvas epidémicas de las CCAA de España. Período 01/01/2020 a 10/06/2021. Construido en base a los datos del número de casos de contagio Covid-19 de la plataforma del CNE. 43
5.2. Comparativo de Curvas epidémicas de las CCAA de Cataluña, Galicia y Madrid. Período 01/01/2020 a 10/06/2021. Construido en base a los datos del número de casos de contagio Covid-19 de la plataforma del CNE. 44
5.3. Porcentaje de casos confirmados en las CCAA de España. Período 01/01/2020 a 10/06/2021. 45
5.4. Acumulado de casos confirmados en las CCAA de España. Período 01/01/2020 a 10/06/2021. 45
5.5. Acumulado de casos confirmados en las CCAA de Cataluña, Galicia y Madrid. Período 01/01/2020 a 10/06/2021. 46
5.6. Incidencia acumulada de casos confirmados a 15 días en las CCAA de España por cada $100,000$ habitantes. Período 19/02/2020 a 03/06/2021. 46
5.7. Incidencia acumulada a 15 días de casos confirmados en Cataluña, Galicia y Madrid por cada $100,000$ habitantes. 19 de febrero 2020 a 18 de mayo de 2021. 47
5.8. Curvas de incidencia acumulada y su derivada a 15 días de casos confirmados en Cataluña. Salto entre las cuervas de $K = 1$. 48

5.9. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo LM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 51
5.10. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo SAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 52
5.11. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo KAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 52
5.12. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo GLS-AR(1) y GLS-ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. . . 53
5.13. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo iGLS-AR(1) y iGLS-AR(P) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 53
5.14. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo LM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 54
5.15. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo SAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 55
5.16. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo KAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 55
5.17. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo GLS-AR(1) y GLS-ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 56
5.18. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo iGLS-AR(1) y iGLS-AR(P) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 56
5.19. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo LM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 57
5.20. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo SAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 58
5.21. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo KAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 58
5.22. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo GLS-AR(1) y GLS-ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 59

5.23. Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo iGLS-AR(1) y iGLS-AR(P) con las observaciones en la comunidad de Madrid por cada 100,000 habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021. 59

A.1. Resultados resumen del modelo LM representado por PC. ARMA(1,1). $\delta = 1$ 71
A.2. Resultados resumen del modelo LM representado por BSP.ARMA(1,1). $\delta = 1$ 72
A.3. Comportamiento de las PC del modelo GSAM. ARMA(1,1). $\delta = 1$ 74
A.4. Comportamiento de las BSP del modelo GSAM. ARMA(1,1). $\delta = 1$ 75
A.5. Resultados resumen del modelo GKAM. ARMA(1,1). $\delta = 1$. 76
A.6. Resultados resumen del modelo LM representado por PC. ARMA(1,1). $\delta = 0{,}7$. 77
A.7. Resultados resumen del modelo LM representado por BSP. ARMA(1,1). $\delta = 0{,}7$ 78
A.8. Comportamiento de las PC del modelo GSAM. ARMA(1,1). $\delta = 0{,}7$ 80
A.9. Comportamiento de las BSP del modelo GSAM. ARMA(1,1). $\delta = 0{,}7$ 81
A.10. Resultados resumen del modelo GKAM. ARMA(1,1). $\delta = 0{,}7$ 82
A.11. Resultados resumen del modelo LM representado por PC. AR(1). $\delta = 1$. 83
A.12. Resultados resumen del modelo LM representado por BSP. AR(1). $\delta = 1$. 84
A.13. Comportamiento de las PC del modelo GSAM. AR(1). $\delta = 1$. 86
A.14. Comportamiento de las BSP del modelo GSAM. AR(1). $\delta = 1$. 87
A.15. Resultados resumen del modelo GKAM. AR(1). $\delta = 1$. 88
A.16. Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. AR(1). $\delta = 1$. 89
A.17. Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. AR(1). $\delta = 0{,}15$. . . . 90
A.18. Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. ARMA(1,1). $\delta = 1$. 91
A.19. Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. ARMA(1,1). $\delta = 0{,}7$. 91
A.20. Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. AR(1). $\delta = 1$. 92
A.21. Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. AR(1). $\delta = 0{,}15$. 92
A.22. Predicciones de los modelos a horizonte $h = 10$. AR(1). $\delta = 1$. 93
A.23. Predicciones de los modelos a horizonte $h = 10$. AR(1). $\delta = 0{,}15$. 93

B.1. Resultados gráficos del modelo FLM para la comunidad de Cataluña. Función *fregre.lm*. 96
B.2. Resultados gráficos del modelo SAM para la comunidad de Cataluña. Función *fregre.gsam*. 97
B.3. Resultados gráficos del modelo GLS-AR(1) para la comunidad de Cataluña. Función *fregre.gls*. 98
B.4. Resultados gráficos del modelo GLS-ARMA(1,1) para la comunidad de Cataluña. Función *fregre.gls*. 99
B.5. Resultados gráficos del modelo FLM para la comunidad de Galicia. Función *fregre.lm*. . 101
B.6. Resultados gráficos del modelo SAM para la comunidad de Galicia. Función *fregre.gsam*. 102
B.7. Resultados gráficos del modelo GLS-AR(1) para la comunidad de Galicia. Función *fregre.gls*. 103
B.8. Resultados gráficos del modelo GLS-ARMA(1,1) para la comunidad de Galicia. Función *fregre.gls*. 104
B.9. Resultados gráficos del modelo FLM para la comunidad de Madrid. Función *fregre.lm*. . 106
B.10. Resultados gráficos del modelo SAM para la comunidad de Madrid. Función *fregre.gsam*. 107
B.11. Resultados gráficos del modelo GLS-AR(1) para la comunidad de Madrid. Función *fregre.gls*. 108
B.12. Resultados gráficos del modelo GLS-ARMA(1,1) para la comunidad de Madrid. Función *fregre.gls*. 109

Índice de tablas

4.1. Promedio de Número de bases seleccionados por el criterio de GCCV en $n_B = 402$ para una proporción de linealidad es $\delta = 1$ y $n_B = 389$ cuando la proporción del linealidad es $\delta = 0{,}70$ del Modelo (a). ARMA(1,1). *($n_B < 500$ por generación de proceso no estacionario).* 33
4.2. Error cuadrático medio del parámetro β. Modelo (a), $n_B = 402$ para una proporción de linealidad es $\theta = 1$ y $n_B = 389$ cuando la proporción del linealidad es $\theta = 0{,}70$. ARMA(1,1). 35
4.3. Error cuadrático medio del parámetro ϕ y θ. Modelo (a), $n_B = 402$ réplicas para una proporción de linealidad es $\delta = 1$ y $n_B = 389$ cuando la proporción del linealidad es $\delta = 0{,}70$. ARMA(1,1). 35
4.4. Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 402$. ARMA(1,1). $\delta = 1$. 38
4.5. Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 402$. ARMA(1,1). $\delta = 0{,}70$. 39

5.1. Predicciones a horizonte $H = 1, ..., 7$ de las curvas de incidencia a 15 días de la comunidad de Cataluña con saltos $k = 1$. 47
5.2. Error cuadrático medio de las predicciones de incidencia acumulada de contagio de la enfermedad Covid-19 a horizonte $H = 1, ..,7$ para las comunidades de Cataluña. Período 04/06/2020 al 10/06/2020. 49
5.3. Error cuadrático medio de las predicciones de incidencia acumulada de contagio de la enfermedad Covid-19 a horizonte $H = 1, ..,7$ para las comunidades de Galicia. Período 04/06/2020 al 10/06/2020. 50
5.4. Error cuadrático medio de las predicciones de incidencia acumulada de contagio de la enfermedad Covid-19 a horizonte $H = 1, ..,7$ para las comunidades de Madrid. Período 04/06/2020 al 10/06/2020. 50

A.1. Promedio de Número de bases seleccionados por el criterio de GCCV en $n_B = 500$ del Modelo (a). AR(1). 65
A.2. Promedio de Número de bases seleccionados por el criterio de GCCV en $n_B = 500$ del Modelo (a). AR(2). 65
A.3. Error cuadrático medio del parámetro β. Modelo (a), $n_B = 500$. AR(1) 66
A.4. Error cuadrático medio del parámetro β. Modelo (a), $n_B = 500$. AR(2). 66
A.5. Error cuadrático medio del parámetro ϕ. Modelo (a), $n_B = 500$. AR(1). 66
A.6. Error cuadrático medio del parámetro ϕ_1 y ϕ_2 . Modelo (a), $n_B = 500$. AR(2). 67
A.7. Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(1). $\delta = 1$. 67
A.8. Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(1). *delta* $= 0{,}15$. 68

A.9. Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(2). $\delta = 1$. 69
A.10. Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(2). $\delta = 0{,}15$. 70

B.1. Error predicciones de los modelos FRM para la comunidad de Cataluña. Período 04/06/2020 al 10/06/2020. 110
B.2. Error predicciones de los modelos FRM para la comunidad de Cataluña. Período 04/06/2020 al 10/06/2020. 110
B.3. Error predicciones de los modelos FRM para la comunidad de Madrid. Período 04/06/2020 al 10/06/2020. 111
B.4. Modelo autoregresivo de los errores AR(1) de los FRM de Cataluña. 111
B.5. Modelo autoregresivo de los errores AR(1) de los FRM de Galicia. 111
B.6. Modelo autoregresivo de los errores AR(1) de los FRM de Madrid. 112
B.7. Modelo autoregresivo de medias móviles de los errores ARMA(1,1) de los FRM de Cataluña. 112
B.8. Modelo autoregresivo de medias móviles de los errores ARMA(1,1) de los FRM de Galicia. 112
B.9. Modelo autoregresivo de medias móviles de los errores ARMA(1,1) de los FRM de Madrid. 112

Resumen

Resumen en español

EL objetivo de este trabajo es evaluar y comparar los diversos modelos de regresión funcional e incorporar una versión iterativa que contemple estructuras de dependencias temporal de los errores. Para los fines, se propone un algoritmo de simulación con varios escenarios y un caso de aplicación a datos del Covid-19.

English abstract

The objective of this work is to evaluate and compare the various functional regression models and incorporate an iterative version that includes temporal dependency structures of the errors. For the purposes, a simulation algorithm with several scenarios and a case of application to Covid-19 data is proposed.

Capítulo 1

Introducción

El análisis de datos funcionales (FDA) es uno de los campos innovadores que en los últimos ha despertado mayor interés a nivel teórico-práctico para los investigadores en áreas de la matemática aplicada , estadística y de otras áreas. Estos han colaborado para adaptar metodología matemática, estadística clásica y multivariante al razonamiento de datos funcionales, ya que los objetivos fundamentales de FDA son los mismos: identificar características *(comportamiento y patrones)*, explicar variabilidad, corroborar o descartar dependencia y hacer comparaciones entre dos o más grupos. El pensamiento de datos funcionales se extiende a una nueva dimensión de datos aleatorios *(curvas funcionales)* que evolucionan en un tiempo y/o espacio continuo observado/medido de manera discreta. Su utilidad ha tenido importantes aportes en ciencias como la econometría, la medicina, la epidemiología, ingeniería financiera y otras.

Aspectos teóricos profundos de FDA se pueden estudiar en *Ramsay y Silverman (2002,2005)* quienes expanden el concepto de FDA desde una perspectiva paramétrica. En este se puede encontrar métodos de suavizado, interpolación y de transformación para describir y explicar la variabilidad de un conjunto de datos representados en una base funcional. Parte esencial del texto son los capítulos de modelos de regresión funcional (FRM), los cuales, son una extensión de los métodos y técnicas de la estadística clásica, donde la variable funcional es una covariable que da respuesta a la variabilidad de un escalar. Con este tipo de casos se trabajará en la parte práctica de este documento, no obstante, los autores tratan otros modelos, como el de respuesta funcional. Otro punto de vista es el de *Ferraty y Vieu (2006)* con métodos no paramétricos para el análisis y modelado de datos funcionales en un espacio semimétrico. La aproximación de curvas mediante técnicas de ponderación local como la tipo Kernel, suavizado y otros, son de los principales métodos no paramétricos que se destacan en la exploración de datos funcionales. Además de eso, los autores proponen métodos no paramétricos de predicción, esenciales cuando las variables tiene una relación no estructural, lo que permite realizar modelados de datos mucho menos estrictos que en el contexto paramérico y con resultados asintóticos. Entre los documentos más recientes, correspondientes a los últimos 10 años, que se pueden estudiar está *Cuevas (2014)* quien precisa las ideas de los principales métodos del FDA de diversos autores que van desde el tratamiento, ajustes de parámetros centralizados hasta regresión y clasificación. Otros autores para fines de estudio se encuentran *Hsing y Eubank (2015)*, *KoKoszka y Reimherr (2017)*, etc.

De los modelos de regresión funcional (FRM) se encuentra una variedad de artículos con resultados interesantes de aplicación a datos simulados y reales. Está el caso de *Cardot y et al. (1998)* quienes describen como opera el modelos lineal funcional (FLM) y mediante simulación analizan la convergencia de los estimadores a partir de componentes principales (kn) y el aumento de n número de muestras. *McKeague y Qian (2011)*, siguen la misma linea de FLM aplicada a la medicina para evaluar políticas de tratamientos de enfermedades a partir de los genes. *Goia (2011)* utiliza los FLM en predicciones

de series de tiempo, como caso práctico hace pronósticos de la demanda de calor de un sistema de calefacción usando criterio de mínimos cuadrados penalizados (PLS) y base funcional de B-spline. En los modelos de regresión no paramétrico funcional (FRNP) se puede consultar a *Ferraty y et al. (2011)*, quienes analizan la existencia de un umbral en las variables para datos de estructuras complejas, su valor es obtenido usando criterio de validación cruzada (CVC). *Attouch y et al. (2011)* utilizan los mismos modelos FRNP robustos en datos espaciales. En los modelos de mínimos cuadrados generalizado funcional (GLS) se puede estudiar a *Kariya y et al. (2004)*; en los modelos de regresión aditiva funcional (FAR) se destaca Ferraty y et al. (2011). Mientras que *Febrero y et al. (2011)*, *Müller and Yao (2008)* y *Aneiros et al. (2011)* hacen aportes sustanciales a los modelos aditivos generalizados funcionales (FGAM).

En la práctica tratar datos funcionales requiere de operaciones estadísticas matemáticas manualmente complejas, en la mayoría de los casos. En vista de eso, se han desarrollado paquetes que aprovechan esas propiedades para aplicarlas eficientemente a un conjunto de datos funcionales. *Ramsay y et al. (2009)* realizan estos cálculos usando en computación el paquete fda del software R. Esta facilidad ha permitido hacer análisis en áreas de medicina, sismología, meteorología, fisiología y muchos otros campos. *Ferraty y Vieu (2006)* en su página web *(http://www.lsp.ups-tlse.fr/staph/npfda)* muestran métodos de análisis funcional no paramétrico con ejemplos de códigos empleados en ciencias aplicadas (medioambiental, quimiométrica, biométrica, medicina, econometría, etc).

Un paquete interesante de los últimos años es el *fda.usc* del software R publicado en un artículo de *Febrero y Oviedo (2012)* y de acceso en la web *(https://www.jstatsoft.org/article/view/v051i04)* con ejemplos prácticos de uso; detalles de las funciones y versión de actualización se pueden encontrar en la página *(https://www.rdocumentation.org/packages/fda.usc/versions/2.0.2)*. El paquete fda.usc propone nuevas funcionalidades de forma intuitiva para tratar datos funcionales, también integra las funciones principales del paquete fda, nlme y otros haciéndolo mucho más versátil al momento de tratar datos funcionales. Este paquete será útil para ejemplos prácticos de exploración y regresión funcional aplicados a un conjunto de datos simulados y reales a lo largo del desarrollo de este documento con el fin de hacer análisis exhaustivo y llegar a conclusiones.

En este proyecto se pretende hacer una revisión de los principales modelos de regresión funcional (FRM) con respuesta escalar propuestos por los diversos autores. El capítulo 2 introduce conceptos básicos de exploración de datos funcionales y varios ejemplos usando el paquete *fda.usc*. El capítulo 3 hace una revisión de los modelos de regresión funcional (FRM), con la finalidad de dar a conocer su estructura y principales características. El Capítulo 4 aplica dichos modelos a un conjunto de datos simulados bajo diversos escenarios, comparando los parámetros estimados y su capacidad predictiva. Finalmente, el Capítulo 5 aborda un caso real de los datos de contagio del Covid-19 en las Comunidades Autónomas (CCAA) de España y predice la incidencia acumulada de contagio bajo la premisa de dependencia temporal de los errores. Para este tipo de casos se propone aplicar modelos de mínimos cuadrados generalizados (GLS), su versión iterativa iGLS. Así como los modelos lineales funcionales (FLM), no paramétricos funcionales (FNP) y aditivos generalizados funcionales (FGAM) con su versión iterativa para obtener predicciones con resultados comparables. Luego de razonar las propiedades de estos modelos, elegir el más apropiado quedará sustentado con los resultados de la simulación y aplicación desarrolladas en los dos últimos capítulos.

Capítulo 2

Análisis exploratorio de datos funcionales

Para futuros análisis es necesario saber como definen los investigadores los datos funcionales, en la teoría matemática estos tienen dimensión infinita, es razonable decir que llevarlos a la práctica bajo estas condiciones sería inmanejable. La exploración de datos funcionales trabaja sobre muestras de dimensión finita debidamente representadas en un espacio funcional que aprecie las principales características y se evite en mayor medida la perdida de información.

En esta primera etapa de exploración se pondrá en contexto las notaciones que están relacionadas con cálculos de matemática avanzada y la estadística clásica y multivariante para tratar las curvas representadas en una base de funciones.

2.1. Conceptos de datos funcionales

El objetivo principal de esta sección es dar a conocer algunas definiciones de datos funcional y variable funcional según diversos autores y asociar esas definiciones a la estructura computacional que propone el paquete complementario de R fda.usc con el fin de introducir y/o construir los conjunto de datos funcionales *(datasets funtional)* y familiarizar a los lectores.

- Una variable funcional es una variable aleatoria donde las observaciones de una rejilla (t_{min}, t_{max}) están expresadas como $X(t_j)$ de forma continua. En *Ferraty-Vieu (2006)* define el concepto de variable funcional, dato funcional y conjunto de datos funcionales de la forma siguiente:

  - ***Definición 2.1.*** *Una variable aleatoria X es una variable funcional $(f.v)$ si toma valores en el espacio de dimensión infinita (o espacio funcional). Una observación x de X se llama dato funcional.*
  - ***Definición 2.2.*** *Un conjunto de datos funcionales $x_1, ..., x_n$ es la observación de n datos funcionales $X_1, ..., X_n$ idénticamente distribuidos como X.*

  Son curvas que suelen ser discretizadas, transformadas y suavizadas para adaptarlas al procesamiento de datos funcionales.

- *Ramsay y et al. (2005)* define el concepto de análisis de datos funcionales (FDA) como una perspectiva de pensamiento donde las funciones de datos observados en n pares (t_j, X_j) continuo

(tiempo, espacio, frecuencia, peso, etc) son vistas como entidades individuales, no como una secuencia de observaciones individuales. A modo de ejemplo, los autores presentan diversos casos prácticos de conjunto de datos con estructura funcional y su tratamiento aplicado. Tales son los casos de *"Datos sobre Crecimiento de Niñas"*, el cual corresponde a 10 niñas y su medida de altura en 31 edades. Estas curvas *son 10 observaciones funcionales de* $altura_i(t)$ donde la variación entre ellas fue tratada con métodos de derivadas. *"Datos de Fabricación en Estados Unidos"* en cada curva con un registro largo se observan replicaciones no necesariamente independientes, con tendencia y heterocedasticidad, producto de eso, fueron tratadas con transformación logarítmica obteniendo líneas con mejores propiedades que describen la variabilidad en el tiempo. *" Datos de entrada/salida de una refinería de petróleo"* estos datos funcionales se refieren a pares de entrada y salida. Finalmente, *"Datos sobre cómo caminan los niños"* estos datos funcionales tienen una naturaleza periódica cerradas que involucran un ejercicio para recopilar el ángulo formado de rodilla/cadera en un ciclo de marcha.

- Investigadores como *Aguilera (2009)* en la introducción de tesis llega a conclusiones importes sobre qué son datos funcionales. *Los datos en lugar de ser un conjunto de vectores, como en el análisis multivariante clásico, son un conjunto de curvas. En la mayoría de aplicaciones, las cuervas muestrales proceden de la observación de un proceso estocástico en tiempo continuo.*

- Según la teoría estadística, *Cuevas (2014)* clasifica de forma simple FDA como *"los problemas estadísticos en los que los datos disponibles consisten en una muestra de n funciones* $x_1 = x_1(t), ..., x_n = x_n(t)$ *definido en un intervalo compacto de la linea real, digamos* [0,1]. *Es afín a problemas inferencia en procesos estocásticos donde la información muestral viene dada por una trayectoria parcial* $x(t), t \in [0, T]$ *del proceso estocástico* $\{X(t), t \geq 0\}$*."*

Con la *data "aemet"* del paquete de R *(fda.usc)* se presenta el siguiente ejemplos gráficos de un conjunto de datos que tiene una estructura funcional. aemet es una data formato lista que contiene una serie resumen diario de 73 estaciones meteorológicas españolas para el período 1980-2009. De la lista se escogió el conjunto de datos de temperatura (en grados Celsius) que tiene preestablecido el formato fdata. A lo largo de este capítulo, también se verán ejemplos prácticos como medidas resumen, representación en base, métodos de suavizado, entre otros. Necesarios para entender y razonar los resultados de los capítulos siguiente de simulación y aplicación a datos reales.

```
data("aemet")
datos<-aemet$temp
plot(datos)
```

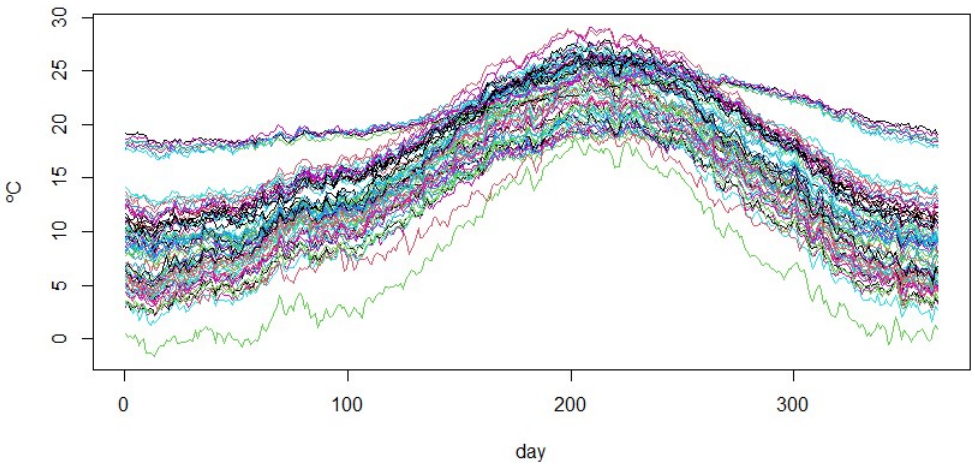


Figura 2.1: Temperaturas medias (en grados Celsius) diarias de 73 estaciones meteorológicas españolas. Período 1980-2009. *"data aemet".*

En la figura 2.1 se muestran 73 curvas de temperaturas medias con oscilaciones como dientes de cierra por su variabilidad en intervalos de tiempos cercanos. En cuanto a su comportamiento, se evidencia un patrón de campana donde las temperatura media van subiendo hasta mediados del período llegando a valores máximos que rondan entre 16° y 29° y luego comienza a descender.

Para tratar la discontinuidad de las curvas más adelante se tratará el como representar en bases funcional y métodos de suavizado que eviten al máximo la perdida información.

2.1.1. Estadísticos descriptivos

En la estadística clásica y en el análisis multivariante toda exploración de datos inicia por al menos saber cual es la media que representa los datos, la variabilidad y la relación que guardan dos o más variables, si esto es posible en los dos escenarios anteriores, también es posible para datos funcionales $x_1, x_2, ..., x_n$ *en* $t \in [a, b]$. *Carrillo y Garatejo (2016)* quienes citan a *Ramsay y Silverman (2005)* describen las expresiones siguientes:

- *Media:* $\overline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$
- *Varianza:* $var_x(t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(t) - \overline{x}(t))^2$
- *Desviación estándar:* $Ds(t) = \sqrt{var_x(t)}$
- *Covarianza:* $cov_x(t_1, t_2) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(t_1) - \overline{x}(t_1))(x_i(t_2) - \overline{x}(t_2))$
- *Correlación:* $corr_x(t_1, t_2) = \frac{cov_X(t_1,t_2)}{\sqrt{var_X(t_1)var_X(t_2)}}$

2.2. Espacio funcional

La definición 2.1 habla sobre un espacio de dimensión infinita (espacio funcional), una variable aleatorias es una variable funcional si toma valores en un espacio funcional que mejor describa su comportamiento. Para más profundidad puede consultarse a *Muscat (2014)* y para aspectos puntuales la presentación de *Febrero-Bande (2020)*. Estos espacios pueden ser de tres tipos: Métrico o semimétrico, de Banach o normado y de Hilbert. A continuación de forma precisa se citan algunas características particulares y propiedades importantes de los espacios funcionales.

I ***Espacio métrico o semimétrico:*** *En estos se dispone únicamente de las distancias, la cual, proporciona la medida de que tan cerca se encuentran los elementos entre si, lo relevante del espacio métrico, no es la estructura interna de los puntos de la curva, sino, su relación con otros puntos exteriores a lo largo de ella. En* $\mathbb{R}^2$. $d : F * F \to \mathbb{R}$*).* Propiedades principales.

1. $d(X, Y) \geq 0$ *(no negativo)*
2. $d(X, Y) = 0 \Longleftrightarrow X = Y$
3. $d(X, Y) = d(Y, X)$ *(simetría)*
4. $d(X, Z) \leq d(X, Y) + d(Y, Z)$ *(desigualdad triangular)*

II ***Espacio de Banach o normado:*** *El uso de una norma agrega la propiedad de escalabilidad.* $\mathcal{L}_P[S, \mu] = f : S \to \mathbb{R} s.t. \int |f(t)|^P d\mu < \infty$ *sabiendo que* (S, μ) *es una medida en espacio con* $1 \leq P < \infty$ *(espacio normado o seminormado).*

Formalmente, *Muscat (2014)* define un **espacio normado** X como un espacio vectorial sobre $F = \mathbb{R}$ o $\mathbb{C}$ con una función llamada norma $\|.\| : X \to \mathbb{R}$ tal que para cualquier $x, y \in X, \lambda \in \mathbb{F}$ se cumplen al menos las siguientes propiedades

1. $\|\alpha X\| = |\alpha|\|X\|$ *(estabilidad positiva)*

2. $\|X+Y\| \leq \|X\| + \|Y\|$ *(desigualdad triangular)*
3. *Si* $\|X\| = 0 \Rightarrow X = 0$

III ***Espacio de Hilbert*** $\mathcal{L}$ ***:*** *es un espacio vectorial con una función de producto interior donde no solo se tiene el concepto de longitud, también se tiene el concepto de ortogonalidad entre vectores.*

Propiedades principales:

1. $\langle X, X\rangle \geq 0$ *(positivo) y* $\langle X, X\rangle = 0 \Longleftrightarrow X = 0$
2. $\langle X, Y\rangle = \langle Y, X\rangle$ *(Simetría)*
3. $\langle \alpha X + \beta Y, Z\rangle = \alpha\langle X, Z\rangle + \beta\langle Y, Z\rangle$ *(lineal)*

Consideramos un espacio de Hilbert H *separable con un producto interior* $\langle\cdot,\cdot\rangle$ *que genera una norma* $\|\cdot\|$*, y se denota por* $\mathcal{L}$ *el espacio de operadores lineales acotados (continuos) con su norma.* Esta definición y más detalles puede consultar *Horváth-Kokoska (2012)* en el cap.2.

$$\|\psi\|\mathcal{L} = sup\{\|\psi(x)\| : \|x\| \leq 1\}$$

$$\psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j\rangle f_j, x \in H \tag{2.1}$$

donde $\psi \in \mathcal{L}$ es el operador, ν_j y f_j son bases ortonormales y λ_j es una secuencia real que converge a 0. El espacio de Hilbert $\mathcal{L}^2$ es el espacio esencial para el manejo de múltiples procedimientos del análisis de datos funcionales.

El Espacio $\mathcal{L}^2$

- El espacio $\mathcal{L}^2 = \mathcal{L}^2([0,1])$ es el conjunto de funciones de valor real medible x definida en un intervalo $[0,1]$ que satisface $\int_0^1 x^2(t)dt < \infty$ separable como el producto interior.

$$\langle x, y\rangle = \int x(t)y(t)dt. \tag{2.2}$$

Para un operador integral en $\mathcal{L}^2$ la expresión 2.2 es igual a

$$\psi(x)(t) = \int \psi(t,s)x(s)ds, x \in \mathcal{L}^2. \tag{2.3}$$

Para un operador tipo Kernel $\psi(\cdot,\cdot)$ la expresión 2.2 es igual a

$$\int\int \psi^2(t,s)dtds < \infty. \tag{2.4}$$

si se cumple esta condición entonces

$$\|\psi\|_S^2 = \int\int \psi^2 dtds.$$

si la función del operador en 2.3 es simétrica y en la expresión 2.4 es ≥ 0 *(positiva)* entonces se resolverían como una sumatoria

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j \nu_j(S) \quad en \quad \mathcal{L}^2([0,1] * [0,1]). \tag{2.5}$$

Función de media y covarianza en $\mathcal{L}^2$

De un total de curvas de la población N se selecciona aleatoriamente una muestra $X_1, X_2, ..., X_N$ iid. en $\mathcal{L}^2$, donde supone independencia entre cada curva X_i y son integrable al cuadrado. Define los siguiente parámetros de la forma:

- $\mu(t) = E[X(t)]$ ***(función media)*** ;
- $c(t,s) = E[(X(t) - \mu(t))(X(s) - \mu(s))]$ ***(función covarianza)***;
- $C = E[\langle (X - \mu), \cdot \rangle (X - \mu)]$ ***(operador covarianza)***

Considerando una muestra de curvas los estimadores se calculan de de la forma siguiente:

- $\hat{\mu}(t) = N^{-1} \sum_{i=1}^{N} X_i(t)$ ***función media muestral*** ;
- $\hat{c}(t,s) = N^{-1} \sum_{i=1}^{N} X_i(t) - \hat{\mu}(t))(Xi(s) - \hat{\mu}(s))]$ ***función covarianza muestral***;
- $\hat{C} = N^{-1} \sum_{i=1}^{N} \langle X_i(t) - \hat{\mu}, x \rangle (X_i - \hat{\mu}), x \in \mathcal{L}^2$ ***operador covarianza muestral***

El estimador de parámetro μ es consistente $E(\hat{\mu}) = \mu$, insesgado $E\|\hat{\mu} - \mu\|^2 = 0(N^{-1})$, los $X_1, X_2 \in \mathcal{L}^2$ integrables al cuadrado son independientes, si $E(X_1) = 0$ entonces $E[\langle X_1, X^2 \rangle] = 0$. Los estimadores covarianza escalar $\hat{c}(t,s)$ y el de covarianza de Hilbert $\hat{c}$ son estimadores sesgados que depende de una norma adecuada para calcular las distancia, es preferible asumir que es conocida e igual a cero.

Que la media de X_i sea cero en todos los casos, no es cierto, las suposiciones anteriores pueden convertirse en $X_i = \mu + Y_i$, donde $EY_i = 0$, en $\mathcal{L}^2$ esto significa que $EX_i(t) = \mu(t)$ para todo $t \in [0,1]$. Las variables se refieren a una media ajustada $X_i - \mu$, para casos de muestras $X_i - \overline{X}$. La expansión de la media y covarianza para aplicable solo para observaciones X_i de medias constantes puede consultarse en *Horváth y Kokszka (2012)*, además de su planteamiento de hipótesis para contrastar si la media es constantes o tienen puntos de cambio.

```
data("aemet")
datos<-aemet$temp
plot(datos,col= 3)
lines(func.mean(datos),col= "red",lwd=2)
legend(x=1,y= 25,legend = "media",col = "red",lty = 1, lwd = 1)
```

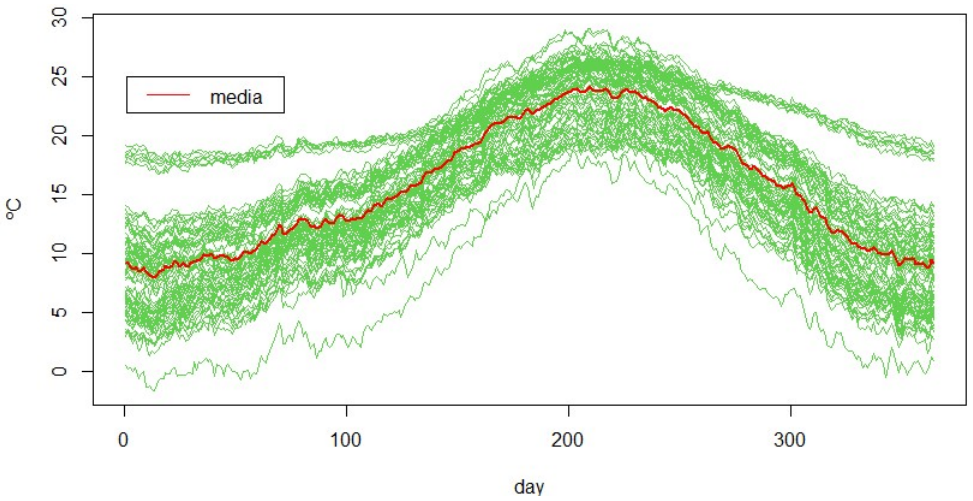


Figura 2.2: Media de temperaturas medias(en grados Celsius) diarias de 73 estaciones meteorológicas españolas. Período 1980-2009. *data "aemet".*

En la figura 2.2 se muestra el conjunto de curvas promedio de temperaturas en grados Celsius y su media en color rojo.

2.3. Representación en base

La representación en base es un proceso de combinaciones lineales que permite ver los datos como funciones adaptables y almacenables para obtener información. Existen dos tipos: **Las bases fijas:** de las principales están las de Fourier, B-splines, Wavelets y **las bases basadas en los datos:** Componentes Principales y Mínimos Cuadrados Parciales. Estas bases tienen una sobre la otra peculiares ventajas que particularmente su utilidad dependerá del comportamiento del conjunto de curvas. Cada registro n pares discreto t_j, y_j funciona como una entidad individual, donde y_j *es un instante de la función* t_j. Para fines aplicativos se trabajaría con muestras de datos funcionales n_i pares $t_{ij}, y_{ij}, j = 1, ..., n_i$, comúnmente aplicados a muestras de tipo serie temporal, pero también funcionan bastante bien para tipo espacial, frecuencia, peso, índice, ratio, tasa, etc.

Las bases fijas: son un conjunto de funciones básicas conocidas ϕK, matemáticamente independiente y con propiedades de aproximación de una suma ponderada o una combinación lineal de números K suficientemente grandes de esta función. Esto puede consultarse en *Ramsay-Silverman (2005)*. Expresado de forma siguiente.

$$x(t) = \sum_{k=1}^{K} C_k \phi K(t). \tag{2.6}$$

donde $x(t)$ es la función de expansión lineal, K son las funciones de la base ϕ_k, C es el vector de longitud K y ϕ vector funcional de la base ϕ_k. Su forma matricial está dada por

$$x = C^{'} \phi = \phi^{'} C. \tag{2.7}$$

Para un conjunto de datos su representación en bases no es derivable si la misma representación funciona bien, es precisa en su estimación y tiene *oscilaciones pequeñas pero en altas frecuencias* (curvas suaves).

$$D\hat{x}(t) = \sum_{k=1}^{K} C_k D\phi K(t) = \hat{C}^{'} D\phi(t). \tag{2.8}$$

- **Bases de Fourier**: son bases periódicas donde los t_j están igualmente espaciados. Estas son muy eficientes en L^2 (espacio de Hilbert). Su función lineal de expansión está dada por

$$\hat{x}(t) = C_0 + C_1 \sin \omega t + C_2 \cos \omega t + C_3 \sin 2\omega t + C_4 \cos 2\omega t + ... \tag{2.9}$$

  donde *la base* $\phi_0(t) = 1$,$\phi_{2_{r-1}}(t) = \sin r\omega t$ *y* $\phi_{2_r}(t) = \cos r\omega t$. ω *determina el período* $\frac{2\pi}{\omega}$.

  Una de las ventajas de esta base de funciones es que su derivada es facilmente calculable.

$$D \sin r\omega t = r\omega \cos r\omega t$$

$$D \cos r\omega t = -r\omega \sin r\omega t$$

- **Bases de B-spline**: son esenciales en datos no periódicos y funcionan combinando polinomios de manera flexible para los subintervalos que figuran como nodos. Cada nodo consta de grado de polinomio que determinan los grados de libertas *(orden de los polinomios más el número de*

puntos de corte interiores).

Los parámetros de una función de spline *un nodo por punto de ruptura es el orden orden más el número de nodos interiores,* $m + L - 1$.

$$S(t) = \sum_{k=1}^{m+L-1} C_k B_k(t, \tau) \tag{2.10}$$

B-spline definida positiva donde $B_k(t, \tau)$ *indica el valor en* t *de la función de B-spline definida por la secuencia de interrupción* τ *y* k *es el número de nodos en* t.

Con el conjunto de datos *aemet* se verá un ejemplo de código de representación por B-spline usando la función *create.bspline.basis y Data2fd* del paquete *(fda de R)*, para representar en otras bases fijas están las funciones *create.constant.basis, create.exponential.basis, create.fourier.basis, create.monomial.basis, entre otras* del paquete anteriormente mencionado. Estas funciones de representación se encuentran compiladas en una función llamada *create.fdata.basis* del paquete *(fda.usc de R)* con el argumento *type.basis* para seleccionar el tipo de representacion en base que sugiere el conjunto de datos, además, para representar en bases que dependen de los datos se encuentran las funciones *create.pc.basis y create.pls.basis* del mismo paquete.

```
data("aemet")
datos<-aemet$temp
bsp<-create.bspline.basis(rangeval = datos$rangeval, nbasis = 11)
bsp_aemet<-Data2fd(argvals = datos$argvals,y=t(datos$data),basisobj = bsp)
plot(datos[1],col="black",lwd=2)
lines(bsp_aemet[1],col=2,lwd=2)
legend(x=1,y=18,legend = c("data","bsp"),col = 1:2,lty = 1,lwd = 1)
```

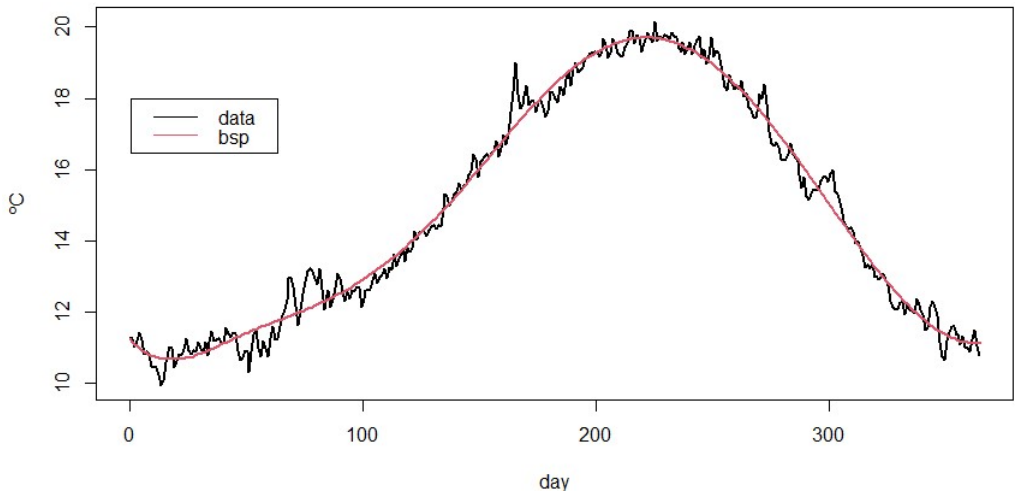


Figura 2.3: Temperaturas (en grados Celsius) diarias de una estación meteorológica española y su representación en base de B-spline ($K_x = 11$). Período 1980-2009. *data "aemet".*

En la figura 2.3 se muestra una curva de las 73 curvas promedio de temperaturas diarias de la data *aemet*, encima de ella una curva suave y continua correspondiente a la representación por B-spline con un número de bases $K_x = 11$. Esta representación no sigue los picos y valles cercanos de la curva original, pero si su trayectoria en forma de campana. La perdida de información de esta representación se verá reflejada en las estimaciones de los modelos regresión al momento de comparar los resultados con otra representación con mayor precisión.

- **Bases de Wavelets:** Permite construir *una base para toda función* $(-\infty, \infty)$ *integrables al cuadrado con una función de ondícula madre* ψ. Lo importante de esta representación en base es que su expansión tiene la capacidad para captar la discontinuidad de las curvas y los cambios repentinos con números bajos de coeficiente.

$$\psi_j k(t) = 2^{\frac{j}{2}} \psi(2^j t - k) \tag{2.11}$$

donde j y k son valores enteros.

De todas las bases fijas las más usadas son las anteriormente mencionadas debido a que en aplicación son potencialmente ortogonales ($XX^T = X^T X = I$).

- Otras bases fijas importantes son la exponencial, potencia, polinomial, poligonal, función escalonada, bases constante, empírica y de diseño, las cuales por los nombres se intuye como funcionan sus propiedades al momento de representar un conjunto de curvas.

Las bases basadas en los datos utilizan la información de la matriz de varianza y covarianza del conjunto de datos para representarlos.

- **Base de componentes principales funcionales (FPC):** En $\mathcal{L}^2$ son capaces de *reducir curvas dimensiones infinitas de datos funcionales a vectores p-dimensiones finitas de manera óptima* utilizando una base ortonormal empírica óptima u_k que minimizan a $\hat{S}^2$. . Ver el cap.3 *Horváth-Kokoska (2012)* y los capítulos 8, 9 y 10 de *Ramsay and Silverman (2005).*

$$\hat{S}^2 = \sum_{i=1}^{N} \left\| x_i - \sum_{k=1}^{p} \langle x_i, u_k \rangle u_k \right\|^2 \tag{2.12}$$

Con una aproximación las curvas x_i son reemplazadas por la aproximación $\sum_{k=1}^{p} \langle x_i, u_k \rangle u_k$. Sabiendo que *las funciones* $u_1, u_2, ..., u_p$ *son iguales (hasta un signo) a la función propia normalizada del operador de covarianza muestral.*

$$x_i = [\langle x_i, u_1 \rangle, \langle x_i, u_1 \rangle, ..., \langle x_i, u_p \rangle]^T. \tag{2.13}$$

Una función de componentes principales funcionales (FPC) se define como la función propia del operador de covarianza muestral $\hat{C}$ cuando las observaciones $X_1, X_1, ..., X_N$ tiene la misma distribución aleatoria X con valor en $\mathcal{L}^2$ integrable cuadrado. El producto interior es igual a $\langle X_i, \hat{v}_j \rangle = \int X_i(t)\hat{v}_j(t)dt$. Donde $\hat{v}_j = u_j$ es la contribución del FPC a la curva X_i.

Bajo el supuesto de que $EX_i = 0$ el FPC puede verse como *"la varianza muestral de los datos en la dirección de la función de* X*"* más correlacionada $\langle \hat{C}(x), x \rangle$.

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, x \rangle^2 = \left\langle \hat{C}(x), x \right\rangle \tag{2.14}$$

si se satisface el teorema 3.2 del cap.3 *Horváth-Kokoska (2012).*

Teorema 3.2. *Suponga que* Ψ *es un operador de Hilbert-Schmidt definido positivo simétrico con funciones propias* v_j *y valores propios* λ_j *que satisfacen*

$$sup\,\{\langle \Psi(x), x \rangle : \|x\| = 1, \langle x, v_j = 0, \rangle, 1 \leq j \leq i-1, i < p\} = \lambda_i, \tag{2.15}$$

y el supremo se alcanza si $x = v_i$. La función maximizadora es única hasta el signo.

Suponiendo que se quiere encontrar una dirección ortonormal $\hat{v}_2$ más correlacionada con la varianza de los datos. Se tiene una base en R^N

$$\frac{1}{N}\sum_{i=1}^{N}\|X\|^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle X_i, \hat{v}_j\rangle^2 = \sum_{j=1}^{N}\frac{1}{N}\sum_{i=1}^{N}\langle X_i, \hat{v}_j\rangle^2 = \sum_{j=1}^{N}\hat{\lambda}_j \tag{2.16}$$

Finalmente, los datos funcionales con un FPC especifico se generan de la forma siguiente

$$x_n(t) = \sum_{j=1}^{P} a_j Z_{jn} e_j(t) \tag{2.17}$$

donde a_j son los números reales, para cada n, Z_{jn} son variables aleatorias con media cero y varianza unitaria y e_j son funciones ortogonales con norma unitaria.

```
data("aemet")
datos<-aemet$temp
pc_aemet<-create.pc.basis(fdataobj = datos,l=1:3)
pc_aemet<-fdata2pc(datos,ncomp = 3)
plot(datos[1])
lines(pc$fdataobj.pc[1],col=2)
legend(x=1,y=18,legend = c("data","pc"),col= 1:2,lty = 1,lwd = 1)
```

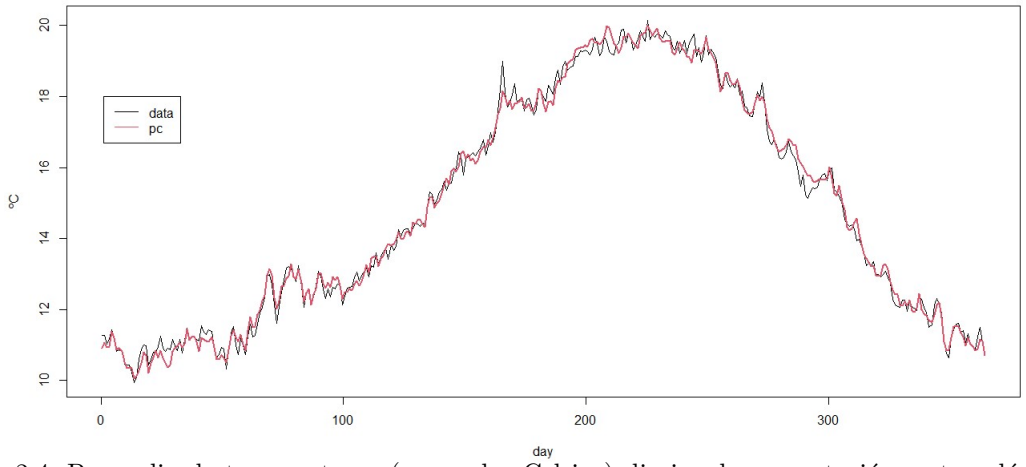


Figura 2.4: Promedio de temperaturas (en grados Celsius) diarias de una estación meteorológica española y su representación en base de componentes principales ($K_x = 3$). Período 1980-2009. *data "aemet".*

En la figura 2.4 se muestra una de las 73 curvas de temperaturas medias de la data *aemet*, encima de ella una curva correspondiente a la representación por componentes principales con un número de bases $K_x = 3$, a diferencia de las bases fijas de B-spline (ver figura 2.3), esta representación trata de seguir los picos y valles de la curva original durante todo el período.

```
plot(pc$basis,col=2:4,xlab = "day")
abline(h=0)
legend(300,-0.10,legend = c("pc1","pc2","pc3"),col = 2:4,lty = 1,lwd = 1)
```

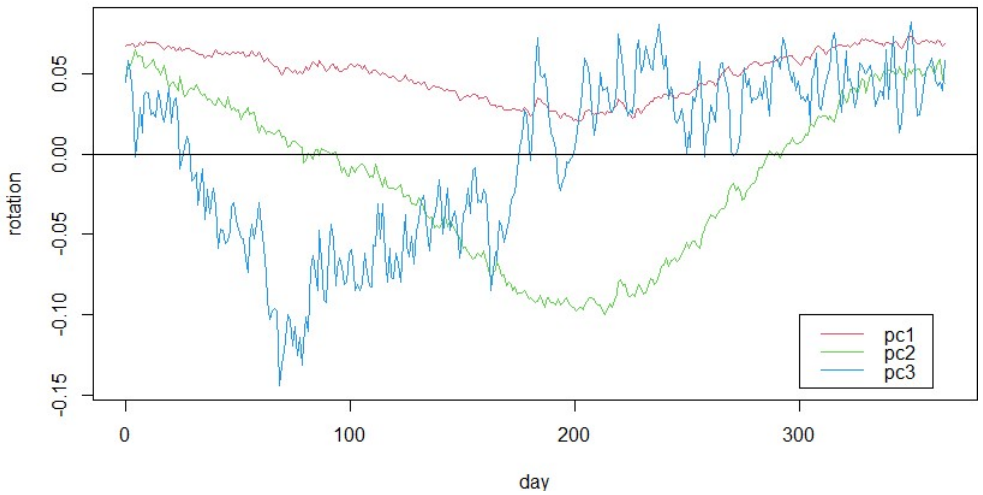


Figura 2.5: Comportamientos de las componentes principales ($K_x = 3$) para representar la *data "aemet"*.

En la figura 2.5 se ve el efecto individual de cada una de las componentes para estimar las curvas de temperaturas medias. La primera componente representa las curvas que a lo largo del período se mantienen por encima de la media del conjunto, la segunda representa las curvas que en los primeros años están por encima de la media, se ponen por debajo durante un largo período y luego vuelven a ponerse por encima de la media en los últimos años. Finalmente, la tercera componente representa las curvas que en un período corto están por encima de la media, se ponen por debajo de la media hasta mediados del período y el resto se mantienen por encima de la media.

- **Base de Mínimos Cuadrados Parciales (PLS):** pueden verse desde una perspectiva exploratoria donde la relación de las matriz de dos variables no excede la importancia de la otra *(análisis simétrico)*, cuando hay una variable que explica el comportamiento de la otra es un *(análisis asimétrico)*, esto puede estudiarse en *Burgillo (2015)*. Los dos tipos de análisis se determinan a partir de la matriz de covarianza de la misma forma y bajo las mismas restricciones que en las bases de componentes principales $X(t) \in \mathcal{L}^2(T)$ e $Y \in (\mathbb{R}^P, \mathcal{L}^2(T))$, solo que en este caso, maximizando la covarianza ente X y Y.

$$\frac{COV^2(\langle b, X\rangle, \langle c, Y\rangle)}{\|b\|^2\|c\|} \tag{2.18}$$

Donde $b_1 \in \mathbb{R}^d$ y $c_1 \in \mathbb{R}^q$ son vectores unitarios que se obtienen maximizando e imponiendo la condición de ortogonalidad. Tienen aplicación en los modelos de regresión, clasificación supervizada y otros, ya que se define de forma equivalente minimizando una suma apropiada de residuos al cuadrado. *Cuevas (2012)* para un modelo de regresión funcional estandar con respuesta escalar $Y = \alpha + \int_0^1 \beta(t)X(t)dt + \epsilon$ utiliza funciones de base $\Psi_i, ..., \Psi_p$ PLS donde los elementos son elegidos iterativamente maximizando en Ψ la función.

$$f(\Psi) = cov(Y - \eta_{p-1}(X), \int_0^1 X(t)\Psi(t))dt, \tag{2.19}$$

donde $\eta_0 = \mathbb{E}(Y)$ y η_{p-1} es la aproximación de $\eta(x)$ basada en las funciones de PLS. Para mayor ampliación ver a *Cuevas (2012)*.

2.4. Suavizado

El suavizado de datos funcionales y su necesidad es producto de la perdida de información de la curvas muestrales discretizadas relacionadas al contexto estadístico. Múltiples procedimientos paramétricos y no paramétricos se han desarrollado para recuperar parte de esa información perdida haciendo uso de la teoría clásica de regresión. Dentro de los métodos habituales en $\mathcal{L}^2$ se encuentra el modelo lineal funcional, mínimos cuadrados penalizados, sesgo y varianza, spline que se pueden consultar en *Hsing-Eubank (2015)* y los métodos *Kernel en Ferraty-Vieu (2006).*

- **Modelo lineal funcional:** tiene una media o pendiente real m en [0,1] donde las observaciones son (t_i, Y_i) discretas *con ruido aleatorio aditivo i.e.*

$$Y_i = m(t_i + e_i) \tag{2.20}$$

  En un espacio de Hilbert esto es

$$Y = \mathcal{F}m + \epsilon, \tag{2.21}$$

  donde $\mathcal{F} \in \mathfrak{B}(\mathbb{H}, \mathbb{Y})$ puede ser conocido, conocido no exactamente y aleatorio.

  El remplazo de m en $t \in [0,1]$ en las observaciones Y_i minimiza sus influencias locales individuales, teniendo resultados mas suaves y menos variables. En un proceso estocástico donde las observaciones toman forma (t_{ij}, Y_{ij}), $i = 1, ..., n$, $j = 1, ..., P$

$$Y_{ij} = x_i(t_{ij}) + \epsilon_{ij} = m(t_{ij}) + x_i(t_{ij}) - m(t_{ij}) + \epsilon_{ij}.$$

- **Mininos cuadrados penalizados:** Este proceso minimiza el criterio de mínimos cuadrados para encontrar a $\hat{m}$ óptimo que satisfaga que $g(t_i) = Y_i$.

$$\hat{m} = argmin_{g \in H_\delta} \sum_{i=1}^{n} (Y - g(t_i))^2 / n. \tag{2.22}$$

  donde

$$\mathbb{H}\delta := \left\{ g \in W_2[0,1] : \int_0^1 |g^{(2)}(t)|^2 dt \leq \delta \right\} \tag{2.23}$$

  Dada la complejidad de cálculo, en la práctica se usa el equivalente de

$$m_\eta = argmin_{g \in W_2[0,1]} \left(n^{-1} \sum_{i=1}^{n} (Y_i - g(t_i))^2 + \eta \int_0^1 |g^{(2)}(t)|^2 dt \right) \tag{2.24}$$

  Esta ecuación proporciona en su primera parte cuan confiable es el ajuste a través de la suma de residuos al cuadrado, en una segunda parte proporciona la suavidad necesaria para definir una curva en términos cuadráticos dado el parámetro regulador η, donde a mayores η mayor suavidad. Para una base H de dimensión finita $\phi_1, ..., \phi_q$ existen procedimientos para encontrar el parámetro óptimo de adaptación como:

  Validación Cruzada : que minimiza

$$CV(\eta) = n^{-1} \sum_{k=1}^{n} (Y_k - \mathcal{F}_k m_\eta^{[k]})^2$$

Validación Cruzada Generalizada: donde se evalúan de residuos $Y-\mathcal{F}m_\eta$ y los elementos de la diagonal $\mathcal{A}(\eta)$ para elegir η.

$$GCV(\eta) = \frac{\|Y - \mathcal{F}m_\eta\|_Y^2}{\left(1 - n^{-1}trace\mathcal{A}(\eta)\right)^2}$$

donde $\mathcal{A}(\eta) = U(U^TU + \eta V)^{-1}U^T$

- **Kernel:** Es un suavizado de la densidad de los datos que selecciona un ancho de banda óptimo *(H ventana óptima)*. Tiene su relevancia en el análisis multivariente de cálculo de estimadores de densidad no paramétrica y su suavizado local que puede llevarse al contexto de datos funcionales.

  En el caso clásico la ponderación local es

$$\Delta_i = \delta_i(x, h, K) = \frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

  donde K es la ponderación Kernel (caja, triangular, cuadrático, gaussiano, etc), h es el parámetro de suavizado (ancho de banda) y x es un número real fijo. Este entorno puede extenderse al caso multivariante donde se tienen vectores fijos $X_1, X_2, ..., X_n \in \mathbb{R}^P$,donde $K\cdot$ esta definida como un producto de funciones p reales del Kernel $K_1, ...K_p$

$$\Delta_i = \frac{1}{p}K \cdot \left(\frac{x - X_i}{h}\right)$$

[1]

  Fácilmente, este caso multivariante, también puede extenderse al caso funcional transformando los n variables aleatorias reales en n cantidades

$$\frac{1}{V(h)}K\left(\frac{d(\chi, \mathcal{X}_i)}{h}\right)$$

  donde d es una distancia semimétrica y K es un núcleo real (asimétrico) y $V(h)$ de $B(\chi, h) = \{\chi' \in E, d(\chi, \chi') \leq h\}$, entonces la ponderación del núcleo funcional es

$$\Delta_i = \frac{K\left(\frac{d(\chi, \mathcal{X}_i)}{h}\right)}{\mathbb{E}\left(K\left(\frac{d(\chi, \mathcal{X}_i)}{h}\right)\right)} \tag{2.25}$$

```
data("aemet")
datos<-aemet$temp
sbsp<-optim.basis(datos,numbasis = seq(5,19,by=2),
+                           lambda = seq(-1,1,by=0.10),
+                           type.basis = "bspline",type.CV = GCV.S)
> sbsp$numbasis.opt # número óptimo de bases
[1] 17
> sbsp$lambda.opt
[1] 1
> sbsp_aemet<-sbsp$fdata.est #
> plot(datos[1],col="black",lwd=2)
> lines(bsp_aemet[1],col=2,lwd=2)
> lines(pc$fdataobj.pc[1],col=3)
> lines(sbsp_aemet[1],col=4,lwd=2)
> legend(x=10, y=19,legend = c("data","bsp","pc","sbsp"),col = 1:4,lty = 1,lwd = 1)
```

[1] $\frac{1}{h_p}$ *para este caso es proporcional al volumen de conjunto donde esta cada X_i*

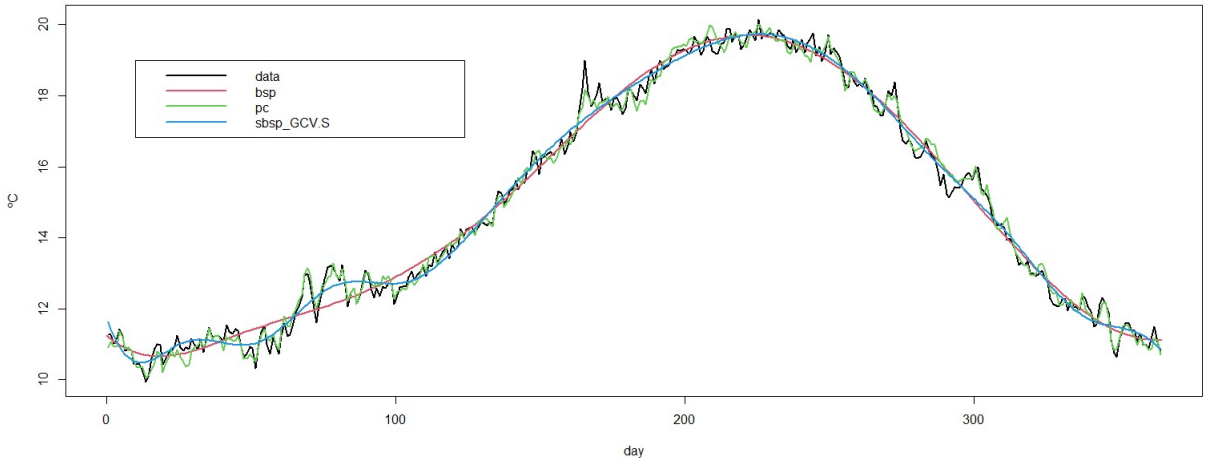


Figura 2.6: Comparativo de la representación en base y el suavizado GCV para las temperaturas medias diarias (en grados Celsius) de una estación meteorológica española. Período 1980-2009. *data "aemet".*

En la figura A.21 se compara la representación en base por B-spline, componentes principales y un suavizado por GVC con base de B-spline para una de las curvas de las temperaturas medias, siendo la representación PC la de mejores resultados. El suavizado obtiene un número óptimo de bases $K_x = 17$, mejora ligeramente la representación la base de B-spline cuando $K_x = 11$ en los primeros años, el resto del período se comportan de manera similar.

```
par(mfrow=c(1,2))
plot(datos,lwd=2, main = "(1)")
plot(sbsp_GCV.S_aemet,lwd=2, main = "(2)")
```

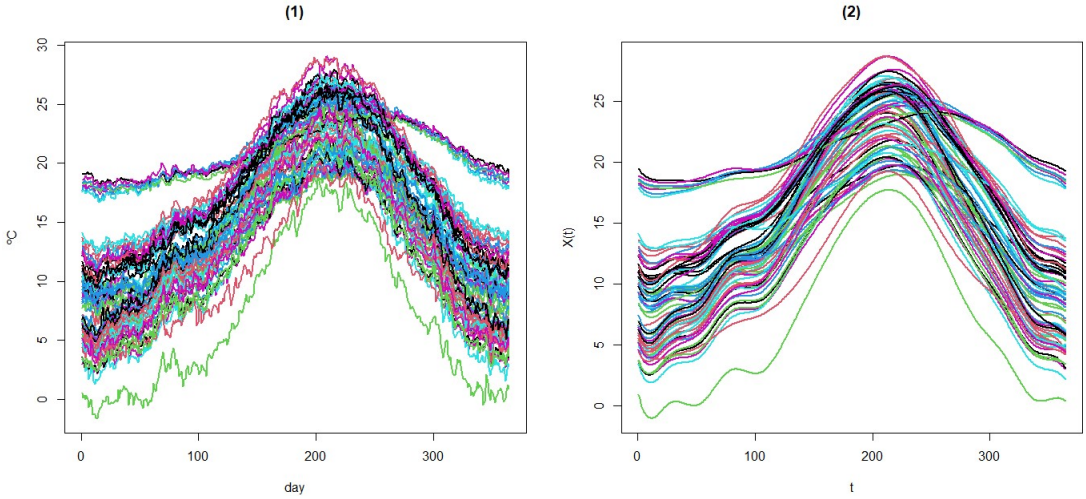


Figura 2.7: En la figura (1): las temperaturas medias (en grados Celsius) diarias de las estaciones meteorológica española. En la figura (2): el suavizado GCV con base de B-spline para las temperaturas medias diarias (en grados Celsius) de las estaciones meteorológica española. Período 1980-2009. *data "aemet".*

2.5. Uso de las derivadas

Las derivadas en el análisis de datos funcionales tiene su relevancia en la parte de exploración debido a que puede destacar componentes de variación e información que no quedan del todo claros en el conjuntos de curvas originales. Usualmente, figuran como transformaciones de las medidas de profundidad de los datos[2] y serán útiles para análisis posteriores como los modelos de regresión, clasificación y otros métodos.

Expresadas las observaciones $X_n(t)$ en una base particular $B_m(t)$ según *Kokoszka y Reimherr (2017)* se debe de *cumplir que la trayectoria de las observaciones son continuamente diferenciables y que las funciones básicas elegidas deben tener al menos K derivadas.* Las bases de Fourier para curvas periódicas cumplen con las dos condiciones y se pueden derivar infinitamente ,sin embargo, para las bases de B-spline solo posee número de derivadas igual a su orden.

Las curvas a menudo presentan cambios o variaciones irregulares con respecto a su comportamientos de su trayectoria y para estabilizarlas se podría considerar la derivación. En modelos de regresión el uso de derivadas es de suma importancia, principalmente en modelos económicos, meteorológicos y demás. En los modelos de regresión lineal puede ser un subproducto, pero también, puede vincularse en el modelo con los valores funcionales. Para ver más información consultarse los capítulos 17, 18 y 19 de Ramsay-Silverman (2005).

[2] $D(\overrightarrow{T}(X)) = \sum_{K=1}^{K} W_k D^k(T^k(X))$, *donde W es el vector de peso, D es la medida de profundidad y K la transformación.*

Bibliografía

Libro Ramsay J.O., Silverman B.W. (2005) Functional Data Analysis. Springer New York. Estados Unidos.

Libro Ferraty F., Vieu P. (2006) Nonparametric Functional Data Analysis. Springer. Estados Unidos.

Tesis Oviedo M., (2011) Utilities for Statistical Computing in Functional Data Analysis: The R Package fda.usc.

Tesis Aguilera M., (2009) Estimación Penalizada con Datos Funcionales. Granada.

Artículo Guccione J.A., Guccione J.J., Espacios Métricos. Departamento de Matemática. Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Presentación Febrero-Bande M. (2020) Computational aspects in Functional Data Analysis. Dpt. de Estatística, Análise Matemática e Optimización. Univ. de Santiago de Compostela.

Capítulo de libro Ramsay J.O., Silverman B.W. (2005) Functional Data Analysis. Cap 3. Springer New York. Estados Unidos.

Capítulo de Libro Horváth L., Kokoska P. (2012) Inference for Functional Data with Applications. Springer New York. Estados Unidos, PP 21-36.

Capítulo de Libro Horváth L., Kokoska P. (2012) Inference for Functional Data with Applications. Springer New York. Estados Unidos, PP 37-43.

Tesis Carrillo A. E., Garatejo O. c. (2016) Análisis de Datos Funcionales Aplicado en Electroencefalogramas: Agrupamiento por K-Medias Funcional. Bogotá, D.C.

Artículo Cuevas A., Febrero M., Fraiman R. (2007) Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics, 22(3): 481-496.

Artículo Febrero M., Galeano P., González W. (2008) Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetría 19(4): 331-345.

Artículo Febrero M., Galeano P., González W. (2007) A functional analysis of NOx levels: location and scale estimation and outlier detection. Computational Statistics (2007) 22: 411-427.

Capítulos de libro Ramsay J.O., Silverman B.W. (2005) Functional Data Analysis. Springer New York. Estados Unidos. Cap. 17, 18 y 19.

Tesis Ginzo-Villamayor M.J., Análisis Geostadístico de datos funcionales. Proyecto Fin de Máster - Máster en Técnicas Estadísticas. Proyecto 388.

Capítulo 3

Modelos de regresión funcional (FRM)

Este capítulo tiene como objetivo poner en contexto los modelos de regresión funcional (FRM) de los cuales podemos encontrar literaturas e investigaciones interesantes. *Ramsay-Silverman (2005)* explican los FRM desde una perspectiva paramétrica, donde se profundiza sobre los regresión lineal con respuesta escalar, con respuesta funcional y variable independiente funcional, respuesta funcional con covariables multivariantes, transformaciones como derivada y otros. Por otro lado, *Ferraty-Vieu (2006)* y *Pons (2011)* hacen grandes aportes a los FRM desde la estadística no paramétrica. *Ferraty (2011)* reúne múltiples investigaciones de diversos autores dedicados y apasionados con este campo del FDA y el FRM, siendo este un aporte importante entre la comunidad de investigadores de la estadística matemática y aplicada, de las investigaciones y aplicaciones sobre FRM que podemos consultar está la de *Aneiros et al. (2011)* en donde se usó del modelos funcional no paramétrico (FNPM) y lineal parcial semifuncional (SFPLM) para predecir de demanda residual en los mercados spot de electricidad, *Attouch et al. (2011)* con los FNPM para datos espaciales, *Bereswill y Johannes (2011)* utilizan los modelos funcionales lineales (FLM) para tratar los efectos de observaciones ruidosas, *Febrero y et al. (2011)* innovan con los modelos aditivos generalizados funcionales(FGAM) aplicados a los datos de *Tecator*, *Ferraty et al. (2011)* con los modelos funcionales aditivos (FAM) para el caso de una respuesta escalar. Además de estas investigaciones, dentro de este mismo documento existen otras investigaciones colaborativas que se pueden consultar. En la actualidad, el aporte a las medicina ha sido sustancial, *Ateba et al. (2020)* utilizó los modelo aditivo espectral generalizado funcional (FGSAM), modelo lineal funcional generalizado (FGLM) y modelo de aditivo de Kernel funcional generalizado (FGKAM) para predecir la incidencia de malaria en Malí. Particularmente, para este capítulo, se priorizará por estudiar los principales modelos de regresión funcional con respuesta un escalar debido a la naturaleza de datos que se utilizaran en el próximo capítulo de aplicación y estimación.

En los modelos de regresión funcional (FRM) se introducen los conceptos de covariables *(variables independientes)* para dar explicación al comportamiento de una variable regresada *(variable dependiente)*, estos conceptos son propios de los modelos de regresión clásico extendidos al pensamiento FDA. Para que sea FRM, es necesario que una de las variables que intervienen sea funcional. Un FRM con respuesta escalar se representa

$$Y = m(Z, \mathcal{X}) + \varepsilon, \tag{3.1}$$

donde $Y \in \mathbb{R}$ *es la variable respuesta (escalar),* m *es la función desconocida de la combinación de covariables,* $Z \in \mathbb{R}^P$ *es la covariables del vector,* $\mathcal{X} = X_j(t)_{j=1}^q$ *son las covariables funcionales y* ε *es el error de media cero.*

3.1. Regresión lineal funcional (FLR)

Las especificaciones de la ecuación anterior 3.1 corresponden a la matriz Z que pertenece a los modelos lineales multivariantes. Logra su extensión para los modelos de regresión lineal funcional (FLR) de Z_i y una aproximación suave mediante interpolación funcional de los parámetros minimizando una expresión cuadrática que se ajuste al vector de valores $E[\hat{y}] = y$. Este problema de minimización puede verse como un producto interior.

$$Y = E[Y|\mathcal{X}] = \alpha + \langle \mathcal{X}, \beta \rangle \tag{3.2}$$

donde α es el intercepto, β es es parámetro funcional estimado.

$$Y = \alpha + \langle \mathcal{X}, \beta \rangle + \varepsilon = \int_a^b \beta(t)\mathcal{X}(t)dt + \varepsilon \tag{3.3}$$

En $\mathcal{L}^2$ donde $T = [a, b]$ el producto interior se define en una integral en el intervalo $\int_a^b$ y su representación en base fija o basada en los datos. *Ramsay y Silverman (2005)* para su ejemplo de *Canadian weather stations* utilizó bases periódicas de Fourier, *Ginzo-Villamayor* utilizó bases de B-spline y las de Fourier para suavizar datos espacio-temporales *(Kriging)*. Para los pares $(X_1, y_1), ..., (X_n, y_n)$ los parámetros pueden escribirse

$$\mathcal{X}_i(t) = \sum_K^{K_x} C_{ik}\psi_k \rightarrow \mathcal{X} = C\Psi(t),$$

$$\beta(t) = \sum_K^{K_\beta} b_k\theta_k \rightarrow \beta = \theta'\mathbf{b}$$

sabiendo que $\mathbf{b} = (Z'Z)^{-1}Z'y, \hat{y} = CJ_{\psi\theta}by = Zby = Z(Z'Z)^{-1}Z'y = Sy$ donde $J_{\psi\theta} = \int \psi(t)\theta(t)dt$. Para más detalles puede consultarse *Ginzo-Villamayor (proyecto 388)*.

3.2. Regresión No paramétrica funcional (FRNP)

Los modelos de regresión no paramétrica funcional (FRNP) son descritos por *Ferraty et al. (2006 y 2011)*. Estos son escenciales para dar respuestas a relaciones complejas de variables con difícil apreciación de la relación entre ellas haciendo uso de métodos convencionales descriptivos y gráficos. La ecuación está dada por

$$Y = m(Z, \mathcal{X}) + \varepsilon$$

Donde Y es una variable aleatoria real , $\mathcal{X}$ es una variable aleatoria en $\mathcal{L}^2$ y ε es una variable aleatoria centrada $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. Sea $m(\mathcal{X}) = \mathbb{E}(Y|\mathcal{X} = x)$ asociado a un espacio *(umbral)* $s \in S = 1, 2$ donde cada par $(\mathcal{X}_i, Y_i)_{i=1}^n$ está definido en $Y_{i,s}^\tau$ para cada $i \in 1, ..., n$ y $s \in S$. En consecuencia el modelo resultante de la expresión de dos componentes dado por

$$Y_i = \sum_{s\in S} Y_{i,s}^\tau = \sum_{s\in S} m_S^\tau(Z_i, \mathcal{X}_i) + \varepsilon_i \quad para \quad i = 1, ..., n$$

Separando los componente se genera una familia de estimación nueva de $\hat{m}^\tau(Z, \mathcal{X}) := \sum_{s\in S} \hat{m}_s^\tau(Z, \mathcal{X})$ construidos cada uno a partir del estimador tipo kernel y ancho de banda dado por

$$\hat{m}^\tau(Z, \mathcal{X}) = \frac{\sum_{i=1}^n Y_{i,s}^\tau K(h_s^{-1}\|\mathcal{X} - x\|)}{\sum_{i=1}^n K(h_s^{-1}\|\mathcal{X} - x\|)} \tag{3.4}$$

Donde K es la función kernel y h_s es el ancho de banda. Los parámetros umbral τ y ancho de banda h_s determinados por validación cruzada (CV).

Luis Fidel Herrera de los Santos

3.3. Regresión aditiva funcional (FAR)

Los modelos de regresión aditivo funcional (FAR) por su flexibilidad tienen mayor capacidad para explicar y predecir que los FLR, dada su descomposición aditiva para modelar relaciones no lineales. Los últimos avances de *Ferraty et al. (2011)* muestran que *para un conjunto de datos centrados* (X, Y) *en* $Hx\mathbb{R}$*, donde* $H = h : \int_I h^2(t)dt < +\infty$*,* $I \in \mathbb{R}$ *es el intervalo en* $\mathcal{L}^2$ *con producto interior* $\langle g, f\rangle = \int_I g(t)f(t)$ *y norma* $\|g\|^2 = \langle g, g\rangle$. El problema de regresión esta dado por

$$Y = m(Z, \mathcal{X}) + \varepsilon$$

Donde $m(Z, \mathcal{X}) = \mathbb{E}[Y|\mathcal{X}, Z]$, se asume que $\mathbb{E}[\varepsilon|\mathcal{X}] = 0$ y la $\mathbb{E}[\varepsilon^2|\mathcal{X}] < \infty$. La función de regresión desconocida de m por aproximación se expresa como una suma finita de términos

$$m(Z, \mathcal{X}) \approx \sum_{j=1}^{m} g_j^*(\langle \theta_j^*, \mathcal{X}\rangle) \tag{3.5}$$

donde $\theta_j^* \in H$ con $\|\theta_j^*\|^2 = 1$, g_j^* para $j = 1, ..., m$ son funciones univariadas que la suma de las mismas determinan a Y.

$$Y_i = \sum_{j=1}^{m^*} \hat{g}_j \hat{\theta}_j(\langle \hat{\theta}_j, \mathcal{X}_i\rangle) \tag{3.6}$$

La estimación combina aproximación de direcciones de spline normalizados y la estimación de regresión Kernel de Nayadara-Watson como procedimientos no paramétricos que mejoran los resultados de las predicciones luego de calcular los residuos y estimar un modelo entre los residuos y los regresores funcionales quedando regresión funcional 3.6. Su expansión y nommenclatura ajustada a 3.1 está dada por

$$\mathbb{E}[Y|\mathcal{X}, Z] = \left(\alpha + \langle Z, \gamma\rangle + \sum_{j=1}^{q} \langle \mathcal{X}^j, \beta_j\rangle\right) = \left(\alpha + Z^T\gamma + \sum_{j=1}^{q} \int_{a_j}^{b_j} X^j(t)\beta_j(t)dt\right)$$

3.4. Regresión aditiva generalizada funcional (FGAM)

Los modelos de regresión aditiva generalizada (FGAM) propuestos por *Febrero y et al. (2011)* y *Febrero y et al. (2013)* donde la idea principal es *adaptar los modelos aditivos generalizados multivariante a covariables de datos funcionales*, estos modelos son una extensión de los FGLM cuando el vinculo o la función de enlace g conocido o no pertenece a un procedimiento de una de las funciones de distribución de la familia exponencial. Se estiman utilizando las propiedades de los modelos no paramétrico, no está restringidos a relaciones lineales entre covariables, ni al espacio de Hilbert $\mathcal{L}^{\in}$, más bien, utiliza funciones de suavizado no paramétricos del modelo FAR tipo kernel y ancho de banda obtenidos mediante validación cruzada generalizada (GCV) aplicable para cualquier espacio por la información obtenida de las distancias entre covariables, por tanto, son más flexibles.

Sea Y una variable respuesta escalar que pertenece la familia de la distribución exponencial, $Z = \left\{\mathcal{X}^i\right\}_{i=1}^{P}$ es el conjunto de covariables en P espacio infinito. El modelo FGLM se escribe $\mathbb{E}[Y|\mathcal{X}, Z] = \mu = g^{-1}(\eta_{\mathcal{Z}}) = g^{-1}(\alpha + \langle Z, \beta\rangle)$. Donde $\langle Z, \beta\rangle = \sum_{i=1}^{P} \langle \mathcal{X}^i, \beta_i\rangle$ y g es el enlace funcional. La expansión de los modelos FGLM adaptar al modelo FGAM cuando hay información suficiente de g de la siguiente forma

$$\mathbb{E}[Y|\mathcal{X}, Z] = \mu = g^{-1}(\eta_{\mathcal{Z}}) = g^{-1}\left(\alpha + \sum_{k=1}^{M} r_k(Z^k) + \sum_{j=1}^{P} f_j(\mathcal{X}^j)\right) \tag{3.7}$$

donde $r_k(\cdot)$ es la función suave del escalar, $f_j(\cdot)$ es la función parcial desconocida suave estimada a partir de las puntuaciones *(score)* de la función de componentes principales (FPC) $\mathcal{X}^j$ y una función suave ξ_j^k de los modelos FAR de *Müller and Yao (2008)* para respuesta escalar de la forma $f_j(\mathcal{X}^j) := \sum_{k=1}^{K} f_j^k(\xi_j^k)$. A esta alternativa *Febrero y et al. (2013)* la denominó Modelo aditivo espectral generalizado funcional (FGSAM). Estas propiedades pueden aprovecharse de forma práctica para FRM de respuesta escalar obviando la función de enlace g quedando un ***modelo aditivo espectral funcional (FSAM)*** y su expresión de la forma siguiente.

$$\mathbb{E}[Y|\mathcal{X}, Z] = \mu = \eta_{\mathcal{Z}} = \alpha + \sum_{k=1}^{M} r_k(Z^k) + \sum_{j=1}^{P} f_j(\mathcal{X}^j) \tag{3.8}$$

Para mayor ampliación del tema se puede estudiar *Febrero y et al. (2011, 2013)* y *Müller and Yao (2008)*.

3.5. Modelos de regresión funcional con errores dependientes

Los modelos revisados anteriormente tiene una característica en común y es que los errores $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ se asumen que cumplen con las propiedades de los supuestos de independencia, homocedasticidad donde $\mathbb{E}(\varepsilon) = 0$, $\mathbb{V}ar(\varepsilon) = \sigma^2$ y $Cov(\varepsilon) = (\varepsilon\varepsilon^{'}) = \Omega \in S(n)$ definida positiva, en la práctica estas restricciones son difíciles de cumplir, principalmente cuando el dato no es necesariamente oportuno producto de diversas eventualidades en el contexto. Los modelos de mínimos cuadrados generalizados (FGLS) son la alternativa cuando las propiedades anteriores son cuestionables, están basados en el estimador de Gauss-Markov (GME) e incorporan la estimación de componentes de dependencia dentro del modelo mediante estructuras de covarianza de para el termino de error. *Kariya y Kurata (2004)* define *la estimación de mínimos cuadrados generalizados (GLSE) con matriz de covarianza desconocida reemplazada por un estimador apropiado.* Un modelo de regresión lineal funcional (FLM) extendido del contexto de la estadística clásica esta dado por

$$Y = \alpha + \langle \mathcal{X}, \beta \rangle + \varepsilon = \alpha + \int_a^b \beta(t)\mathcal{X}(t)dt + \varepsilon$$

$$\mathbb{E}(\varepsilon) = 0 \qquad Cov(\varepsilon) = E(\varepsilon\varepsilon') = \Omega \in S(n)$$

Donde $S(n)$ denota el conjunto de $n \times n$ matrices definidas positiva, el términos de error ε con matriz de covarianza $\Omega = \sigma^2\Sigma$ dada por una función de parámetro desconocido θ, pero estimable $\Omega = \sigma^2\Sigma(\theta)$ como una matriz diagonal aplicable para familia de modelos equi-correlacionado, heterocedástico, regresión aparentemente no relacionada (SUR), modelo de error y modelo de Anderson (AR) donde hay una matriz regresora X y la matriz de covarianza anteriormente mencionada.

Estimador de Gauss-Markov (GME)

$$b(\Sigma) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

Este modelo GME $b(\Sigma)$ no es factible cuando la matriz Σ es desconocida, para este caso lo ideal es un GLSE $b(\hat{\Sigma})$.

Función desconocida de parámetro estimable θ.

$$\Omega = \Omega(\theta)$$

donde $\Omega(.)$ es la forma funcional conocida. La estructura de covarianza es

$$\Omega = \sigma^2\Sigma(\theta)$$

con inversa $\Sigma(\theta)$

$$\Sigma(\theta)^{-1} = I_n + \lambda_n(\theta)C \qquad y \quad \theta \in \Theta \subset R^1$$

donde $\lambda_n(\theta)$ es una función continua de valor real en Θ y C es una matriz simétrica conocida que se diagonaliza Γ (ortogonal) de la siguiente manera

$$\Gamma' C \Gamma = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix} \equiv D \quad con \quad d_1 \leq ... \leq d_n.$$

Transformando a LMR

$$\Gamma' y = \Gamma' X\beta + \Gamma'\varepsilon \tag{3.9}$$

Para los errores la estructura de covarianza $Cov(\Gamma'\varepsilon) = \sigma^2\Gamma'\Sigma(\theta)\Gamma$, donde la inversa de $(\Gamma'\Sigma(\theta)\Gamma)^{-1} = I_n + \lambda_n(\theta)D$ es la matriz diagonal con j-enésimo elemento diagonal $1 + \lambda_n(\theta)d_j$ que incluye la familia de modelos anteriormente mencionados.

- **Modelo de error AR(1):** cuando en el modelo FLM el termino de error ε es generado por un proceso autoregresivo estacionario e invertible de orden $(p = 1)$ (AR(1)) donde el primer retardo *(pasado inmediato)* de error ε_{j-1} es estadísticamente significativo en la predicción de ε_j *(presente inmediato)* como una esperanza condicionada $E[\varepsilon_j|\varepsilon_{j-1}]$.

$$\varepsilon_j = \phi\varepsilon_{j-1} + \xi_j \qquad con \quad |\phi| < 1(j = 0, \pm1, \pm2, ...), \tag{3.10}$$

para $i > 1$ la expresión es de forma

$$\varepsilon_j = \sum_{i=1}^{p} \phi\varepsilon_{j-i} + \xi_j \qquad con \quad |\phi| < 1(j = 0, \pm1, \pm2, ...), \tag{3.11}$$

la matriz de covarianza se expresa como

$$\Gamma = \tau^2\Phi \qquad \Phi = \Phi(\phi) = \frac{1}{1-\phi^2}\left(\phi^{|i-j|}\right) \tag{3.12}$$

Otras propuestas con respecto a los modelos AR y su expansión a FDA que pueden consultarse, tal es el caso de los procesos Hilbertianos auto regresivos (ARH) y los procesos Hilbertianos autoregresivos con variables exógenas (ARHX) propuestos por *Damon y Guillas (2005)*, los modelos de dependencia de *Hörmann y Kokoszka (2010)* aplicados, tanto, a lineales como a dependencia no lineal, el análisis de dependencia desde un contexto no paramétrico de *Battey y Sancetta (2013)*.

Un FLM que en termino de error ε es generado por un proceso más complejo en términos de parámetros, pero, de igual manera estacionario e invertible denominado modelo *autoregresivo de medias móviles* de orden $(p = 1, q = 1) = \mathbf{ARMA}(1, 1)$, por tanto, el ε_j es explicado de forma lineal por un valor del pasado (ε_{j-1}) y un valor del pasado de un proceso de ruido blanco (a_{j-1}) de forma

$$\varepsilon_j = \phi\varepsilon_{j-1} + \xi_1 + \theta\xi_{j-1} \qquad con \quad |\phi, \theta| < 1(j = 0, \pm1, \pm2, ...), \tag{3.13}$$

para $i > 1$ la expresión es de forma

$$\varepsilon_j = \sum_{i=1}^{p} \phi\varepsilon_{j-1} + \xi_j - \sum_{i=1}^{q} \theta_1\xi_{j-1} \qquad con \quad |\phi, \theta| < 1(j = 0, \pm1, \pm2, ...), \tag{3.14}$$

- **Modelo Equi-correlacionado:** el termino de ε del modelo 3.3 satisface

$$Var(\varepsilon_i) = \tau^2 \qquad Cov(\varepsilon_i, \varepsilon_j) = \tau^2\theta \quad (i \neq j), \tag{3.15}$$

para algunos $\tau^2 > 0$ y $\theta \in \left(-\frac{1}{n-1}, 1\right) \equiv \Theta$, significa que, la matriz de covarianza Ω es

$$\Omega = \tau^2 \begin{pmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & & \vdots \\ \vdots & & \ddots & \theta \\ \theta & \cdots & \theta & 1 \end{pmatrix} \tag{3.16}$$

- **Modelo heterocedástico:** El modelo 3.3 siguen la siguiente estructura

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} : nx1, \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} : nxk, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix} : nx1 \tag{3.17}$$

y matriz de covarianza *Omega*

$$\Omega = \begin{pmatrix} \sigma_1^2 I_{n_1} & & 0 \\ & \ddots & \\ 0 & & \sigma_p^2 I_{n_p} \end{pmatrix} \in S(n) \tag{3.18}$$

donde p varianzas son distintas.

- **Modelo de regresión aparentemente no relacionadas (SUR):** Cuando el 3.3 sigue una estructura

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} : nx1, \quad X = \begin{pmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_p \end{pmatrix} : nxk, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} : kx1, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix} : nx1 \tag{3.19}$$

con estructura de covarianza

$$\Omega = \Sigma \otimes I_m \qquad y \quad \Sigma = (\sigma_{ij}) \in S(p), \tag{3.20}$$

donde $\otimes$ es el producto Kronecker para matrices $P = (p_{ij}) : axb$ y $Q = (q_{ij}) : cxd$ de la siguiente manera

$$P \otimes Q = \begin{pmatrix} P_{11}Q & \cdots & P_{1b}Q \\ \vdots & & \vdots \\ P_{a1}Q & \cdots & P_{ab}Q \end{pmatrix} : acxbd \tag{3.21}$$

Mínimos cuadrados generalizados (GLS)

La extension práctica del modelo clásico GLS de *Kariya and Kurata (2004)* a funcional se puede consultar en *Oviedo y et al. (2018)*, la propuesta fue utilizada para estimar la tasa de influenza en Galicia (España) con variables multivariantes y funcionales, aproximados por métodos de estimación de los modelos de mínimos cuadrados generalizados funcionales FGLS. La extensión en un espacio de Hilbert $\mathcal{L}^2$ está dada por

$$GLS(\beta,\phi) = (y - \langle \mathcal{X}, \beta\rangle)'\Sigma(\phi)^{-1}(y - \langle \mathcal{X}, \beta\rangle) \tag{3.22}$$

$\mathcal{X}_i$ y β dada una muestra se aproximan

$$\mathcal{X}_i(t) \approx \sum_{k}^{K_x} C_{ik}\psi_k(t), \quad \beta(t) \approx \sum_{k}^{K_\beta} b_k\varphi_k(t), \tag{3.23}$$

de forma matricial

$$X = C\Psi, \quad B = b'\varphi \tag{3.24}$$

donde X es una matriz nxM en $[a,b]$ y C es una matriz nxK con representación en base, ψ es la matriz K_xxM. B es la matriz $1xM$ con evaluaciones del parámetro β en en la cuadrícula, φ es la matriz $K_{\beta xM}$ con evaluaciones de la base φ_j y b es el vector de coeficientes de β en la base.

Los términos pueden ser aproximados al modelo lineal multivariado de la forma $\{\langle \mathcal{X}_i, \beta\rangle\} \approx C\Psi\varphi' b = Zb$, suponiendo que φ se conoce, se puede definir $W = \Sigma(\phi)^{-1}$ donde b y su MELI esta dada por

$$b_\Sigma = (Z'WZ)^{-1}Z'Wy,$$

con covarianza

$$Cov(b_\Sigma) = \sigma^2(Z'WZ)^{-1},$$

por tanto el ajuste se obtiene por la matriz H

$$\hat{y} = Z(Z'WZ)^{-1}Z'Wy = Hy$$

El criterio de estimación de los parámetros K_x y K_β relacionados con las bases para $\mathcal{X}$ Y β se obtiene minimizando

$$\min_{K_x,K_\beta,b,\phi} GLS = \min_{K_x,K_\beta,b,\phi} (y - Zb)'\Sigma(\phi)^{-1}(y - Zb)$$

dado los errores dependientes, este proceso de generación de datos utiliza *validación cruzada correlacionada generalizada (GCCV)* de la forma siguiente

$$GCCV(K_x, K_\beta, b, \phi) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i,b})^2}{\left(1 - \frac{tr(G)}{n}\right)^2}$$

Mínimos cuadrados generalizados iterativos (iGLS)

Los Modelos de regresión iterativo funcional de mínimos cuadrados generalizados (FIGLS) son una version alternativa de los FGLS que separa la estimación de la estructura de dependencia temporal correspondiente un **AR o ARMA** *(ver expresiones 3.11 y 3.14)* de los parámetros asociados a la regresión K_x, K_β, b de forma iterativa para reducir la carga computacional, esto y más detalles puede consultarse en *Oviedo et al. (2018)*. Es posible que en la práctica la mayoría de los modelos de dependencia se resuelvan con un modelo autorregresivo (AR) de orden $P = 1, 2$ a través de combinaciones

lineales de las componentes del pasado *(retardos)* como una esperanza condicional y la componente del termino aleatorio con las propiedades de ruido blanco, este número de parámetros se define a través de criterios, donde el más conocido es el *Criterio de Información de Akaike (AIC)*, definido el número de parámetros obteniendo el menor AIC. Para número de parámetros altos es factible aprovechar las propiedades conjuntas de los modelos AR y MA, las cuales permiten reducir conformemente el número de parámetros utilizando en AIC anteriormente mencionado o el *Criterio de Información Bayesiana de Schwarz (BIC)*, donde de varios modelos generados el de menor valor en ambos criterios $0 \leq p \leq P \quad y \quad 0 \leq q \leq Q$.

Bibliografía

Libro Ramsay J.O., Silverman B.W. (2005) Functional Data Analysis. Springer New York. Estados Unidos.

Libro Ferraty F., Vieu P. (2006) Nonparametric Functional Data Analysis. Springer. Estados Unidos.

Libro Pons O., (2011) Functional Estimation for Density, Regression Models and Processes. World Scientific Publishing Co. Pte. Ltd. Estados Unidos.

Tesis Oviedo M., (2011) Utilities for Statistical Computing in Functional Data Analysis: The R Package fda.usc.

Artículo Ferraty F., Goia A., Salinelli E., Vieu P. (2011) Recent Advances on Functional Additive Regression. Recent Advances in Functional Data Analysis and Related Topics, pp 97-102

Artículo Ferraty F., Martínez-Calvo A., Vieu P. (2011) Thresholding in Nonparametric Functional Regression with Scalar Response. Recent Advances in Functional Data Analysis and Related Topics, pp 103-109

Artículo Müller, H.-G., Stadtmüller, U. (2005) Generalized functional linear models. The Annals of Statistics. Vol. 33, No. 2, 774-805.

Artículo James G.M., Generalized Linear Models with Functional Predictors.Marshall School of Business, University of Southern California.

Capítulo de libro Harezlak J., Randolph T.W., (2011) Structured Penalties for Generalized Functional Linear Models (GFLM).F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 161-167.

Capítulo de libro Febrero-Bande M., González-Manteiga W., (2011) Generalized Additive Models for Functional Data. F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 91-96.

Artículo Febrero-Bande M., González-Manteiga W., (2013) Generalized Additive Models for Functional Data.Dpt. of Statistics and O.R. Faculty of Mathematics. Campus Vida 15782. Santiago de Compostela. Spain.

Artículo Damon J., Guillas S., (2005) Estimation and Simulation of Autoregressive Hilbertian Processes with Exogenous Variables. Statistical Inference for Stochastic Processes (2005) 8: pp 185-204.

Artículo Hörmann S., Kokoszka P., (2010) Weakly Dependent Functional Data. The Annals of Statistics Vol. 38, No. 3, Institute of Mathematical Statistics. pp 1845?1884.

Artículo Battey H., Sancetta A., (2013) Conditional estimation for dependent functional data. Journal of Multivariate Analysis.

Artículo Müller, H.G. and Yao, F. (2008) Functional Additive Model. Journal of the American Statistical Association, 103(484): 1534-1544.

Artículo Oviedo M., Febrero-Bande M., Muñoz M. P., Dominguez A. (2018) Predicting seasonal influenza transmission using functional regression models with temporal dependence. PLoS ONE 13(4): e0194250. https://doi.org/10.1371/journal.pone.0194250

Artículo Chen R., Tsay R. Functional-Coefficient Autoregressive Models. Published by: Taylor y Francis, Ltd. on behalf of the American Statistical Association. *https://www.jstor.org/stable/2290725.* Journal of the American Statistical Association , Mar., 1993, Vol. 88, No. 421(Mar., 1993), pp. 298-308.

3.5. MODELOS DE REGRESIÓN FUNCIONAL CON ERRORES DEPENDIENTES

Artículo Ateba F., Febrero M., Sagara I.,... (2020) Predicting Malaria Transmission Dynamics in Dangassa, Mali: A Novel Approach Using Functional Generalized Additive Models. International Journal of Environmental Research and Public Health.

Capítulo 4

Simulación

La simulación de los datos para los modelos de regresión funcional (FRM) se encuentran en Cardot y et al. (2003), específicamente, para modelos lineales (LM). En el proceso se generaron $n_B = 500$ réplicas de un proceso aleatorio de Wiener con longitud $M = 101$ definidos en un intervalo entre $[0, 1]$. Para fines de comparación, se pusieron a prueba varios modelos de regresión funcional (FRM) con respuesta escalar donde la viabilidad y eficiencia de cada uno está sujeta a los siguientes escenarios: 1. "la relación lineal entre las variables está controlada por un parámetro integrado δ *(proporción de linealidad)* multiplicado por una componente no lineal", 2. "el cambio de la señal de ruidos (snr) como efecto en la dispersión en la estructura de dependencia de los errores en adición a la variable respuesta existe" y 3. " la creación de una versión iterativa para cada uno de los modelos que contempla la estimación del parámetro ϕ y θ correspondiente a la estructura de covarianza del error ε en las predicciones." Los resultados obtenidos en ambos escenarios serán presentados en tablas simultáneamente y de forma cruzada. La ecuación general de los modelos FLM está dada por

$$Y = E[Y|\mathcal{X}] = \langle \mathcal{X}, \beta \rangle + \varepsilon$$

Ecuación del modelo FLM controlado por el factor δ que regula la proporción de linealidad.

$$Y = E[Y|\mathcal{X}] = \delta\langle \mathcal{X}, \beta \rangle + (1 - \delta)\|\mathcal{X}\| + \varepsilon$$

Si el factor que pondera la linealidad es $\delta = 1$ la componente no lineal $\|\mathcal{X}\|$ se hace cero resultando un modelo lineal, si $\delta < 1$ tiene como resultante un modelo con proporcionalidad lineal menor. El β teórico es un parámetro con representación positivas construido a partir del *modelo (a)* donde con tres componentes principales tiene representación exacta, una alternativa de construcción es el modelo (b) donde se necesita un número mayor de componentes para estar bien representado.

a) $\beta(t) = 2\sin(0{,}5\pi t) + 4\sin(1{,}5\pi t) + 5\sin(2{,}5\pi t), \quad t \in [0, 1]$

b) $\beta(t) = log(15t^2 + 10) + \cos(4\pi t), \qquad t \in [0, 1]$

El término de error ε es una variable aleatoria perturbada por un modelo temporal autoregresivo (AR) con parámetro ϕ de orden $(p = 1, 2)$ en adicción al modelo anterior . Esta simulación de modelos con dependencia de errores está incluidas en código de R elaborado por *Oviedo y Febrero en "https://github.com/moviedo5/fda.usc"*, en una nueva actualización, se incorporó un modelo autoregresivo de medias móviles (ARMA) de parámetros ϕ y θ de orden $(p = 1, q = 1)$ con el objetivo de comparar el efecto dependencia de cada uno de los procesos, la capacidad de los modelos para estimar los parámetros y para predecir valores futuros a un horizonte $yp + h; \quad h = 1, .., 10$.

La perturbación incluida de los procesos temporales son estrictamente estacionarias con variabilidad controlada por efecto de su desviación teórica multiplicada por una señal de ruido (snr) de la forma

$\mathbb{V}ar[\varepsilon] = snr\mathbb{V}ar[\langle \mathcal{X}, \beta \rangle]$. El cálculo de la desviación $sd[\varepsilon]$ para cada uno de los modelos según *Mauricio (2007), Metcalfe y et al. (2009) y Monsalve y et al. (2015)* está dado por

- **Autoregresivo de orden 1** $AR(p = 1)$

  con varianza

$$\sigma_Y^2 = \frac{\sigma_A^2}{1 - \phi_1 \rho_1} = \frac{\sigma_A^2}{1 - \phi_1^2}$$

  la desviación control está dada por

$$ds[\varepsilon] = \sqrt{\frac{snr\mathbb{V}ar[\langle \mathcal{X}, \beta \rangle]}{1 - \phi_1^2}}$$

- **Autoregresivo de orden 2** $AR(p = 2)$

  con varianza

$$\sigma_Y^2 = \frac{\sigma_A^2}{1 - \phi_1 \rho_1 - \phi_2 \rho_2} = \frac{\sigma_A^2}{1 - \phi_1^2 - \phi_2^2}$$

  la desviación control está dada por

$$ds[\varepsilon] = \sqrt{\frac{snr\mathbb{V}ar[\langle \mathcal{X}, \beta \rangle]}{1 - \phi_1^2 - \phi_2^2}}$$

- **Autoregresivo de medias móviles** $ARMA(p = 1, q = 1)$

  con varianza

$$\sigma_Y^2 = \frac{1 - 2\phi_1\theta_1 + \theta_1^2}{1 - \phi_1^2}\sigma_1^2 = \left[1 + \frac{(\phi_1 - \theta_1)^2}{1 - \phi_1^2}\right]\sigma_A^2$$

  la desviación control está dada por

$$ds[\varepsilon] = \sqrt{\left[1 + \frac{(\phi_1 - \theta_1)^2}{1 - \phi_1^2}\right] * snr\mathbb{V}ar[\langle \mathcal{X}, \beta \rangle}$$

Para el conjunto de datos se utilizó representación en base de B-spline (BSP) y de componentes principales (FPC) expresada como ($\Psi = \varphi$) con un mismo número de componentes ($K_x = K_\beta$) entre $5 - 11$ y de $1 - 8$, respectivamente. El número óptimo de componentes es determinado por un método alternativo al de validación cruzada generalizada (GCV) cuando los errores son dependientes. Para estos casios se recomienda usar validación cruzada correlacionada generalizada (GCCV) propuesto por *Carmack (2012)*. Para estimar el parámetro mininizador $\hat{\Theta}$ de la forma siguiente.

$$GCCV(\Theta) = \frac{1}{n}\frac{\sum_{k=1}^{n}(y_k - \hat{\mathfrak{f}}(x_k))^2}{(1 - \frac{tr[2SC - SCS']}{n})^2} \tag{4.1}$$

Donde $\hat{\mathfrak{f}}(x) = \sum_{i=1}^{n} w_i y_i$ es el estimador de la función de suavizado lineal $\mathfrak{f}$, $w_1, ..., w_n$ son los pesos de la función de x, los grados de libertad residuales cuando hay correlación están definidos por $n - tr[2SC - SCV']$.

Este método es usado por *Oviedo y et al. (2018)* para estimar los parámetros $K_x, K_\beta, \mathbf{b}, \phi$ de la forma siguiente.

$$GCCV(K_x, K_\beta, \mathbf{b}, \phi) = \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i,\mathbf{b}})^2}{(1 - \frac{tr[G]}{n})^2} \tag{4.2}$$

donde $G = 2H\sum(\phi) - H\sum(\phi)H$; H es la matriz sombrero y la traza de G estima los grados de libertad del modelo.

A continuación se presentan los resultados correspondientes al modelo donde β fue construido de la forma (a) y el error es generado por un proceso AR(p = 1,2) y ARMA(1,1).

	PC						**BSP**					
	$\delta = 1$			$\delta = 0.70$			$\delta = 1$			$\delta = 0.70$		
snr	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$
0.05	2.2	2.1	2.1	2.2	2.1	2.1	4.1	4.2	3.9	4.1	4.0	3.9
0.10	2.1	2.1	2.0	2.1	2.1	2.1	3.9	3.8	3.8	3.9	3.9	3.8

Tabla 4.1: Promedio de Número de bases seleccionados por el criterio de GCCV en $n_B = 402$ para una proporción de linealidad es $\delta = 1$ y $n_B = 389$ cuando la proporción del linealidad es $\delta = 0{,}70$ del Modelo (a). ARMA(1,1). *($n_B < 500$ por generación de proceso no estacionario).*

En la tabla 4.1, se muestra el número promedio de componentes y de bases obtenidos mediante el criterio de GCCV. Con un número de réplicas razonables para la representación de las curvas por componentes principales (PC) se obtiene un número promedio de K_x que se aproximan conformemente al valor exacto ($PC = 3$) de componentes vistos en la metodología, por otro lado, para bases de B-spline (BSP) el K_x para representar el conjunto de curvas tiende al valor mínimo de bases especificado, esto significa que con un número mayor de réplicas ($n_B > 500$) se tendrían valores promedios más satisfactorios. Dentro de eso, se destaca que en los modelos donde la proporción de linealidad (δ) es baja el número promedio de bases de PC y BSP es menor, al menos para los modelos donde la dependencia temporal de los errores incluida viene dada por un AR(1) y AR(2), para un ARMA(1,1) el número promedio es relativamente igual. Los resultados de procesos con errores dependientes correspondientes a un AR(1) y AR(2) están en las tablas A.1 y A.2 del apéndice.

$$\mathbb{E}\left[\|\beta - \hat{\beta}\|^2\right]$$

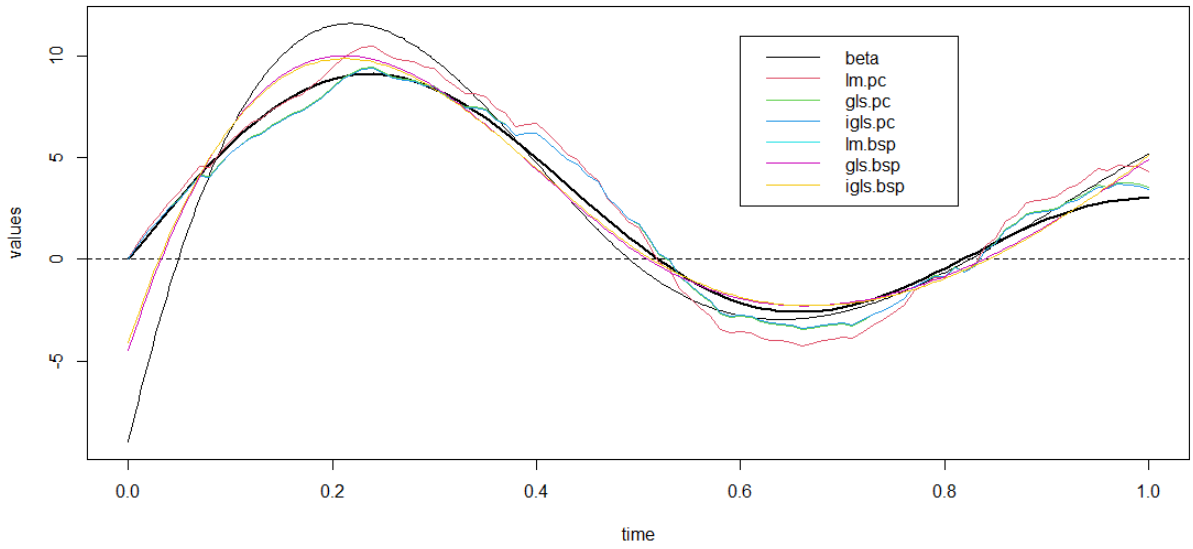


Figura 4.1: Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. ARMA(1,1). $\delta = 1$.

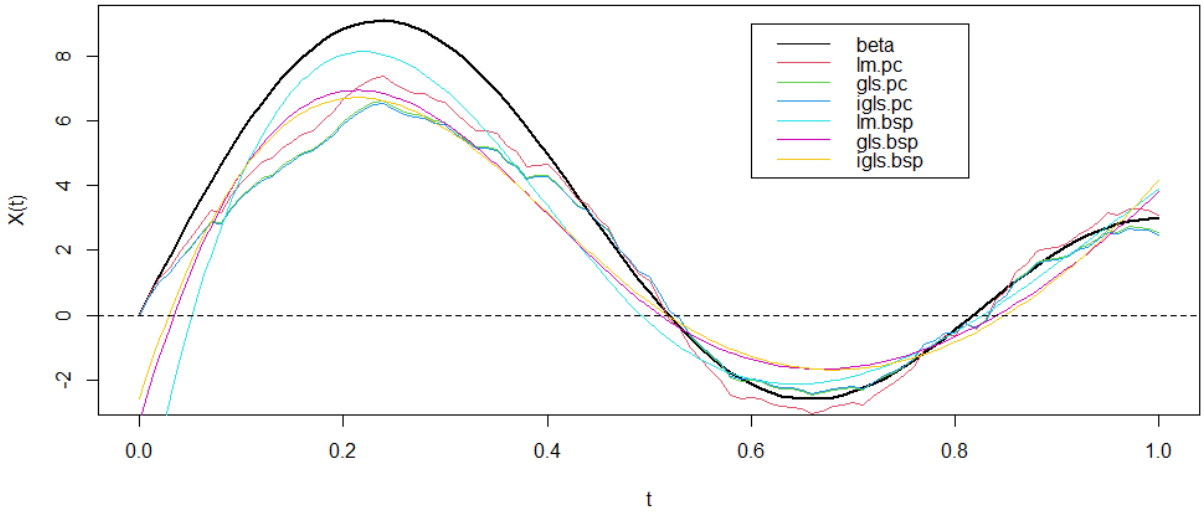


Figura 4.2: Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. ARMA(1,1). $\delta = 0{,}7$.

En las figuras 4.1 y 4.2 se puede apreciar el efecto de la proporción de linealidad (δ) de la relación entre las variables para estimar el β *(línea negra)* . Cuando $\delta = 1$ los $\hat{\beta}$ tejen toda la curva, sin embargo, una leve reducción del $\delta = 0{,}7$ hace que las estimaciones queden por debajo en la cresta de la curva del β y mejorando su aproximación al final. Los resultados para un modelo AR(1) pueden consultarse en las figuras A.16 y A.17 del apéndice.

		snr = 0.05						**snr = 0.10**					
		$\delta = 1$			$\delta = 0.15$			$\delta = 1$			$\delta = 0.15$		
Base	**Modelo**	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$	$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$	$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$	$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$
PC	LM	0.442	0.440	0.457	0.939	0.937	0.991	4.070	4.066	**4.075**	4.128	4.119	**4.126**
PC	GLS-ARMA(1,1)	**0.403**	**0.403**	**0.438**	**0.534**	**0.553**	**0.883**	4.068	4.067	**4.075**	**4.126**	4.119	4.128
PC	iGLS-ARMA(1,1)	**0.403**	**0.403**	**0.438**	**0.534**	**0.553**	**0.883**	4.068	4.067	**4.075**	**4.126**	4.119	4.128
BSP	LM	0.482	0.494	0.513	1.145	1.179	1.294	**4.065**	**4.061**	4.082	**4.126**	4.108	4.133
BSP	GLS-ARMA(1,1)	0.422	0.434	0.495	0.684	0.740	1.138	4.068	**4.061**	4.081	4.127	**4.105**	4.132
BSP	iGLS-ARMA(1,1)	0.422	0.434	0.495	0.684	0.740	1.138	4.067	**4.061**	4.081	4.127	**4.105**	4.132

Tabla 4.2: Error cuadrático medio del parámetro β. Modelo (a), $n_B = 402$ para una proporción de linealidad es $\theta = 1$ y $n_B = 389$ cuando la proporción del linealidad es $\theta = 0{,}70$. ARMA(1,1).

En la tabla 4.2 se muestran varios escenarios de comparación del error cuadrático medio (MSE) del parámetro β y su estimación. El conjuntos curvas representadas en base PC con un nivel de ruido $snr = 0{,}05$ presentan resultados más favorables que si son representadas por BSP con $snr = 0{,}05$ en todos los escenarios haciendo uso de los modelo GLS e iGLS, con $snr = 0{,}10$ la efectividad para estimar el parámetro disminuye y no se define un patrón; el efecto de disminuir a δ afecta en los resultados de estimación del parámetro como era de esperarse. Se puede observar que un aumento en los parámetros ϕ y θ como efecto de dependencia temporal de los errores de los modelos AR Y ARMA incluidos provoca un aumento *(empeoran)* en los resultados del MSE para todos los modelos estimados, sin embargo, los modelos con capacidad para detectar dependencia GLS e iGLS funcionan de con mayor eficiencia que los modelos lineales LM, aún cuando la dependencia es casi cero. Ver más resultados en las tablas A.3 y A.4 del apéndice.

$$\mathbb{E}\left[(\phi - \hat{\phi})^2\right]$$

$$\mathbb{E}\left[(\theta - \hat{\theta})^2\right]$$

		PC						**BSP**					
		$\delta = 1$			$\delta = 0.70$			$\delta = 1$			$\delta = 0.70$		
snr	**Modelo**	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}7$
0.05	GLS-ARMA(1,1)	0.042	0.013	**0.005**	0.074	0.020	0.006	0.042	0.014	**0.005**	0.074	0.019	0.006
0.05	iGLS-AR(1,1)	**0.039**	**0.012**	**0.005**	**0.065**	0.018	0.006	**0.035**	**0.012**	**0.005**	**0.056**	0.016	0.006
0.10	GLS-AR(1,1)	0.055	0.014	**0.005**	0.074	0.015	**0.005**	0.057	0.014	**0.005**	0.075	0.015	**0.005**
0.10	iGLS-ARMA(1,1)	0.052	0.013	**0.005**	0.067	**0.014**	**0.005**	0.050	**0.012**	**0.005**	0.062	**0.013**	**0.005**
snr	**Modelo**	$\theta_1 = 0{,}3$	$\theta_1 = 0{,}5$	$\theta_1 = 0{,}7$	$\theta_1 = 0{,}3$	$\theta_1 = 0{,}5$	$\theta_1 = 0{,}7$	$\theta_1 = 0{,}3$	$\theta_1 = 0{,}5$	$\theta_1 = 0{,}7$	$\theta_1 = 0{,}3$	$\theta_1 = 0{,}5$	$\theta_1 = 0{,}7$
0.05	GLS-AR(1,1)	0.124	0.266	0.498	0.193	0.355	0.631	0.126	0.266	0.495	0.191	0.351	0.629
0.05	iGLS-AR(1,1)	**0.111**	**0.252**	**0.482**	**0.164**	0.332	0.606	0.106	0.248	0.473	0.158	0.319	0.594
0.10	GLS-AR(1,1)	0.119	0.259	0.491	0.166	0.309	0.564	0.122	0.260	0.489	0.168	0.309	0.562
0.10	iGLS-AR(1,1)	0.109	0.246	0.475	0.145	**0.290**	**0.544**	**0.100**	**0.243**	**0.468**	**0.138**	**0.283**	**0.535**

Tabla 4.3: Error cuadrático medio del parámetro ϕ y θ. Modelo (a), $n_B = 402$ réplicas para una proporción de linealidad es $\delta = 1$ y $n_B = 389$ cuando la proporción del linealidad es $\delta = 0{,}70$. ARMA(1,1).

En las tabla 4.3 se muestran los resultados del error cuadrático medio (MSE) para los parámetros ϕ y θ con su estimación, obtenidas de los modelos GLS e iGLS, los cuales son propicios para detectar esa dependencia, modelar y predecir bajo este escenarios. En general, los resultados obtenidos presentan mejora *(disminuye)* al aumentar *(efecto de autocorrelación)* el valor de ϕ de los modelos AR(p = 1,2) y

ARMA(1,1) incluidos. Caso contrario para el parámetro θ del modelos ARMA(1,1) que al aumentar el valor del efecto aleatorio los resultados del MSE también aumentan *(empeoran)*. El MSE del parámetro ϕ aumenta cuando el *snr* aumenta y $\phi = 0{,}3$ (*es pequeño)*, sin embargo, para valores de $\phi = 0{,}5$ y $\phi = 0{,}7$ las diferencias del MSE son mínimas, por otro lado, el MSE del efecto aleatorio del parámetro θ disminuye (mejora) cuando *snr* aumenta de 0.05 a 0.10 entre modelos del mismo tipo, tanto en el escenario de proporcionalidad lineal $\delta = 1$ como cuando proporcionalidad disminuye $\delta = 0{,}70$. Las tablas A.5 y A.6 del apéndice muestran que cuando los errores son generados por proceso un AR(1) y AR(2) no hay patrón de variación notable de los resultados del MSE para los parámetros ϕ al aumentar la señal de ruido *snr* de 0.05 a 0.10 y la proporción de linealidad $\delta = 1$, bajo ese mismo escenario, pero con $\delta = 0{,}15$ el MSE obtiene mejores resultados al aumentar el *snr* 0.05 a 0.10 y los valores de ϕ presentan autocorrelación alta.

A continuación se presentan los resultados del error cuadrático medio de las predicciones (MSPE) para cada uno de los modelos vistos a horizonte 1, 5 y 10, en donde se comparan los resultados bajo escenarios específicos como: una variación en proporción de linealidad δ y las señal de ruido (snr), la base en la que se representa el conjunto de datos y la dependencia temporal o no de los errores. Entre los modelos a comparar se encuentra el modelo lineal funcional clásico FLM que contempla independencia en los errores *(fregre.lm de R)*, modelos lineales donde hay dependencia de error como el de mínimos cuadrados generalizado funcional (GLS) *(fregre.gls de R)* y su versión iterativa (iGLS) *(fregre.igls de R)*; se integró los modelos de regresión aditiva espectral generalizada funcional (GSAM) *(fregre.gsam de R)* y los modelos aditivos de kerner generalizados funcional (GKAM) *(fregre.gkam de R)* con capacidad para predecir relaciones no lineales entre covariables, cada uno con una versión iterativa que agrega a las estimaciones las predicciones de la dependencia temporal de los errores manualmente.

$$MSPE = \frac{1}{n_B}\sum_{b=1}^{n_B}\left(y^b_{n+h} - \hat{y}^b_{n+h}\right)^2$$

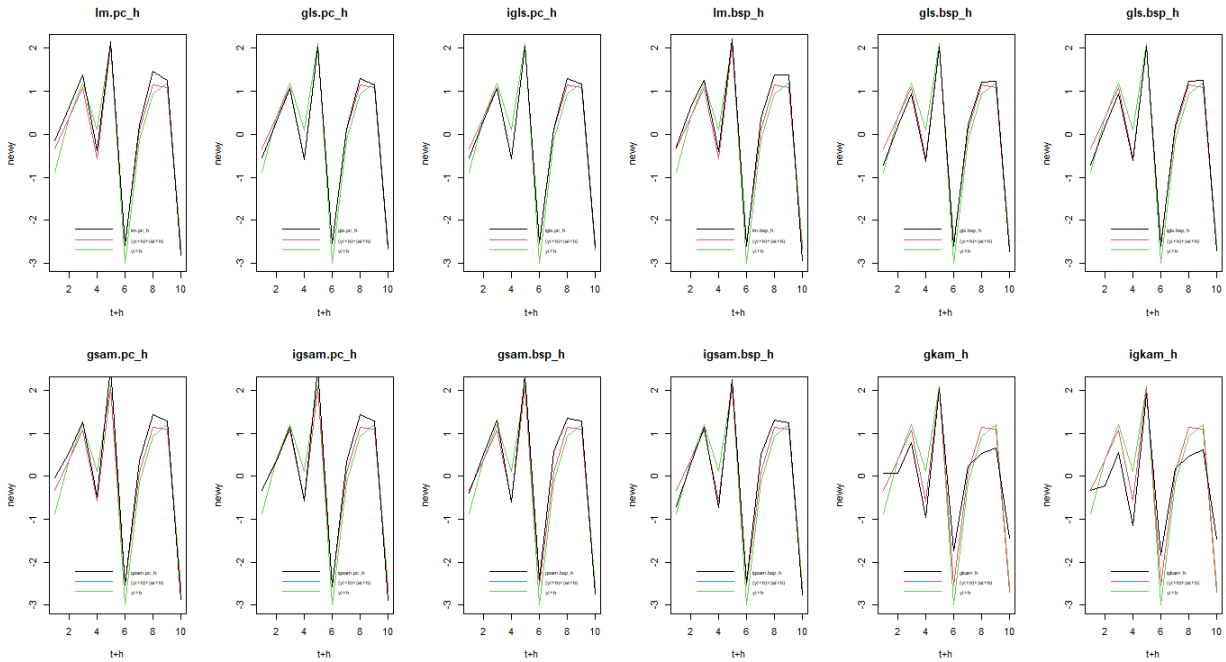


Figura 4.3: Predicciones de los modelos a horizonte $h = 10$. ARMA(1,1). $\delta = 1$

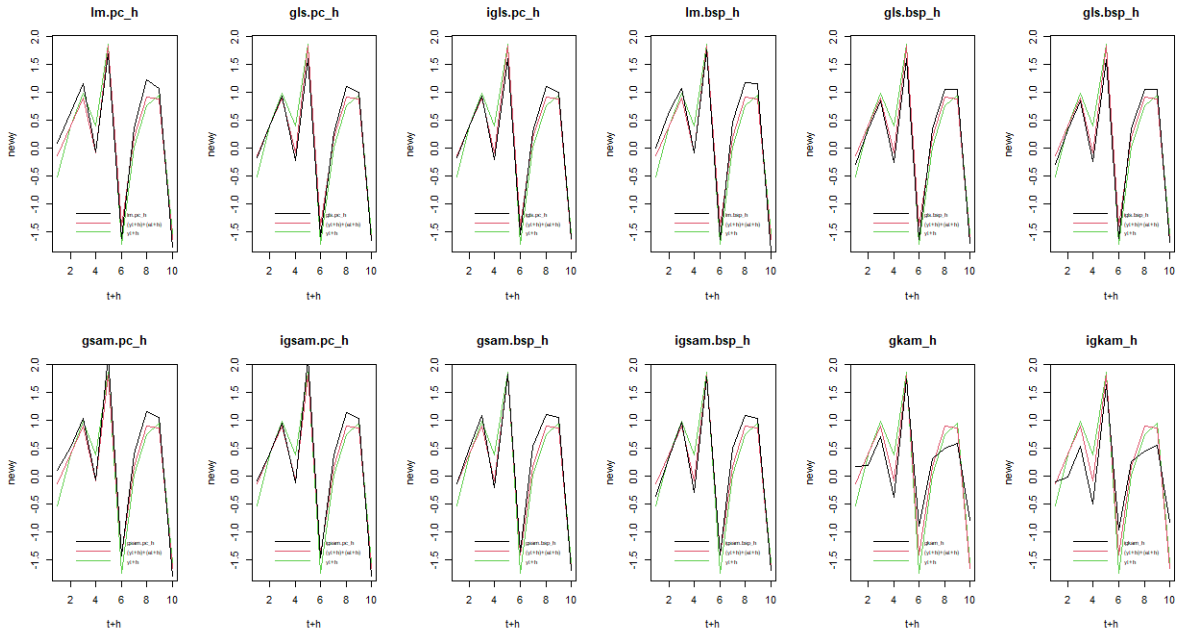


Figura 4.4: Predicciones de los modelos a horizonte $h = 10$. ARMA(1,1). $\delta = 0{,}7$.

En las figuras 4.3 y 4.4 se hace una comparación gráfica de las predicciones $(\hat{y} + h)$ *(negro)* de los modelos a horizonte $h = 1, ..., 10$ con las observaciones $(y+h)$ *(verde)* sin el efecto de la dependencia de error generado por un proceso temporal estacionario y con el efecto $(y+e)+h)$ *(rojo)*. Estos resultados se pueden interpretar de varias manera, comprobar por la distancia mínima cual de los modelos tuvo mejores resultados, además, comprobar si el contexto y las especificaciones metodológicas de aplicación para cada uno de los modelos tienen concordancia con estos resultados. Los modelos GLS, IGLS,GSAM E iGSAM con base de componentes principales tienen mejores resultados que el resto de modelo cuando los modelos son enteramente lineales. La reducción de proporción de linealidad permitida de $\delta = 0{,}7$ para conservar la estacionalidad no son suficiente para lograr mejoras sustanciales de los modelos no paramétricos. Otros resultados gráficos pueden verse en las figuras A.22 y A.23.

			$ARMA(1,1)$, $\delta = 1$								
			$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$			$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$			$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$		
snr	**Modelo**	**Bases**	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10
0.05	LM	PC	0.086	0.078	0.085	0.116	0.116	0.132	0.290	0.268	0.269
0.05	GLS-ARMA(1,1)	PC	**0.080**	0.078	**0.084**	0.098	**0.115**	0.132	**0.148**	**0.253**	**0.266**
0.05	iGLS-ARMA(1,1)	PC	**0.081**	0.078	**0.084**	0.097	**0.115**	0.132	**0.148**	**0.253**	**0.267**
0.05	GSAM	PC	0.089	0.079	0.086	0.118	0.117	0.133	0.294	0.266	0.274
0.05	iGSAM-ARMA(1,1)	PC	0.093	**0.075**	0.095	0.103	0.164	0.153	0.158	0.278	0.296
0.05	LM	BSP	0.087	0.079	0.087	0.117	0.117	0.131	0.295	0.271	0.279
0.05	GLS-ARMA(1,1)	BSP	**0.081**	0.078	0.086	0.097	0.117	**0.130**	0.150	**0.251**	0.270
0.05	iGLS-ARMA(1,1)	BSP	0.082	0.079	0.087	**0.096**	0.117	**0.130**	0.150	**0.251**	0.271
0.05	GSAM	BSP	0.091	0.080	0.089	0.118	0.121	**0.126**	0.302	0.269	0.278
0.05	iGSAM-ARMA(1,1)	BSP	0.093	0.079	**0.084**	**0.095**	0.160	0.149	0.162	0.293	0.315
0.05	GKAM		0.472	0.361	0.426	0.437	0.404	0.557	0.506	0.475	0.677
0.05	iGKAM-ARMA(1,1)		0.479	0.376	0.428	0.434	0.417	0.576	0.500	0.465	0.745
0.10	LM	PC	**0.158**	0.157	0.171	0.261	0.246	0.269	0.499	**0.518**	0.532
0.10	GLS-ARMA(1,1)	PC	**0.157**	0.159	0.172	**0.215**	**0.242**	0.267	0.294	0.530	0.527
0.10	iGLS-ARMA(1,1)	PC	**0.157**	0.159	0.172	**0.215**	**0.242**	0.267	0.294	0.529	0.527
0.10	GSAM	PC	0.159	**0.155**	0.176	0.268	0.254	0.271	0.513	0.530	0.538
0.10	iGSAM-ARMA(1,1)	PC	0.204	0.189	0.171	0.233	**0.241**	**0.224**	0.288	0.586	0.662
0.10	LM	BSP	0.164	0.164	0.168	0.273	0.251	0.269	0.506	0.524	0.525
0.10	GLS-ARMA(1,1)	BSP	0.162	0.162	0.168	0.225	0.245	0.267	0.301	0.529	**0.520**
0.10	iGLS-ARMA(1,1)	BSP	0.162	0.163	0.168	0.225	0.245	0.268	0.301	0.528	**0.519**
0.10	GSAM	BSP	0.173	0.163	0.172	0.269	0.261	0.270	0.505	0.541	0.545
0.10	iGSAM-ARMA(1,1)	BSP	0.208	0.183	**0.157**	0.243	0.251	0.226	**0.277**	0.563	0.655
0.10	GKAM		0.716	0.483	0.367	0.522	0.406	0.436	0.887	0.793	0.692
0.10	iGKAM-ARMA(1,1)		0.719	0.508	0.370	0.508	0.421	0.446	0.784	0.792	0.722

Tabla 4.4: Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 402$. ARMA(1,1). $\delta = 1$.

			$ARMA(1,1)$, $\delta = 0{,}70$								
			$\phi_1 = 0{,}3$ $\theta_1 = 0{,}3$			$\phi_1 = 0{,}5$ $\theta_1 = 0{,}5$			$\phi_1 = 0{,}7$ $\theta_1 = 0{,}7$		
snr	**Modelo**	**Bases**	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10
0.05	LM	PC	0.059	0.046	0.054	0.075	0.078	0.071	0.147	0.130	0.153
0.05	GLS-ARMA(1,1)	PC	0.056	0.046	0.055	0.063	0.077	0.070	**0.087**	0.130	0.151
0.05	iGLS-ARMA(1,1)	PC	0.059	0.046	0.055	0.063	0.077	0.070	**0.087**	0.130	0.152
0.05	GSAM	PC	**0.048**	**0.041**	**0.047**	0.067	**0.070**	**0.066**	0.138	**0.127**	**0.145**
0.05	iGSAM-ARMA(1,1)	PC	**0.044**	**0.037**	**0.047**	**0.052**	0.082	0.075	**0.089**	0.149	0.152
0.05	LM	BSP	0.060	0.046	0.053	0.074	0.079	0.071	0.144	0.132	0.158
0.05	GLS-ARMA(1,1)	BSP	0.058	0.047	0.053	0.062	0.078	0.069	**0.086**	0.132	0.152
0.05	iGLS-ARMA(1,1)	BSP	0.058	0.047	0.053	0.062	0.078	0.068	**0.086**	0.131	0.153
0.05	GSAM	BSP	**0.049**	**0.041**	**0.046**	0.066	**0.070**	**0.064**	0.139	**0.126**	0.150
0.05	iGSAM-ARMA(1,1)	BSP	**0.046**	**0.040**	**0.042**	**0.048**	0.082	0.074	**0.086**	0.145	0.156
0.05	GKAM		0.233	0.177	0.207	0.215	0.199	0.276	0.247	0.238	0.339
0.05	iGKAM-ARMA(1,1)		0.239	0.190	0.207	0.215	0.205	0.285	0.243	0.232	0.368
0.10	LM	PC	0.095	0.085	0.096	0.127	0.128	0.127	0.276	0.273	0.267
0.10	GLS-ARMA(1,1)	PC	**0.094**	**0.085**	0.096	**0.105**	0.128	0.127	0.157	0.262	0.266
0.10	iGLS-ARMA(1,1)	PC	**0.094**	**0.084**	0.096	**0.105**	0.128	0.127	0.157	0.261	0.266
0.10	GSAM	PC	0.111	0.093	**0.075**	0.126	**0.105**	**0.086**	0.213	**0.252**	**0.246**
0.10	iGSAM-ARMA(1,1)	PC	0.105	0.098	**0.083**	0.119	0.124	**0.114**	**0.144**	0.291	0.334
0.10	LM	BSP	0.096	**0.085**	0.096	0.128	0.131	0.131	0.281	0.276	0.266
0.10	GLS-ARMA(1,1)	BSP	**0.094**	**0.084**	0.096	**0.105**	0.130	0.131	0.160	0.261	0.265
0.10	iGLS-ARMA(1,1)	BSP	**0.093**	**0.085**	0.095	**0.105**	0.130	0.131	0.160	0.261	0.265
0.10	GSAM	BSP	0.113	0.092	**0.074**	0.134	**0.108**	**0.088**	0.208	**0.257**	0.253
0.10	iGSAM-ARMA(1,1)	BSP	0.110	0.093	**0.081**	0.119	0.129	0.118	**0.143**	0.282	0.338
0.10	iGKAM		0.353	0.238	0.180	0.261	0.205	0.218	0.444	0.387	0.348
0.10	iGKAM-ARMA(1,1)		0.355	0.252	0.181	0.254	0.215	0.223	0.388	0.389	0.363

Tabla 4.5: Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 402$. ARMA(1,1). $\delta = 0{,}70$.

En la tabla 4.4 se muestran los resultados del MSPE donde los errores fueron generados por un proceso ARMA(1,1) y la relación entre covariables es lineal $\delta = 1$, se observa que los resultados de los modelos GLS e iGLS superan el resto de modelos en todos los escenarios cuando la señal de ruido $snr = 0{,}05$ y el valor predicho es a horizonte $t + 1$, sin embargo, al aumentarla a $snr = 0{,}10$ los modelo lineales tienen resultados del MSPE relativamente iguales para valores de ϕ y θ pequeños, cuando aumenta el valor de estos parámetros las diferencias son más marcadas a favor de los modelos GLS e iGLS de igual manera para horizonte $t + 1$, a horizonte $t + 5$ y $t + 10$ los modelos tienen resultados similares. En la tabla 4.5 se reducen las especificaciones de la proporción de linealidad a $\delta = 0{,}70$ y los modelos GSAM e iGSAM superan al resto cuando $snr = 5$, sin embargo, para $snr = 10$ los modelos con mejores resultados son los GLS e iGLS a horizonte $t + 1$, a horizontes más largos $t + 5$, .,10 modelos GSAM e iGSAM vuelven y mejoran la precision con resultados que superan al resto de modelos. La reducción de la proporción de linealidad permitida hasta $\delta = 0{,}70$ para tener un número razonable de muestras de errores generadas por un procesos ARMA(1,1) estacionarios no permite apreciar la bondad del modelo GKAM e iGKAM, sin embargo, para los procesos autoregresivos AR(p=1,2) donde el $\delta = 0{,}15$ es cercano a cero los modelos no paramétricos tuvieron resultados más favorables a todos los horizontes de predicción, en caso contrario, cuando $\delta = 1$ los GLS e iGLS tienen mejores resultados y a todos los horizontes de predicción. Para ver estos resultados ir a las tablas A.7, A.8, A.9 y A.10.

Bibliografía

Artículo Cardot H., Ferraty F., Sarda P. (2003) Spline Estimators for the Functional Linear Model. https://www.researchgate.net/publication/228727533. Statistica Sinica 13(2003), pp. 571-591.

Revista Febrero-Bande, M. y Oviedo de la Fuente, M. (2012). Computación estadística en análisis de datos funcionales: El paquete R fda.usc. Revista de Software Estadístico, 51(4):1-28. http://www.jstatsoft.org/v51/i04/.

Libro Mauricio J.A (2007) Introducción al Análisis de Series de Tiempo. Universidad de Complutense. Madrid, España.

Libro Metcalfe A, V., Cowpertwait P.S.P. (2009) Introductory Time Series with R. https://www.researchgate.net/publicat
.

Artículo Oviedo M., Febrero M., Muñoz M.P., Dominguez A. (2018) Predicting seasonal influenza transmission using functional regression models with temporal dependence. PLoS ONE 13(4): e0194250. https://doi.org/ 10.1371/journal.pone.0194250

Artículo Carmack P.S., Spence J.S., Schucany W.R (2012). Generalised correlated cross-validation, Journal of Nonparametric Statistics, Vol.24, No.2, pp. 269-282, DOI: https://doi.org/10.1080/10485252.2012.655733

Libro Cryer J.D., Chan K-S. (2008) Time Series Analysis with Applications in R. Springer Science+Business Media, LLC. New York, Estados Unidos.

Capítulo 5

Aplicación de FRM para predecir la incidencia acumulada de contagio del Covid-19

En esta parte del documento se utilizar los modelos de regresión funcional (FRM) para aplicarlos datos reales y para predecir la incidencia de contagio del Covid-19 en las comunidades autónomas (CCAA) de España. El origen de este mal que azota al mundo inició en la provincia Hubei-China a finales de diciembre del año 2019, luego de identificar unos casos de neumonía agrupados de forma inusual en una comunidad de la provincia. Su rápida expansión y forma de propagación provocó que se denominara como "*brote de virus*" SARS-CoV-2 y su enfermedad Covid-19 por al *Organización Mundial de la Salud (OMS*) el 7 de enero del 2020[1]. El estado de emergencia sanitaria se reconoció como pandemia global el 11 de marzo de 2020; mediante el "*Real Decreto 463/2020*"[2] del 14 de marzo del 2020 se declara el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el COVID-19 en España de cara a preservar la salud y la economía del país.

La "nueva normalidad" a la que se adaptan las CCAA de España durante todo este tiempo de pandemia depende del seguimiento de registro de nuevos casos y la velocidad con la que propaga el virus en tiempo/espacio dentro de las mismas. Con esta información los investigadores y organismos de salud construyen indicadores de alerta y elaboran estrategias locales para reducir en número de casos, evitar las oleadas, mantener la salud y la dinámica de la economía. Como fuentes de información fidedignas el estado se apoya de Red Nacional de Vigilancia Epidemiológica (RENAVE) con el Sistema de Vigilancia de España (SiViES). Este sistema está implementado en una plataforma informática virtual gestionada por el Centro Nacional de Epidemiología (CNE) donde se cargan los datos e informaciones del Covid-19 reportados de toda España. De las informaciones que se pueden consultar en esta plataforma está la tasa de incidencia a 7 y 14 días construidas con criterios y parámetros claramente especificados por las fechas sintomáticas o no, y la confirmación de positividad al momento de hacerse alguna de las pruebas diagnosticas.

El archivo de datos utilizado para el análisis se llama *casos_tecnica_ccaa.csv* que corresponde al "Número de casos por técnica diagnóstica y CCAA (de residencia)"[3] que diariamente se actualiza en el SiViES en horarios de 15:00h a 16:00h. La fecha de corte de la data corresponde al período desde 01/01/2020 hasta 10/06/2021, esto equivale a un total 527 observaciones diarias para cada una de las 19 CCAA.

[1] https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).

[2] https://boe.es/eli/es/rd/2020/03/14/463.

[3] https://cnecovid.isciii.es/covid19/#documentaci %C3 %B3n-y-datos.

El registro de contagios diarios *(variable temporal)* depende de la información suministrada por la persona "afectada", en ese sentido, la fecha con la que se registra el caso de contagio está determinado por los siguientes criterios:

- Si al momento de aplicar una prueba diagnostica se registra el caso el día que se sintieron los síntomas.
- De no estar seguro el caso se registra menos 6 días después de la persona ser diagnosticado positivo hasta la fecha del 10 de mayo 2020 y 3 días después de ser diagnosticado a partir del 11 de mayo 2020 en adelante.
- Para las personas asintomáticas la fecha registro coincide con la fecha de diagnostico positivo.

Para construir las tasas se utilizó el registro de padrón municipal del Instituto Nacional de Estadística (INE) la tabla 3.1 "*Población por comunidades y ciudades autónomas y sexo.*"[4]

EL código de predicción utilizado fue proporcionado por *Manuel Oviedo*, se realizó una adaptación de la versión código original de *Manuel Oviedo y Manuel Febrero* para predecir la tasa de crecimiento de casos de confirmados y de muerte mundiales por Covid-19 *(http://modestya.usc.es:3838/covid19prediction/)* utilizando modelos FLM, FKAM y FSAM *(aún no publicado)*, como propuesta se incorporan los modelos GLS, iGLS dado que los datos de contagio del Covid-19 capturados por las entidades de salud no son oportunos al momento de actualizar los cortes diarios, ademas, para fines de comparación se incluyó una versión iterativa que incluye el efecto de dependencia de los errores en las predicciones de los modelos FLM, FKAM y FSAM.

5.1. Metodología

La incidencia acumulada "casos de contagios de Covid-19" a 15 días de registro para cada una de las CCAA de España se tomó a partir de 50 días después de la fecha de inició de registro el 1/01/2020 hasta la fecha de corte 10/06/2020. Esto se debe a los registros escasos de casos que presenta la data en la mayoría de las CCAA en esos primeros días. Con la incidencia acumulada se plantea evaluar la dinámica de aparición de nuevos casos de contagio acumulados en un tiempo para una determinada población en riesgo.

5.1.1. Definición de las variables

- Población total de las CCAA de España $Pob : x_{i=1}^{(m)}$; donde $m \in \{1, ..., 19\}$ población total registrada al 2020.
- Casos registrados acumulados al día $t : x_t^{(j)}$; donde $j \in \{1, ..., 15\}$ casos confirmados.
- Casos nuevos $t : x_t^{(j)} - x_{t-1}^{(j)}$
- Incidencia acumulada casos registrados $H_k : r_k^{(j)}(t) = \frac{x_{t+k}^{(j)} - x_t^{(j)}}{x_{i=1}^{(m)} - x_{t+k}^{(j)}} = \frac{x_{t+50+k}^{(j)} - x_{t+50}^{(j)}}{x_{i=1}^{(m)} - x_{t+50+k}^{(j)}}$ para $t = ..., t_0 - 1$ y número de saltos $k = 1, ..., 5$.

El objetivo principal es obtener los modelos óptimos para predecir la incidencia acumulada de contagio a un horizonte determinado como dinámica de riesgo de contagio *(nuevos casos)* que experimentan las

[4] https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=523.

poblaciones de las CCAA al cabo de 15 días. Para la incidencia de contagio H_1 se puede expresar de la siguiente manera.

$$R_1(0) = \left\{ r_1^j(-14), ..., r_1^j(0) \right\} \tag{5.1}$$

Los ajustes de los modelos de regresión funcional general: $Y = \alpha + \langle \mathcal{X}, \beta \rangle + \varepsilon$ expresados como $r_k^{(j)}(0) = m(R_1(0)) + \varepsilon$, donde m es la función desconocida de diferenciación.

- FLM: usa función lineal: $m(R_1(0)) = \int R_1(t)\beta(t)dt$
- GlS: m usa función lineal y detecta dependencia en los errores (autocorrelación).
- iGlS: m usa función lineal, con una version iterativa para detecta dependencia en los errores (autocorrelación).
- GSAM: m usa combinación aditiva de funciones suaves de su representación en base.
- GKAM = m usa modelo no paramétrico tipo kernel y ancho de banda.

El esquema de predicción puede replicarse para nuevos datos de incidencia de casos de contagio de mañana u otro horizonte de predicción.

5.2. Explotación del conjunto de datos

Durante todo este tiempo de pandemia España ha experimentado 3 oleadas importantes antes de disponer una vacuna y un protocolo de distribución. La primera a mediados de marzo y principio de abril 2020, para épocas de verano las medidas se relajaron un poco. Una segunda oleada para los meses de octubre y noviembre del mismo año. Antes de la fecha de corte de la data, una tercera oleada de contagio, luego de los fiestas de navidad a principios de enero del 2021 y extendiéndose hasta mediados de febrero.

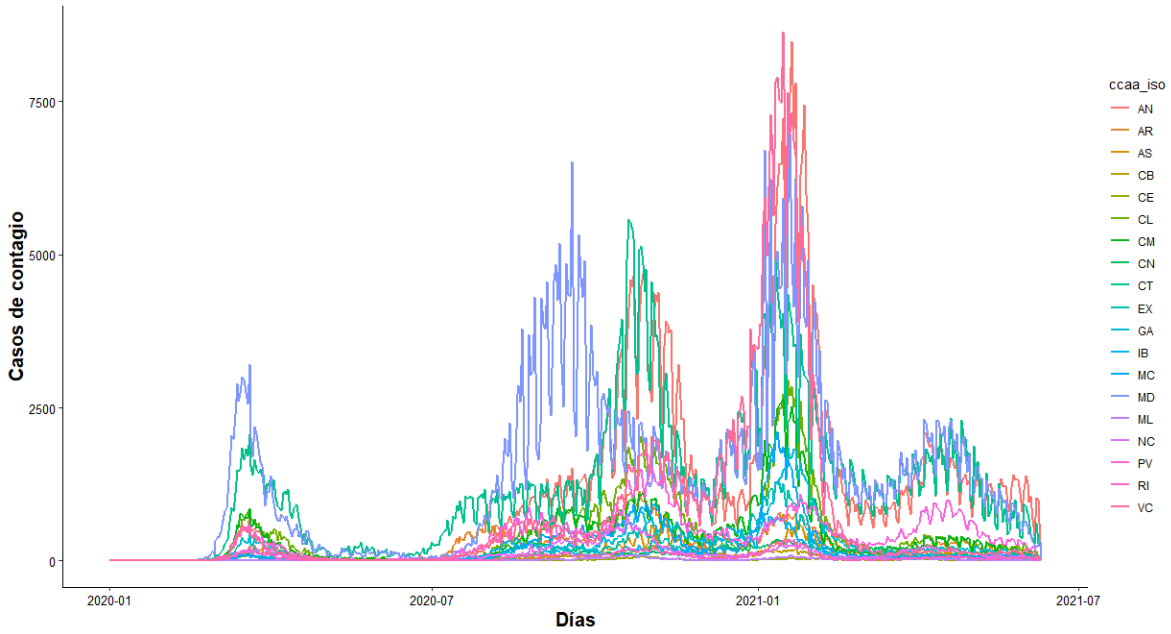


Figura 5.1: Curvas epidémicas de las CCAA de España. Período 01/01/2020 a 10/06/2021. Construido en base a los datos del número de casos de contagio Covid-19 de la plataforma del CNE.

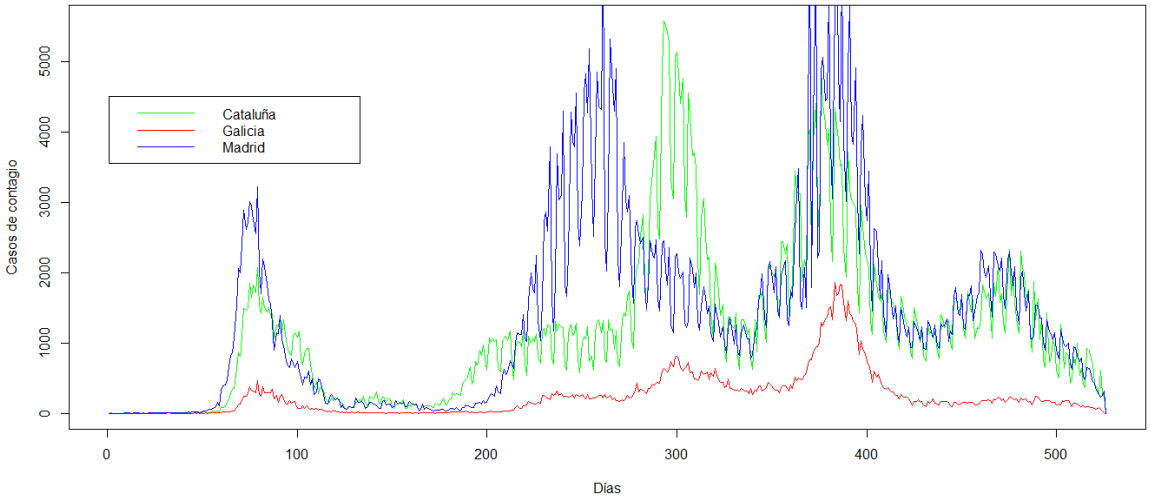


Figura 5.2: Comparativo de Curvas epidémicas de las CCAA de Cataluña, Galicia y Madrid. Período 01/01/2020 a 10/06/2021. Construido en base a los datos del número de casos de contagio Covid-19 de la plataforma del CNE.

Las curvas de la figura 5.1 muestran en su mayoría el comportamiento similar a lo largo del período donde las oleadas afectan a todas las comunidades en tiempos específicos, aunque no en la misma magnitud. Las CCAA con mayor número de habitantes, densamente pobladas[5] y con atractivos turísticos[6] son las más afectados. Un caso particular es el de la comunidad de Madrid *(ver figura 5.2)*, por ser uno de los principales atractivo turístico para resto de países del mundo y con el aeropuerto con mayor número de entradas de España se observa que el aumento en los reportes de casos de contagio comienzan por esta comunidad al menos en las dos primeras oleadas, en ese mismo orden las medidas restrictivas para controlar la situación comienzan a implementarse aquí. Al 10/06/2021 se registra un total de $3,729,468$ casos de contagios, lo que representa aproximadamente un $7{,}86\,\%$ de una población

[5]https://datosmacro.expansion.com/demografia/poblacion/espana-comunidades-autonomas
[6]https://es.wikipedia.org/wiki/Anexo:Aeropuertos_de_Espa%C3%B1a

de 47,450,795 personas registradas en el padrón.

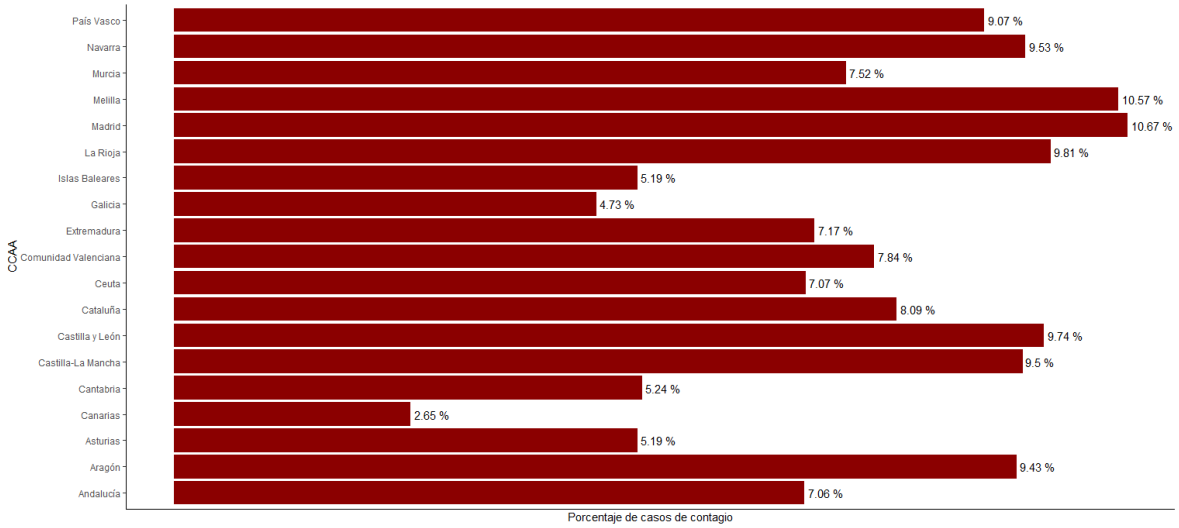


Figura 5.3: Porcentaje de casos confirmados en las CCAA de España. Período 01/01/2020 a 10/06/2021.

Anteriormente se había dicho que las CCAA con mayor número de habitantes, densamente pobladas y con entradas internacionales de turistas son los más afectados. El 7,86 % de afectados de la población general, es excedido por las CCAA que cumplen con al menos dos de estas características, tal es el caso de Madrid, Melilla, Castilla y León, Castilla-La mancha, entre otros. *Ver figura 5.3.*

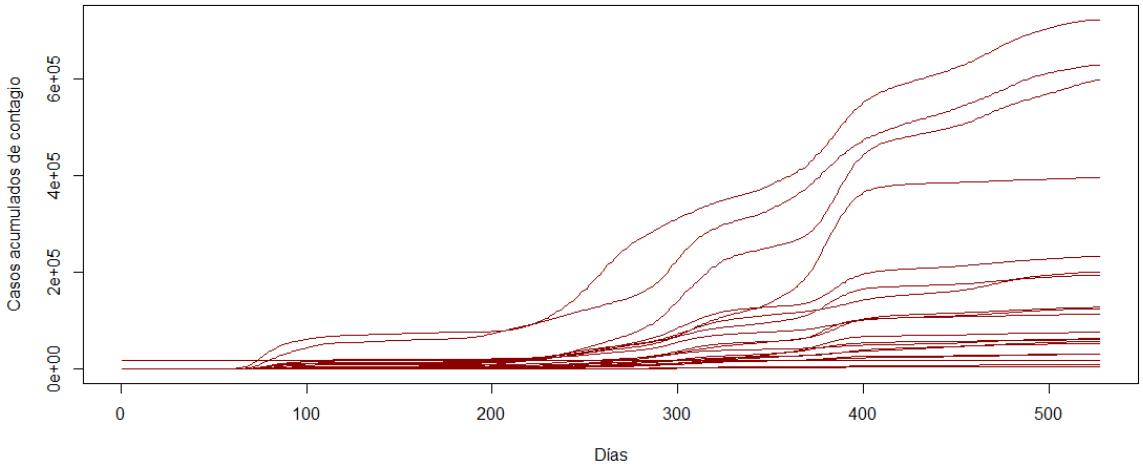


Figura 5.4: Acumulado de casos confirmados en las CCAA de España. Período 01/01/2020 a 10/06/2021.

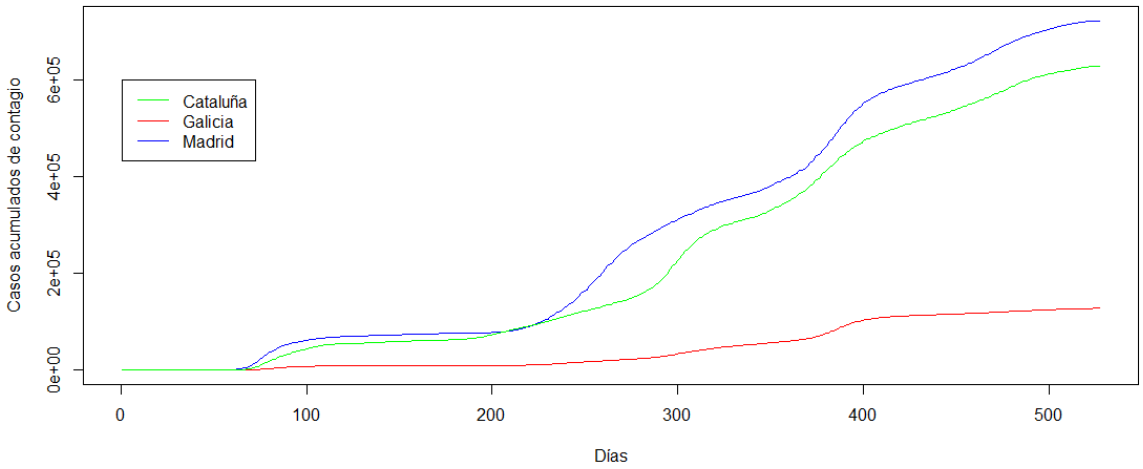


Figura 5.5: Acumulado de casos confirmados en las CCAA de Cataluña, Galicia y Madrid. Período 01/01/2020 a 10/06/2021.

Las figuras 5.4 muestran como curvas suaves el acumulado de casos de contagio a lo largo del período. Claramente, el aumento progresivo de casos en las comunidades con las características anteriormente mencionadas hace que el comportamiento acumulado de las que tienen menos casos se vea como una recta que no se aprecia su pendiente. Las comunidades de Cataluña y Madrid tiene pendientes más pronunciadas que Galicia *(ver figura 5.5)* por ser grandes ciudades, densamente pobladas y con entradas masivas de turistas diarias.

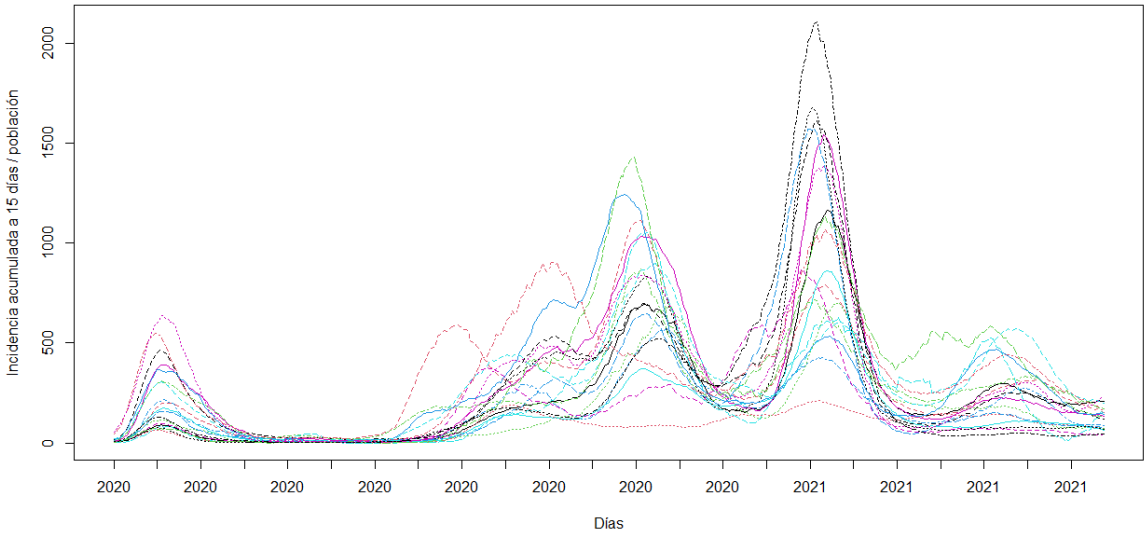


Figura 5.6: Incidencia acumulada de casos confirmados a 15 días en las CCAA de España por cada $100,000$ habitantes. Período 19/02/2020 a 03/06/2021.

El comportamiento de la incidencia de casos de contagio a 15 días con número de saltos $k = 1$ refleja un comportamiento similar que el de los casos diarios, se ven los picos de las oleadas cuando relajan las medidas y un valle cuando las restricciones se ajustan, la variabilidad de las curvas se ha reducido obteniendo curvas más suaves con dientes de sierra menos pronunciados. Con respecto a las curvas diarias el efecto de la población provocó un reducción considerable en las distancias del conjuntos de curvas. Ver figura 5.6.

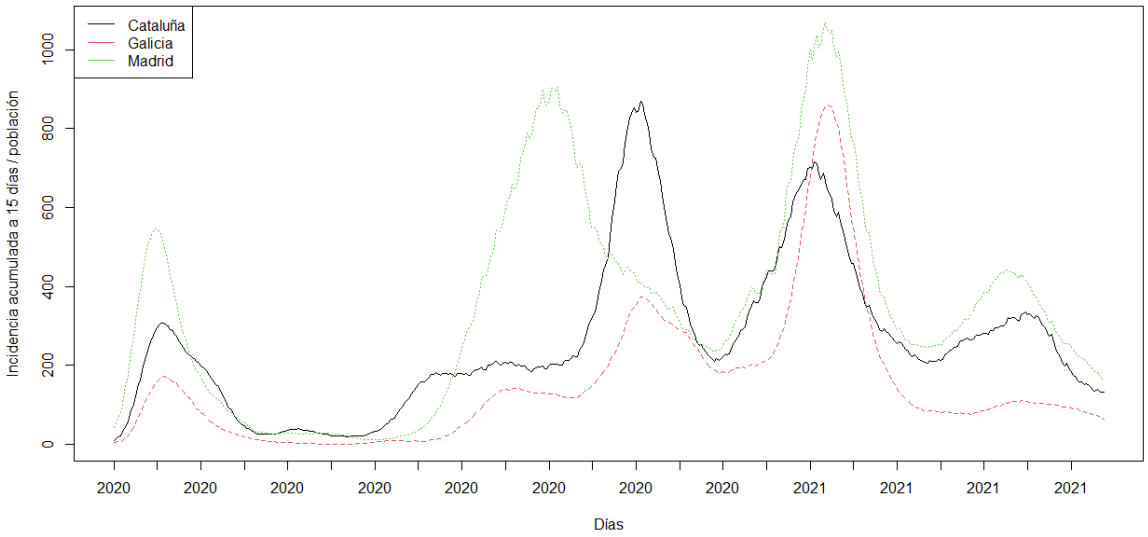


Figura 5.7: Incidencia acumulada a 15 días de casos confirmados en Cataluña, Galicia y Madrid por cada 100, 000 habitantes. 19 de febrero 2020 a 18 de mayo de 2021.

En la figura 5.7 se presenta la incidencia acumulada con respecto a la población de las comunidades de Cataluña, Galicia y Madrid. Se observa que la incidencia en Madrid de los pasados 15 días tiene un efecto desfavorable con respecto a las demás CCAA; Galicia se mantiene por debajo hasta la tercera oleada para épocas de las fiestas de navidad y año nuevo, donde su incidencia de casos supera a Cataluña, pero no a Madrid.

fecha	idx	H1	H2	H3	H4	H5	H6	H7	ccaa
2020-03-20	**16**	**246**	**260**	**277**	288	295	304	308	CT
2020-03-21	17	**260**	**277**	288	295	304	308	307	CT
2020-03-22	18	**277**	288	295	304	308	307	302	CT
2020-03-23	19	288	295	304	308	307	302	299	CT
2020-03-24	20	295	304	308	307	302	299	291	CT
2020-03-25	21	304	308	307	302	299	291	287	CT

Tabla 5.1: Predicciones a horizonte $H = 1, ..., 7$ de las curvas de incidencia a 15 días de la comunidad de Cataluña con saltos $k = 1$.

Para ver la metodología más clara se presenta la tabla 5.1 de las predicciones de incidencia de contagio de Covid-19 a horizonte $H = 1, ..., 7$ de las 6 primeras curvas funcionales pasado 15 días en la comunidad

de Cataluña (CT), cada curva da un salto de $k = 1$ días, esta variación se puede ver en el campo variable $idx = x_t^{(j)}, ...x_{t+k}^{(j)} = 16, 17, ...$, lo que significa que las curvas tamaño 15 aumentan en una unidad de tiempo. A modelo de ejemplo traducimos las predicciones a horizonte $H1 = 246$ como la incidencia acumulada de casos de contagio un día después del 20 de marzo del 2020 por cada 100,000 habitantes. El campo variable fecha sitúa el contexto en el que la población se encuentra, corresponden a días antes de iniciar la Semana Santa y pertenece a la curva tamaño 15 que predice los valores futuros.

5.2.1. Derivadas de las curvas de incidencia

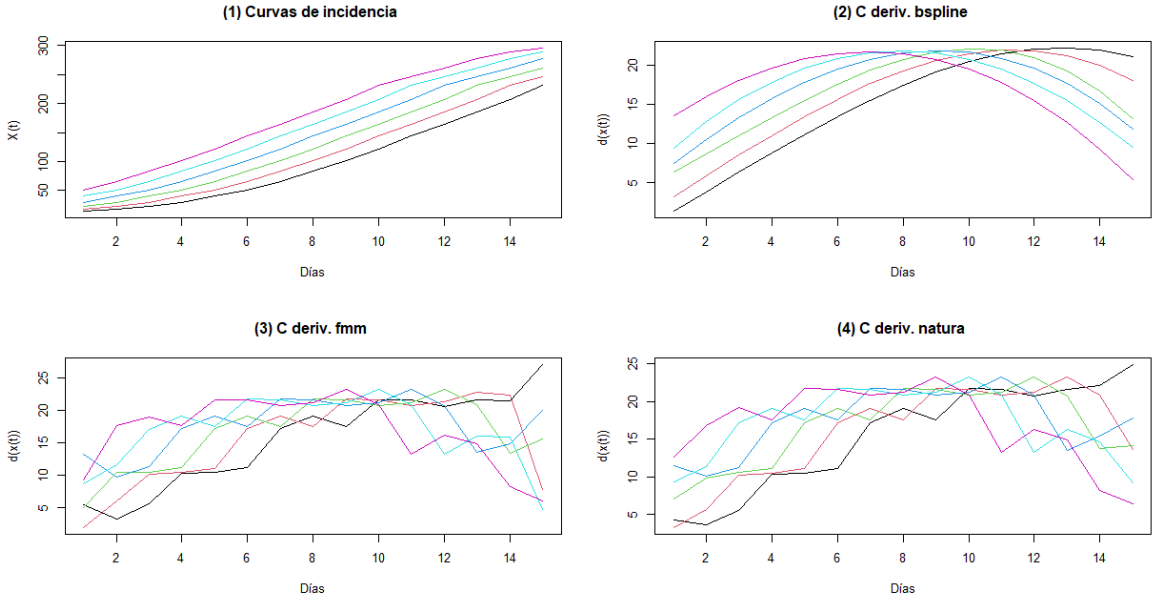


Figura 5.8: Curvas de incidencia acumulada y su derivada a 15 días de casos confirmados en Cataluña. Salto entre las cuervas de $K = 1$.

En la figura 5.8 se muestran las primeras 6 curvas funcionales de incidencia de Cataluña y sus derivadas por diferentes métodos. En el gráfico (1) las 6 curvas que ha su día 15 corresponden a las ultimas fechas del 20 al 25 de marzo del 2020 *(ver fechas tabla 5.1)*, estas tienen un comportamiento creciente durante el transcurso, fechas próximo a semana santa y luego de flexibilizar las medidas restrictivas de la primera oleada. El efecto de cambio de las derivadas en las curvas provocan cambios de dirección. El gráfico (2) que utiliza métodos de B-spline genera curvas suaves que al día 10 cambia de dirección, por otro lado, en los gráficos (3) y (4) las curvas vienen con más oscilaciones que se entremezclan y no permiten claramente diferenciar el comportamiento de entre curvas por la variabilidad aportada por estos métodos de derivación.

5.3. Predicciones de incidencia acumulada de contagio del Covid-19.

Para las predicciones se seleccionaron tres de las CCAA para tener estimaciones de manera local. Los modelos parciales[7] se generaron para las comunidades de Cataluña, Galicia y Madrid ($m = 1, 2, 3$).

[7] *Ver resultados de la aplicación de los modelos FRM en el apéndice B.*

Cada comunidad tiene un total de $i = 434$ de longitud $j = 1, ...n = 15$ construidas a partir de datos diarios correspondientes al 20 de marzo del 2020 hasta el 3 de junio del 2021 en saltos tamaño $k = 1$. El horizonte de predicción corresponde a los siete días después ($H = 1, ..., 7$) de la fecha final, específicamente los días del 4 al 10 de junio del 2021 comparados con las observaciones para misma fecha utilizando el MSEP ya visto en el capítulo de simulación.

En la tabla siguientes, se muestran los resultados de MSPE para la incidencia de contagio de cada uno de los modelos a un horizonte especifico ($H = 1, ..., 7$). Los resultados sombreados con gris claros representa los modelos con mejores resultados en el H especificado y los resultados con gris más oscuro y en negrita el modelo ganador en el horizonte.

5.3.1. Error cuadrático medio de las predicciones.

$$MSPE = \frac{1}{J} \sum_{j=n+1}^{n+J} \sum_{r=1}^{m} \left(y_{r+h}^{m} - \hat{y}_{n+h}^{m} \right)^2 \tag{5.2}$$

	$h+1, ..., H$						
MODELO	H_1 "04/06/2021"	H_2 "05/06/2021"	H_3 "06/06/2021"	H_4 "07/06/2021"	H_5 "08/06/2021"	H_6 "09/06/2021"	H_7 "10/06/2021"
LM	5	11	13	13	21	28	44
LM-AR(1)	3	4	2	6	8	10	24
LM-ARMA(1,1)	4	**0**	2	9	**3**	9	23
SAM	3	8	9	6	110	15	30
SAM-AR(1)	2	2	**1**	6	8	9	22
SAM-ARMA(1,1)	3	1	**0**	8	**3**	9	22
KAM	**2**	5	4	**1**	**3**	7	19
KAM-AR(1)	7	10	13	18	22	28	**46**
KAM-ARMA(1,1)	**2**	4	5	**2**	4	**0**	**12**
GLS-AR(1)	4	16	16	21	25	25	38
GLS-ARMA(1,1)	4	15	15	23	22	26	36
iGLS-AR(1)	4	16	16	21	25	25	38
iGLS-AR(P)	**2**	2	**1**	18	14	15	26
iGLS-ARMA(1,1)	8	3	**0**	26	18	310	23

Tabla 5.2: Error cuadrático medio de las predicciones de incidencia acumulada de contagio de la enfermedad Covid-19 a horizonte $H = 1, ..,7$ para las comunidades de Cataluña. Período 04/06/2020 al 10/06/2020.

En la tabla 5.2 se muestran los resultados del MSEP para la CCAA de Cataluña. A modo general se evidencia mejores resultados de predicción los modelos que no tienen restricciones de linealidad, la cual mejora al incorporarle las predicciones de un modelo AR(1) O ARMA(1,1). Los modelos con mejores resultados son el KAM-ARMA(1,1), seguido del modelo KAM. De cerca le siguen los modelos SAM-AR(1), SAM-ARMA(1,1) con buenas predicciones a lo largo del horizonte de predicción. Los modelos LM-AR(1) y iGLS-AR(P) presentan resultados bastante buenos en horizontes cortos de predicción del 1 al 3 luego aumentan considerablemente su valor de MSEP.

	$h+1,...,H$						
MODELO	H_1 "04/06/2021"	H_2 "05/06/2021"	H_3 "06/06/2021"	H_4 "07/06/2021"	H_5 "08/06/2021"	H_6 "09/06/2021"	H_7 "10/06/2021"
LM	5	9	11	12	16	21	26
LM-AR(1)	**4**	**7**	**8**	10	13	15	17
LM-ARMA(1,1)	**4**	**7**	9	11	14	14	18
SAM	5	**7**	**7**	**6**	**7**	**9**	**11**
SAM-AR(1)	**4**	**7**	**8**	9	12	13	14
SAM-ARMA(1,1)	**4**	**7**	9	10	13	**13**	15
KAM	20	25	27	30	35	41	47
KAM-AR(1)	17	22	25	28	33	37	41
KAM-ARMA(1,1)	19	25	29	34	41	44	51
GLS-AR(1)	**4**	8	9	11	15	17	19
GLS-ARMA(1,1)	**4**	8	10	12	16	16	19
iGLS-AR(1)	**4**	8	9	11	15	17	19
iGLS-AR(P)	**4**	8	9	11	15	17	19
iGLS-ARMA(1,1)	**4**	160	35	18	113	238	177

Tabla 5.3: Error cuadrático medio de las predicciones de incidencia acumulada de contagio de la enfermedad Covid-19 a horizonte $H = 1,..,7$ para las comunidades de Galicia. Período 04/06/2020 al 10/06/2020.

Para la CCAA de Galicia los modelos con mejores predicciones a lo largo del horizonte corresponden a los SAM, seguidos bien de cerca por sus modelos con versión iterativa SAM-AR(1) y SAM-ARMA(1,1). Los modelos LM-AR(1), LM-ARMA(1,1), GLS-AR(1), iGLS-ARMA(1) e iGLS-AR(P) obtienen buenos resultados de predicción, pero solo hasta el horizonte 3.

	$h+1,...,H$						
MODELO	H_1 "04/06/2021"	H_2 "05/06/2021"	H_3 "06/06/2021"	H_4 "07/06/2021"	H_5 "08/06/2021"	H_6 "09/06/2021"	H_7 "10/06/2021"
LM	**8**	14	16	18	27	34	45
LM-AR(1)	**8**	10	**9**	**12**	**14**	**14**	19
LM-ARMA(1,1)	9	**6**	11	**12**	**13**	**15**	18
SAM	12	24	30	36	47	55	69
SAM-AR(1)	12	18	20	25	30	30	38
SAM-ARMA(1,1)	14	15	21	26	28	30	36
KAM	41	35	31	28	21	17	**12**
KAM-AR(1)	52	53	54	55	56	61	68
KAM-ARMA(1,1)	39	43	39	39	40	42	45
GLS-AR(1)	9	26	23	30	36	31	36
GLS-ARMA(1,1)	10	**7**	14	31	36	31	33
iGLS-AR(1)	9	26	22	30	36	31	36
iGLS-AR(P)	9	26	22	30	36	31	36

Tabla 5.4: Error cuadrático medio de las predicciones de incidencia acumulada de contagio de la enfermedad Covid-19 a horizonte $H = 1,..,7$ para las comunidades de Madrid. Período 04/06/2020 al 10/06/2020.

En la CCAA de Madrid los resultados favorecen notablemente a los modelos LM-AR(1) y LM-ARMA(1,1) durante todo el horizonte de predicción con respecto al resto de modelos. Particularmente, para esta comunidad los modelos más próximos en buenos resultados de predicción corresponde los modelos GLS-AR(1), GLS-ARMA(1,1), iGLS-AR(1) e iGLS-ARMA(1,1), únicamente a horizonte de predicción 1.

Predicciones la comunidad de Cataluña

A continuación se presentas los gráficos de las predicciones *(color rojo)* a horizonte $H = 1, ..,7$ con su intervalos de confianza del 95 % *(color verde)* para la CCAA de Cataluña. En estos se muestran los modelos tratados en la teoría y en la simulación individualmente, comparados con los valores observados *(color negro)*.

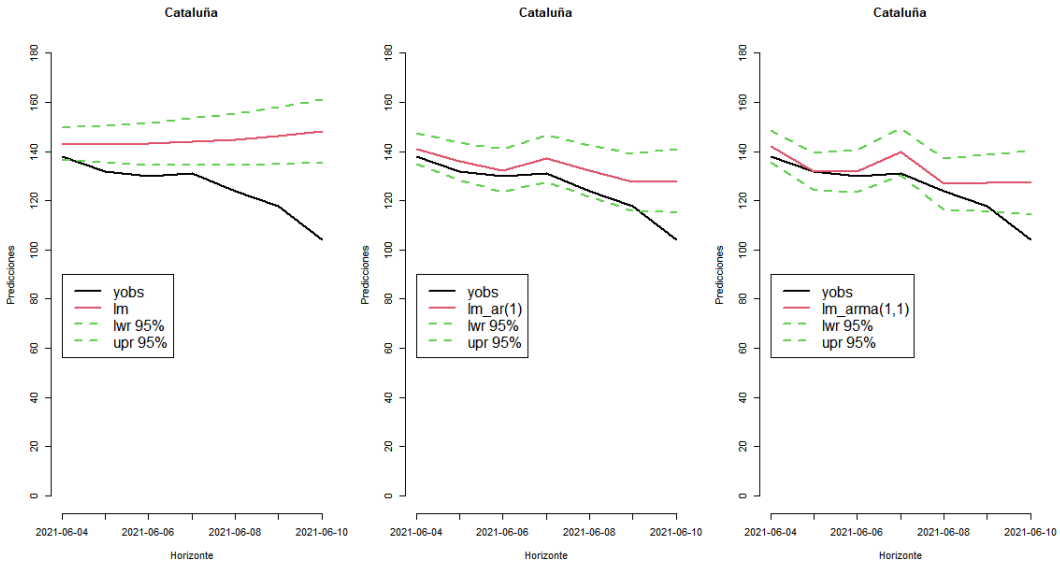


Figura 5.9: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo LM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

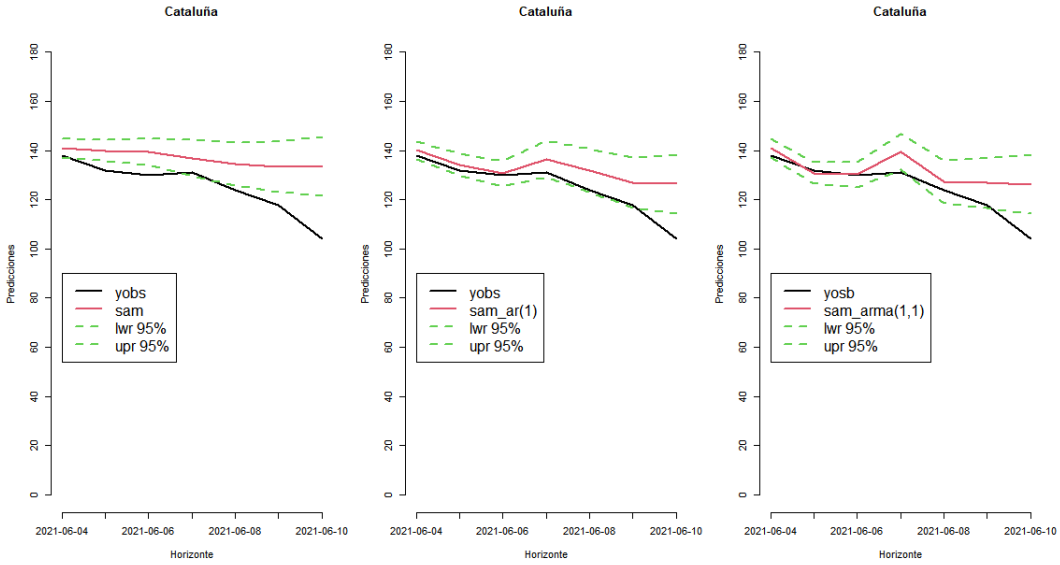


Figura 5.10: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo SAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

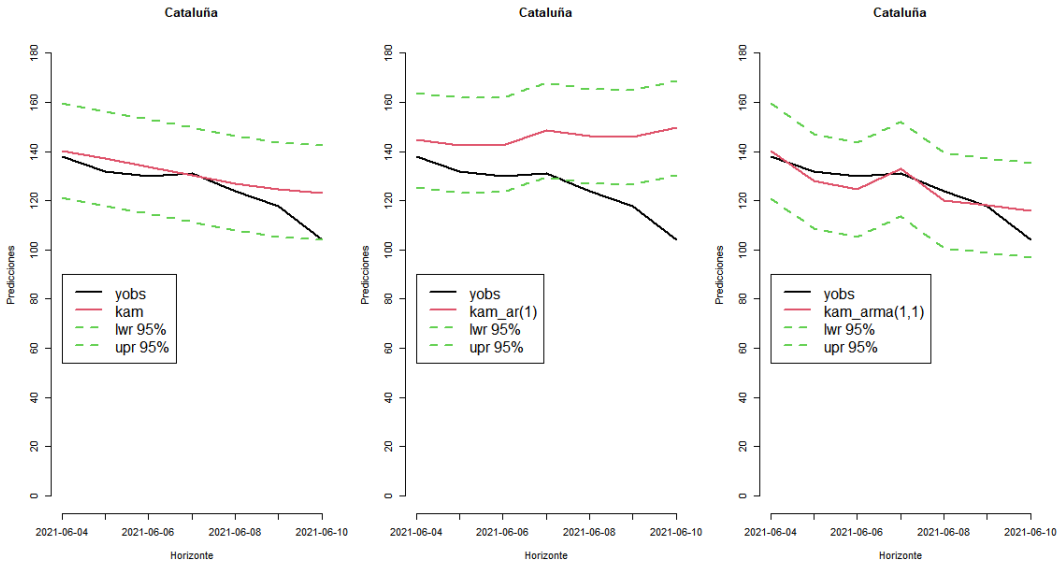


Figura 5.11: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo KAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

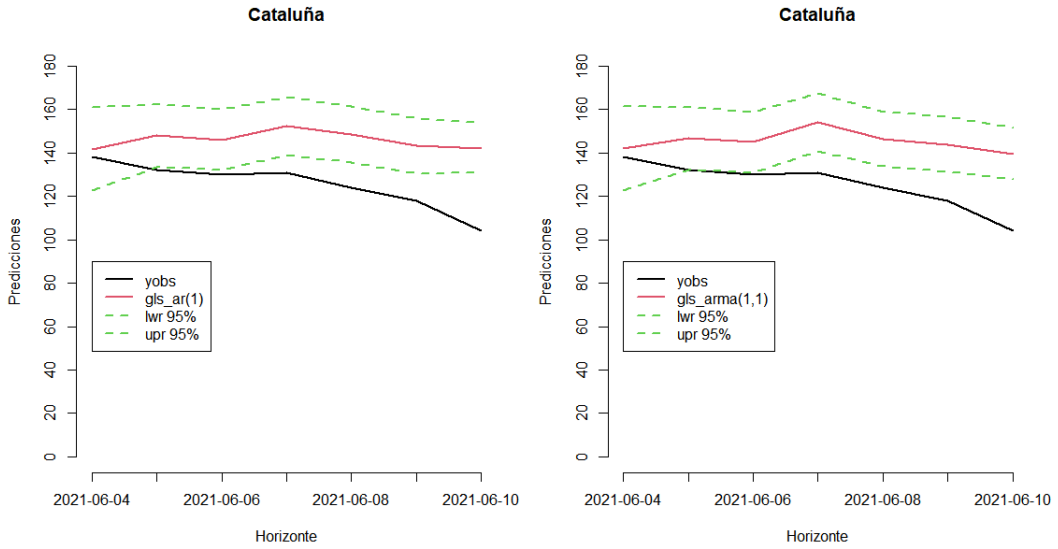


Figura 5.12: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo GLS-AR(1) y GLS-ARMA(1,1) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

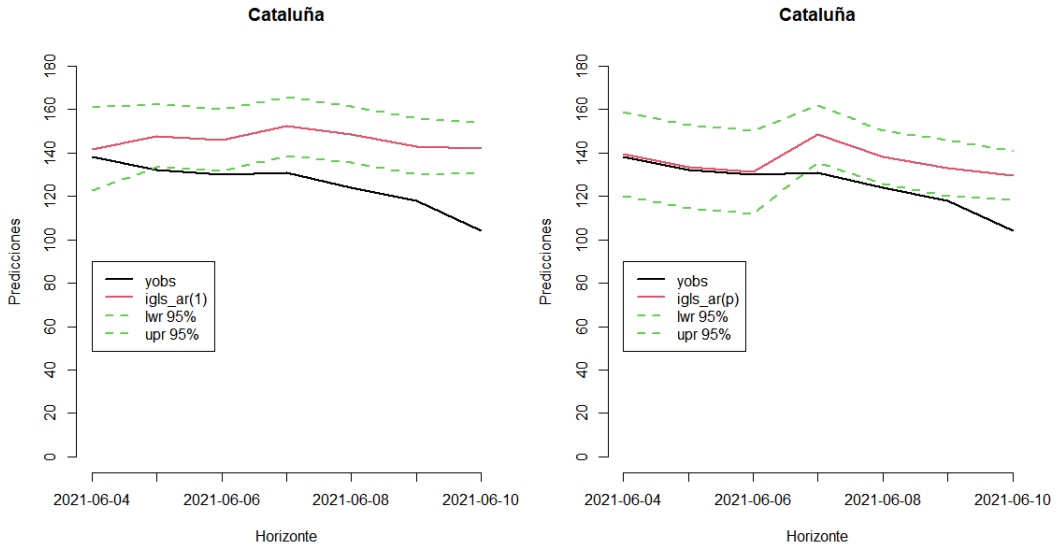


Figura 5.13: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo iGLS-AR(1) y iGLS-AR(P) con las observaciones en la comunidad de Cataluña por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

Los modelos LM (5.9) presenta unas predicciones cercanas pero con tendencia contraria a las observaciones y fuera de los intervalos de confianza. En cuanto a las salidas del modelo, presenta un R^2 ajustado alto y componentes significativas, gráficamente se evidencia violación los supuestos de normalidad, homocedasticidad e independencia de los errores *ver los resultados del modelo en (B.1.1) y las salidas gráficas en (B.1)*. La estructura de autocorrelación de los errores para este modelo es alta

ver (B.4) y (B.7), no tomarlas en cuenta refleja las debilidades en las predicciones, como se ve en la gráfica. Los resultados favorables son notables al incluirle su versión iterativa AR(1) y ARMA(1,1), las cuales se acercan considerablemente a los valores observados, siguen la tendencia, pero no logran captar la caída brusca de la predicción del horizonte $H = 7$. Viendo los gráficos del modelo SAM (5.10), es notable la mejora con respecto a los modelos FLM; sus salidas proporcionan valores similares *ver (B.2) y (B.1.1)* y gráficamente no se evidencia una corrección notable con respecto al comportamiento de los errores. La versión iterativa AR(1) y ARMA(1,1) del SAM mejora considerablemente, pero al igual que los modelos FLM se aleja fuera de los intervalos de predicción del horizonte $H = 7$. El modelo KAM (5.11) es hasta ahora el modelo con los resultados más próximos. Su versión iterativa AR(1) se aleja sobrestimando los valores, cambiando la tendencia y dejando las predicciones a $H = 6, 7$ fuera del límite inferior. Por otro lado, su versión ARMA(1,1) supera todos los modelos anteriores y su versión iterativa tejiendo los valores de las observaciones y manteniendo la tendencia a lo largo de los horizontes de predicción. Con estos resultados se confirma lo esperado de la tabla del MSEP *(5.2)*. Finalmente, los modelo GLS (5.12) e iGLS (5.13) solo tienen resultados razonables a horizontes de predicción cortos. *Ver resultados de los modelos en (B.1.1), (B.3), (B.4) y (B.1.1).*

Predicciones la comunidad de Galicia

A continuación se presentas los gráficos de las predicciones *(color rojo)* a horizonte $H = 1, ..,7$ con su intervalos de confianza del 95 % *color verde* para la CCAA de Galicia. En estos se muestran los modelos tratados en la teoría y en la simulación individualmente, comparados con los valores observados *(color negro)*.

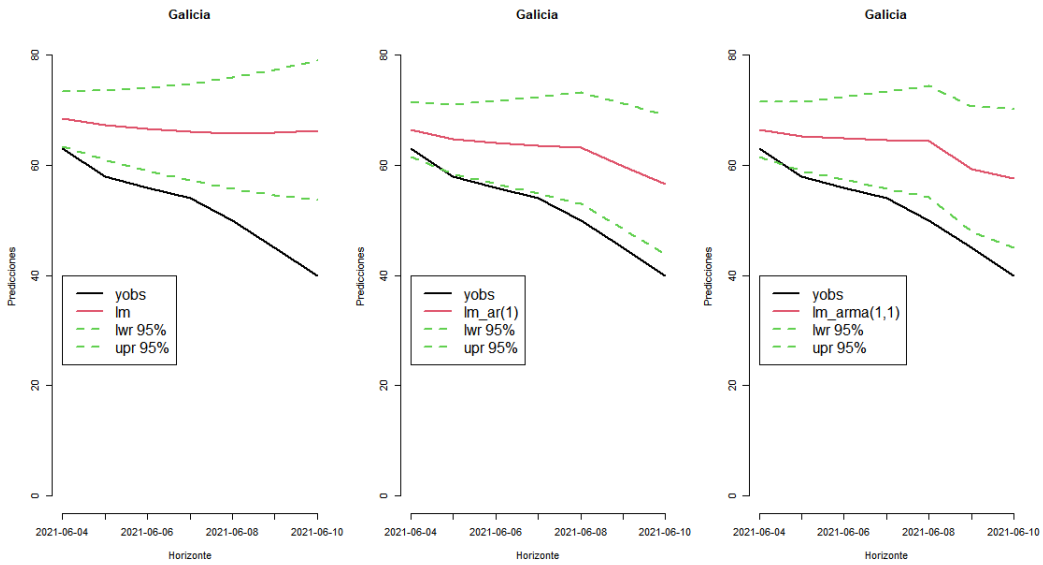


Figura 5.14: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo LM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

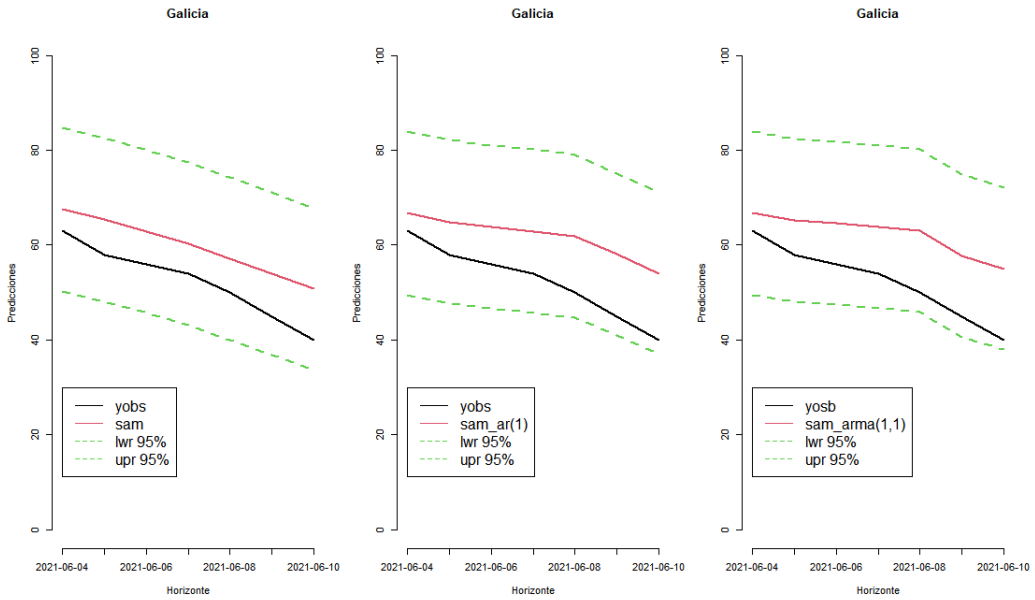


Figura 5.15: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo SAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

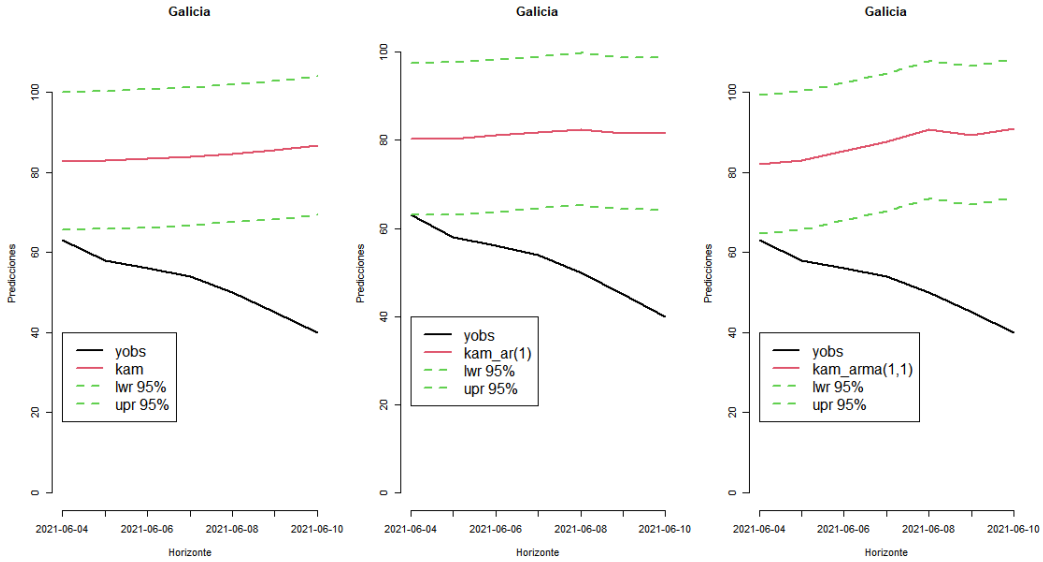


Figura 5.16: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo KAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

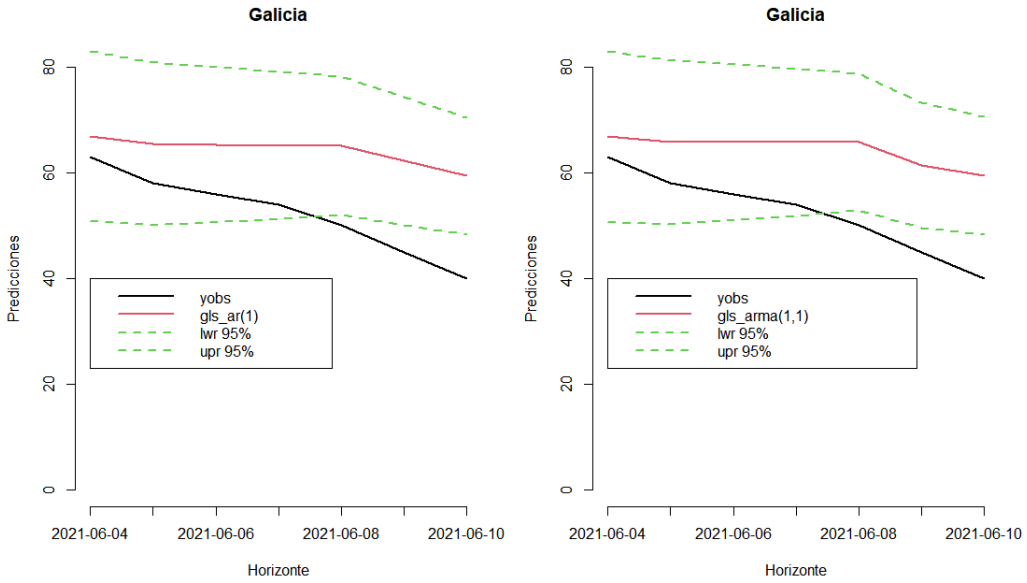


Figura 5.17: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo GLS-AR(1) y GLS-ARMA(1,1) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

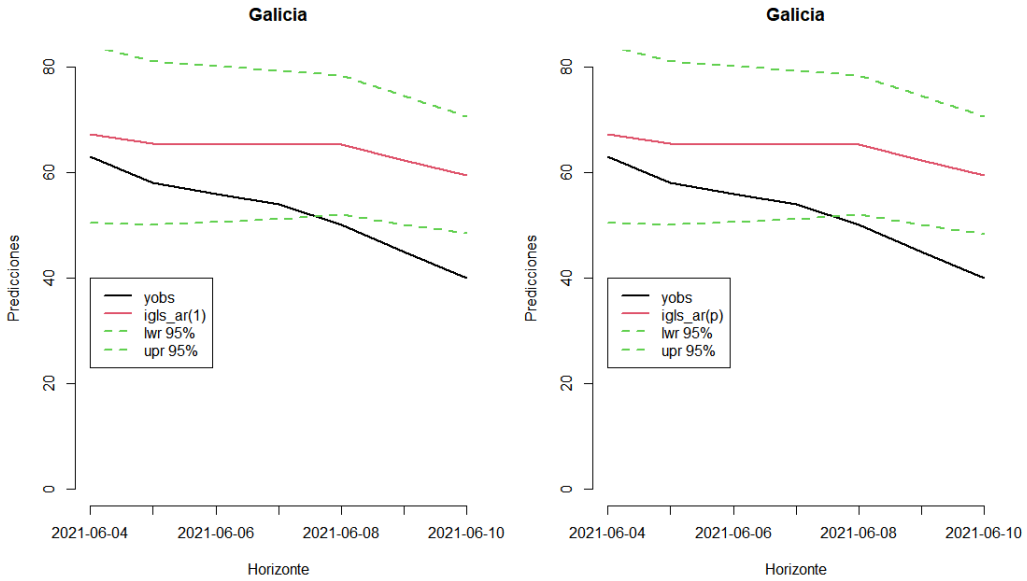


Figura 5.18: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo iGLS-AR(1) y iGLS-AR(P) con las observaciones en la comunidad de Galicia por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

Los resultados de predicción de Galicia para decidir cual es el modelo óptimo difieren de los resultados obtenidos de la CCAA de Cataluña. El modelos FLM (5.14) estima por encima los valores de las observaciones, sus versiones iterativa AR(1) y ARMA(1,1) mejoran considerablemente la tendencia

y su acercamiento al límite inferior. *Ver (B.1.2), (B.5).* Los modelos SAM (5.15), según el MSEP (5.3) son los modelos más acertados a nivel de predicción para CCAA de Galicia. Los resultados de las iteraciones AR(1) y ARMA(1,1) aumentan las distancias con respecto de las observaciones, sin embargo, en su favor detectan los picos de mejor manera y siguen la tendencia. *Ver (B.6) y (B.1.2*). El modelo KAM (5.16) y sus y sus iteraciones AR(1) y ARMA (1,1) en su construcción básica proporciona predicciones fatales, sobrestiman los valores y se amplía las distancia conforme un aumento adicional de H quedando todos los valores fuera de los intervalos de confianza y con tendencia opuesta a lo esperado. Finalmente, los modelos GLS (5.17) e iGLS (5.18) funcionan eficientemente hasta el tercer horizonte de predicción y ligeramente la tendencia a la baja de las observaciones. *Ver resultados y salidas de los modelos en el apéndice (B.1.2), (B.7), (B.8), (B.1.2).*

Predicciones la comunidad de Madrid

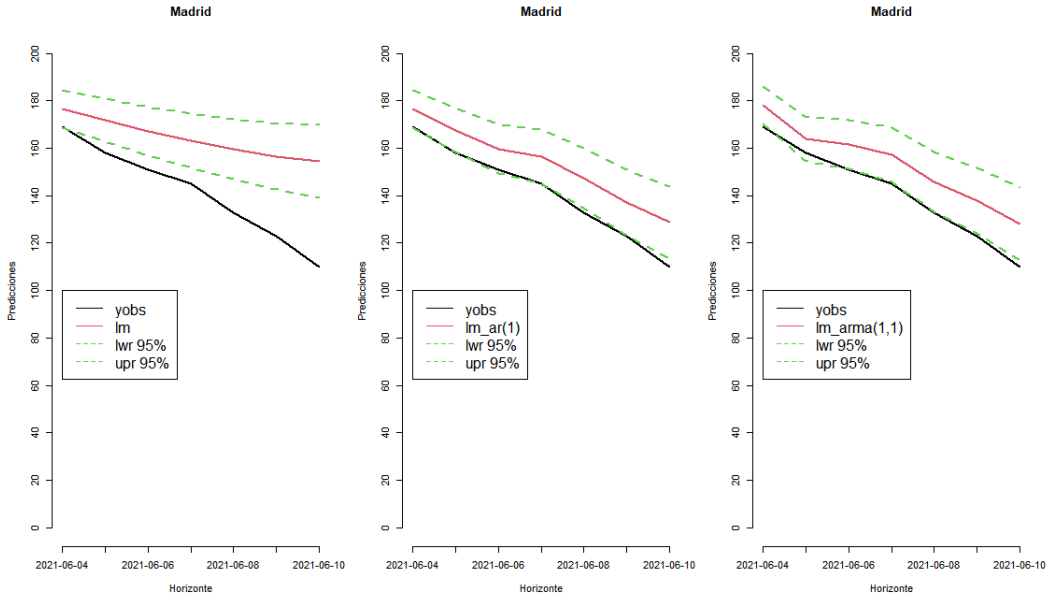


Figura 5.19: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo LM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

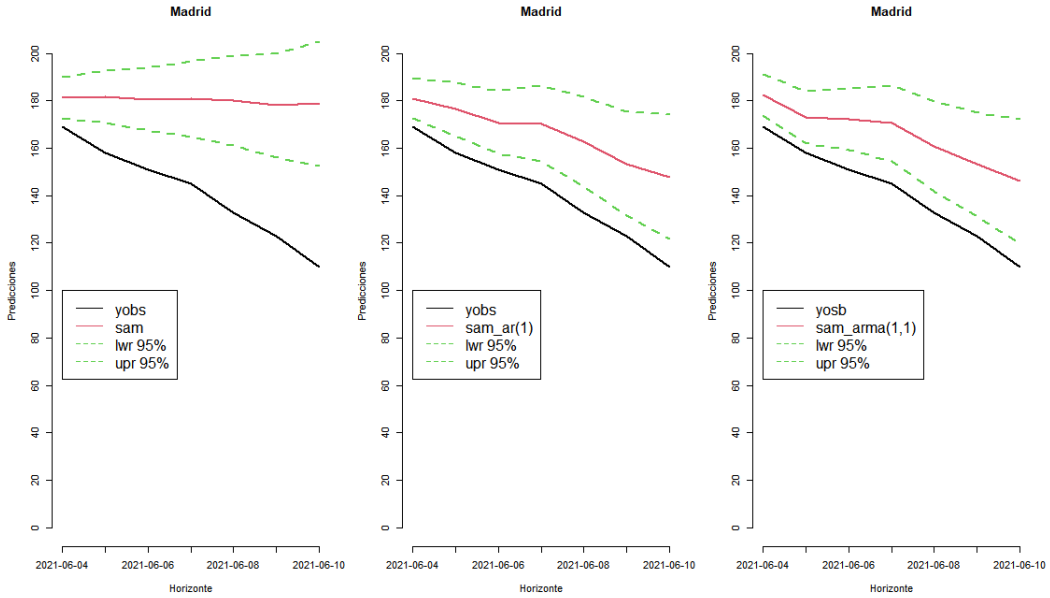


Figura 5.20: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo SAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

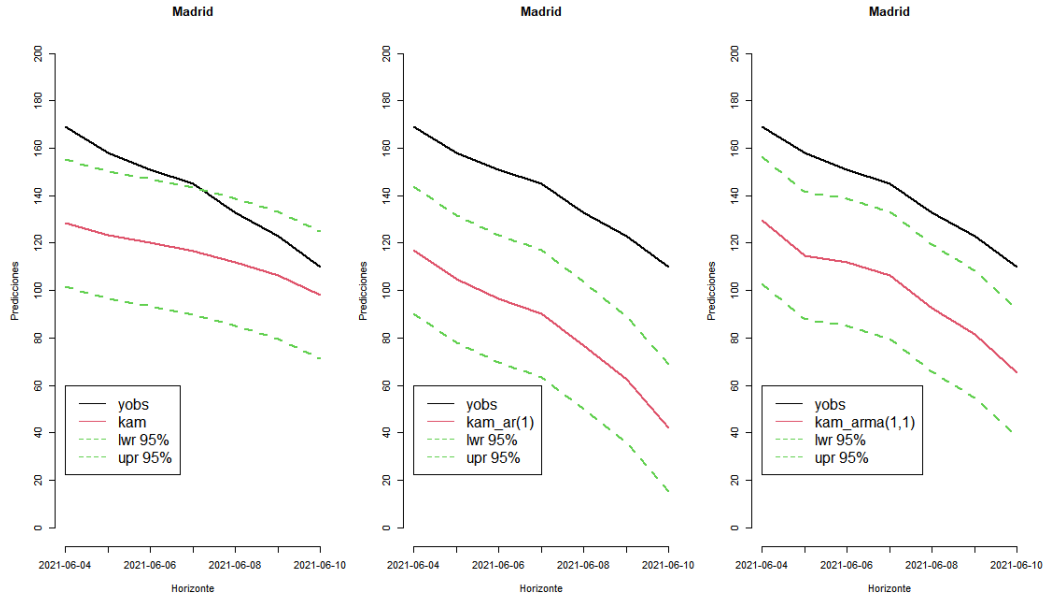


Figura 5.21: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo KAM, su versión iterativa AR(1) y ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

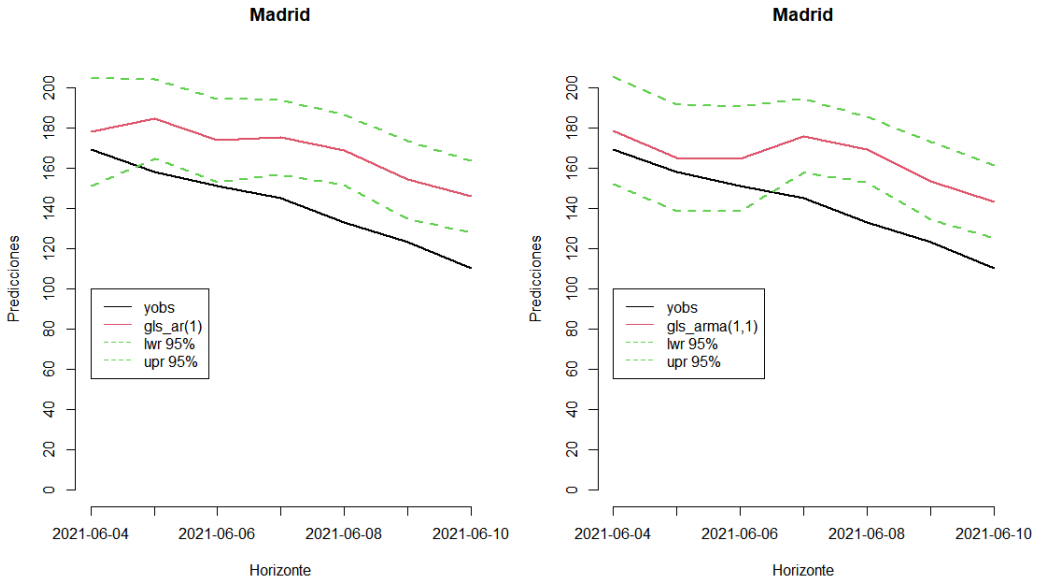


Figura 5.22: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo GLS-AR(1) y GLS-ARMA(1,1) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

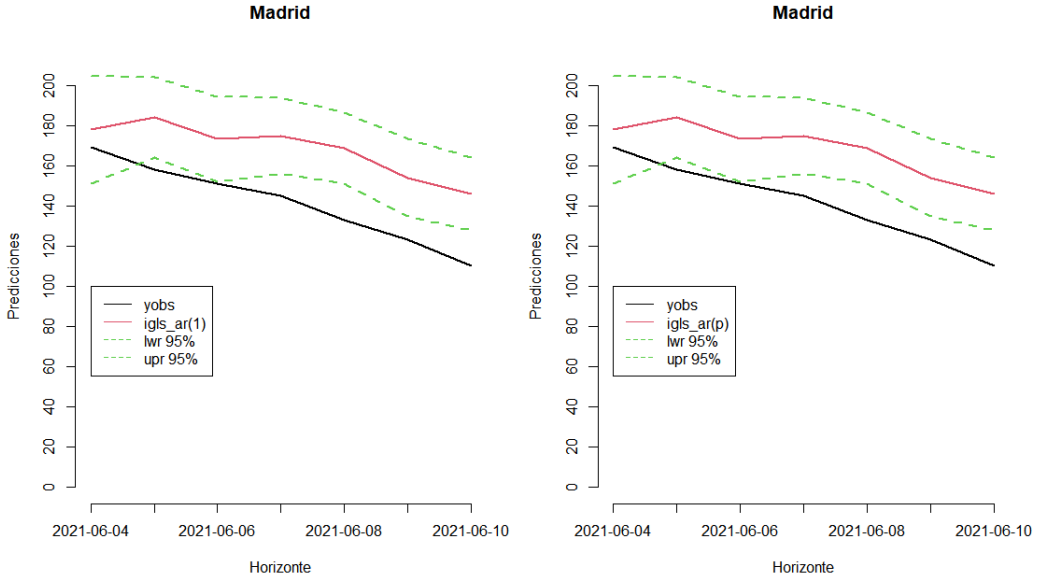


Figura 5.23: Comparativo de las predicciones de incidencia acumulada de contagio de Covid-19 del modelo iGLS-AR(1) y iGLS-AR(P) con las observaciones en la comunidad de Madrid por cada $100,000$ habitantes. $H = 1, ..., 7$. Período 04/06/2021 al 10/06/2021.

En la CCAA de Madrid las predicciones del modelo FLM (5.19) deja fuera del límite todos los, sin embargo, al agregar las predicciones de las iteraciones AR(1) y ARMA(1,1) mejoran considerablemente al modelo inicial. Además de seguir la tendencia conformemente, alcanza los valores del límite inferior

hasta traslaparse en más de 5 valores de los horizontes de predicción. *Ver (B.1.2), (B.9).* El modelo SAM (5.20) tiene un comportamiento similar al de los modelos FLM dado que sobrestima los valores de las observaciones y todos los valores fuera de los intervalos de predicción. Su versión iterativa aproxima los resultados al límite inferior, sin embargo, no tan eficientemente como los modelos FLM. Ver resultados y salidas gráficas del modelo en *B.1.2, B.10.* El modelos KAM (5.21) deja las predicciones se por debajo muy por debajo de las observaciones en los tres primeros valores del horizonte de predicción, el resto los predice favorablemente, las versiones iterativas AR(1) y ARMA(1,1) no proporcionan mejoría en las distancias, más bien se alejan del límite superior. Los modelos GLS (5.22) e iGLS (5.23) funcionan bien el primer horizonte de predicción. De estos modelos los mejores resultados corresponden al modelo iGLS-ARMA(1,1) con a penas los tres primeros valores de horizonte de predicción dentro de los intervalos de confianza. Ver resultados y salidas gráficas del modelo en *(B.1.2), (B.11) y (B.1.2).*

Capítulo 6

Conclusión

Como ya se ha visto en la simulación y en la aplicación a datos reales este documento pone en contexto las bondades de los diversos modelos de regresión funcional para respuesta escalar. Con la premisa de encontrar estructuras de dependencia de los errores, en este caso temporal $AR(1)$, $AR(P)$ y $ARMA(1,1)$. Se propusieron el modelos de mínimos cuadrados generalizados (GLS), ya que contempla estas estructuras de dependencia en modelos lineales y su versión iterativa iGLS que funcionan bastante bien aunque con una carga computacional más elevada. En el campo real, la relación entre las variables trascienden estas perspectivas y supuestos, por tanto, otros modelos de la estadística clásica y no paramétrica extendidos al marco funcional se evaluaron y compararon a partir de las predicciones bajo diversos escenarios llegando las siguientes conclusiones.

En la simulación de FRM para escalar y se demostró que agregando una estructura temporal de error ε, los modelos GLS e iGLS tienen mejores resultados al momento de predecir que el resto de modelo que no contemplan este tipo de estructuras, este resultado se afianza a valores mayores de la autocorrelación incorporada y con la salvedad que en este modelo simple la relación entre las variables es enteramente lineal. Cuando la relación entre las variables no es enteramente lineal *(aumenta o disminuye proporción de linealidad)* se comprueba que los modelos FLM y GLS no funcionan de manera óptima. Como alternativa queda evidenciado que los modelos SAM con sus propiedades de suavizado trabajan bastante bien. Escenarios donde hay evidencia de estructura de dependencia sus versiones iterativas iSAM-AR(1) e iSAM-ARMA(1,1) superan de manera más eficiente el modelo básico y los modelo donde el supuesto de linealidad es esencial. Particularmente, cuando la estructura de dependencia temporal de los errores corresponde a aun modelo AR(1) y AR(2) resultaron ser eficientes las predicciones en todos los horizontes sin importar que el valor del parámetro ϕ sea grande o pequeño. Por otro lado, cuando es un ARMA(1,1) son eficientes a todo horizonte de predicción si el valor de los parámetros de la estructura de dependencia y su variabilidad es baja.

En la aplicación se utilizaron los FRM para estimar la incidencia acumulada de contagio de la enfermedad del Covid-19 en las CCAA de España. Predecir la incidencia de estos casos tiene importancia al momento de aplicar políticas públicas de salud para minimizar las perdidas de vidas humanas, controlar el número de nuevos casos y recuperar la economía de país. Se estimó un modelo local simple para 3 de las 19 CCAA *(Cataluña, Galicia y Madrid)* por la carga computacional que implica modelar el número de curvas para el período. Este proceso de estimación arrojó resultados interesantes. De manera general, se evidenció correlación alta de los errores resultantes de los modelos; las versión iterativa minimiza el MSEP con respecto a los modelos clásicos en la mayoría de los casos. Los modelos óptimos obtenidos mediante el MSEP y los gráficos de predicción varían conforme a la CCAA trabajada.

Los modelos GLS e iGLS propuesto no fueron los más sobresalientes, fueron superados por el modelo

KAM y su versiones iterativas KAM-ARMA(1,1) para predecir los valores futuros de la CCAA de Cataluña. En las perdiciones de la CCAA de Galicia fueron superados por los modelos SAM, SAM-AR(1) y SAM-ARMA(1,1). Finalmente, en la CCAA de Madrid los modelos GLS e iGLS fueron superados por los modelos LM-AR(1) y LM-ARMA(1).

6.1. Investigaciones futuras

Este tesis de iniciación en el análisis de datos funcionales abre nuevas iniciativas de investigación como propuestas futuras de las cuales se resaltan a continuación

- Una ampliación de los procesos de simulación autoregresivos de medias móviles que tome en cuenta estructuras más complejas ARMA(p,q), así como otras estructuras de dependencia citadas en este documento como la espacial y espacio-temporal.
- La ampliación de modelo simple a multivariante donde se incluyan nuevas variables relacionadas al contexto del procesos de vacunación, medidas restritivas, sexo, grupos etareos, etc.
- Aplicar los modelos a el resto de CCAA incorporando nuevas variables y lograr resultados más interesantes.
- Aplicar estos modelos en otros escenarios, grupos pequeños, países como por ejemplo República Dominicana agregando otras variables de interés.
- Profundizar sobre los modelo de regresión no paramétricos con dependencia de los errores de *Opsomer y Wang (1991)* para extender al contexto de datos funcionales.

Apéndices

Apéndice A

Resultados de la simulación

A.1. Tablas

A.1.1. Número de bases estimado

	PC						**BSP**					
	$\delta = 1$			$\delta = 0.15$			$\delta = 1$			$\delta = 0.15$		
snr	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$
0.05	3.3	3.2	3.0	2.9	2.9	2.2	6.3	6.1	5.6	5.5	5.4	5.4
0.10	3.2	3.1	2.9	2.9	2.9	1.7	5.9	5.8	5.5	5.5	5.4	5.5

Tabla A.1: Promedio de Número de bases seleccionados por el criterio de GCCV en $n_B = 500$ del Modelo (a). AR(1).

	PC						**BSP**					
	$\delta = 1$			$\delta = 0.15$			$\delta = 1$			$\delta = 0.15$		
snr	$\phi_1 = 0{,}7*$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$
0.05	3.3	3.3	3.3	2.9	2.9	2.9	6.5	6.5	6.4	5.5	5.4	5.4
0.10	3.2	3.2	3.2	2.9	2.9	2.9	6.0	6.0	6.1	5.5	5.4	5.5

Tabla A.2: Promedio de Número de bases seleccionados por el criterio de GCCV en $n_B = 500$ del Modelo (a). AR(2).

A.1.2. MSE del parámetro β

		snr = 0.05						**snr = 0.10**					
		$\delta = 1$			$\delta = 0.15$			$\delta = 1$			$\delta = 0.15$		
Base	**Modelo**	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$
PC	LM	0.451	0.490	0.949	0.964	1.232	3.278	4.064	4.073	4.237	4.122	4.125	4.259
PC	GLS-AR(1)	0.450	0.465	0.580	0.958	1.001	1.221	4.064	4.074	4.280	4.122	4.126	4.134
PC	iGLS-AR(1)	0.450	0.465	0.580	0.958	1.001	1.221	4.064	4.074	4.280	4.122	4.126	4.134
PC	iGLS-AR(p)	0.450	0.466	0.580	0.963	1.006	1.224	4.064	4.073	4.281	4.122	4.126	4.122
BSP	LM	0.491	0.563	1.382	1.165	1.478	4.377	4.070	4.070	4.396	4.135	4.115	4.412
BSP	GLS-AR(1)	0.490	0.521	0.954	1.157	1.209	1.571	4.071	4.071	4.451	4.135	4.114	4.164
BSP	iGLS-AR(1)	0.490	0.521	0.954	1.157	1.209	1.572	4.071	4.071	4.451	4.135	4.114	4.164
BSP	iGLS-AR(p)	0.490	0.521	0.955	1.162	1.208	1.572	4.070	4.071	4.449	4.134	4.113	4.155

Tabla A.3: Error cuadrático medio del parámetro β. Modelo (a), $n_B = 500$. AR(1)

		snr = 0.05						**snr = 0.10**					
		$\delta = 1$			$\delta = 0.15$			$\delta = 1$			$\delta = 0.15$		
Base	**Modelo**	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$	$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$	$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$
PC	LM	0.442	0.440	0.457	0.939	0.937	0.991	4.070	4.066	4.075	4.128	4.119	4.126
PC	GLS-AR(2)	0.403	0.403	0.438	0.534	0.553	0.883	4.068	4.067	4.075	4.126	4.119	4.128
PC	iGLS-AR(2)	0.403	0.403	0.438	0.534	0.553	0.883	4.068	4.067	4.075	4.126	4.119	4.128
PC	iGLS-AR(p)	0.403	0.403	0.438	0.535	0.555	0.887	4.068	4.067	4.076	4.126	4.118	4.128
BSP	LM	0.482	0.494	0.513	1.145	1.179	1.294	4.065	4.061	4.082	4.126	4.108	4.133
BSP	GLS-AR(2)	0.422	0.434	0.495	0.684	0.740	1.138	4.068	4.061	4.081	4.127	4.105	4.132
BSP	iGLS-AR(2)	0.422	0.434	0.495	0.684	0.740	1.138	4.067	4.061	4.081	4.127	4.105	4.132
BSP	iGLS-AR(p)	0.423	0.434	0.495	0.686	0.741	1.140	4.067	4.061	4.082	4.128	4.104	4.133

Tabla A.4: Error cuadrático medio del parámetro β. Modelo (a), $n_B = 500$. AR(2).

A.1.3. MSE del parámetro ϕ

		PC						**BSP**					
		$\delta = 1$			$\delta = 0{,}15$			$\delta = 1$			$\delta = 0{,}15$		
snr	**modelo**	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$	$\phi = 0{,}1$	$\phi = 0{,}5$	$\phi = 0{,}9$
0.05	GLS-AR(1)	0.004	0.003	0.001	0.012	0.206	0.121	0.004	0.004	0.001	0.012	0.205	0.114
0.05	iGLS-AR(1)	0.004	0.003	0.001	0.013	0.209	0.124	0.004	0.003	0.001	0.013	0.208	0.114
0.10	GLS-AR(1)	0.004	0.003	0.001	0.012	0.167	0.054	0.004	0.003	0.001	0.012	0.167	0.048
0.10	iGLS-AR(1)	0.004	0.003	0.001	0.012	0.170	0.057	0.004	0.003	0.001	0.013	0.169	0.048

Tabla A.5: Error cuadrático medio del parámetro ϕ. Modelo (a), $n_B = 500$. AR(1).

		PC						BSP					
		$\delta = 1$			$\delta = 0{,}15$			$\delta = 1$			$\delta = 0{,}15$		
snr	**Modelo**	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}3$	$\phi_1 = 0{,}7$	$\phi_1 = 0{,}5$	$\phi_1 = 0{,}3$
0.05	GLS-AR(2)	0.004	0.004	0.004	0.423	0.213	0.081	0.004	0.004	0.004	0.422	0.213	0.081
0.05	iGLS-AR(2)	0.004	0.004	0.004	0.428	0.217	0.083	0.004	0.004	0.004	0.426	0.216	0.083
0.10	GLS-AR(2)	0.004	0.003	0.004	0.378	0.177	0.069	0.004	0.003	0.004	0.377	0.177	0.069
0.10	iGLS-AR(2)	0.004	0.003	0.004	0.382	0.180	0.071	0.004	0.003	0.004	0.381	0.180	0.070
snr	**Modelo**	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}4$	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}4$	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}4$	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}2$	$\phi_2 = 0{,}4$	$\phi_2 = 0{,}2$
0.05	GLS-AR(2)	0.004	0.003	0.004	0.030	0.134	0.039	0.004	0.003	0.004	0.030	0.134	0.039
0.05	iGLS-AR(2)	0.004	0.004	0.004	0.031	0.137	0.040	0.004	0.004	0.004	0.031	0.137	0.040
0.10	GLS-AR(2)	0.004	0.004	0.004	0.021	0.111	0.031	0.004	0.004	0.004	0.021	0.111	0.032
0.10	iGLS-AR(2)	0.004	0.004	0.004	0.022	0.113	0.033	0.004	0.004	0.004	0.022	0.113	0.033

Tabla A.6: Error cuadrático medio del parámetro ϕ_1 y ϕ_2 . Modelo (a), $n_B = 500$. AR(2).

A.1.4. MSPE de los modelos

			$AR(1)$, $NSL = 1$								
			$\phi = 0{,}1$			$\phi = 0{,}5$			$\phi = 0{,}9$		
snr	**Modelo**	**Bases**	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10
0.05	LM	PC	0.071	0.062	0.075	0.119	0.114	0.131	2.054	1.980	2.021
0.05	GLS-AR(1)	PC	0.070	0.062	0.074	0.097	0.114	0.130	0.367	1.233	1.777
0.05	iGLS-AR(1)	PC	0.070	0.062	0.074	0.097	0.114	0.130	0.368	1.233	1.772
0.05	iGLS-AR(p)	PC	0.072	0.062	0.074	0.097	0.114	0.130	0.371	1.229	1.782
0.05	GSAM	PC	0.073	0.063	0.075	0.121	0.114	0.132	2.074	1.947	2.047
0.05	iGSAM-AR(1)	PC	0.144	0.124	0.174	0.275	0.260	0.288	0.921	2.805	4.506
0.05	LM	BSP	0.071	0.063	0.076	0.122	0.116	0.130	2.068	2.000	2.076
0.05	GLS-AR(1)	BSP	0.071	0.063	0.075	0.098	0.115	0.129	0.366	1.231	1.794
0.05	iGLS-AR(1)	BSP	0.071	0.063	0.075	0.098	0.115	0.129	0.367	1.230	1.788
0.05	iGLS-AR(p)	BSP	0.072	0.063	0.075	0.098	0.115	0.128	0.372	1.220	1.789
0.05	GSAM	BSP	0.073	0.064	0.077	0.120	0.118	0.129	2.075	2.020	2.089
0.05	iGSAM-AR(1)	BSP	0.154	0.142	0.171	0.273	0.279	0.253	0.785	2.763	4.295
0.05	GKAM		0.375	0.343	0.337	0.426	0.383	0.407	2.675	2.688	2.879
0.05	iGKAM-AR(1)		0.388	0.356	0.333	0.414	0.416	0.433	0.939	1.903	2.755
0.10	LM	PC	0.145	0.138	0.143	0.250	0.267	0.277	3.379	3.409	3.597
0.10	GLS-AR(1)	PC	0.147	0.138	0.144	0.201	0.261	0.275	0.830	2.473	3.283
0.10	iGLS-AR(1)	PC	0.147	0.138	0.144	0.201	0.261	0.275	0.828	2.458	3.263
0.10	iGLS-AR(p)	PC	0.151	0.138	0.144	0.201	0.261	0.276	0.858	2.543	3.300
0.10	GSAM	PC	0.147	0.138	0.148	0.254	0.272	0.277	3.481	3.410	3.713
0.10	iGSAM-AR(1)	PC	0.144	0.124	0.174	0.275	0.260	0.288	0.921	2.805	4.506
0.10	LM	BSP	0.148	0.140	0.144	0.257	0.272	0.279	3.420	3.452	3.631
0.10	GLS-AR(1)	BSP	0.150	0.140	0.144	0.208	0.265	0.276	0.789	2.465	3.303
0.10	iGLS-AR(1)	BSP	0.150	0.140	0.144	0.208	0.265	0.276	0.787	2.450	3.283
0.10	iGLS-AR(p)	BSP	0.153	0.140	0.144	0.209	0.265	0.277	0.803	2.502	3.288
0.10	GSAM	BSP	0.154	0.141	0.147	0.258	0.276	0.278	3.328	3.509	3.743
0.10	iGSAM	BSP	0.154	0.142	0.171	0.273	0.279	0.253	0.785	2.763	4.295
0.10	GKAM		0.494	0.287	0.427	0.561	0.460	0.551	3.557	3.507	4.082
0.10	iGKAM		0.493	0.287	0.425	0.484	0.485	0.567	1.128	3.074	4.594

Tabla A.7: Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(1). $\delta = 1$.

			$AR(1)$, $\delta = 0{,}15$								
			$\phi = 0{,}1$			$\phi = 0{,}5$			$\phi = 0{,}9$		
snr	**Modelo**	**Bases**	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10
0.05	LM	PC	0.095	0.086	0.086	0.097	0.086	0.117	0.239	0.250	0.235
0.05	GLS-AR(1)	PC	0.094	0.086	0.086	0.097	0.086	0.116	0.158	0.239	0.232
0.05	iGLS-AR(1)	PC	0.094	0.086	0.086	0.097	0.086	0.116	0.158	0.239	0.232
0.05	iGLS-AR(p)	PC	0.094	0.086	0.086	0.098	0.087	0.117	0.139	0.213	0.221
0.05	GSAM	PC	0.007	0.006	0.008	0.011	0.011	0.013	0.171	0.171	0.171
0.05	iGSAM-AR(1)	PC	0.007	0.007	0.006	0.009	0.012	0.015	0.052	0.114	0.191
0.05	LM	BSP	0.094	0.086	0.086	0.100	0.086	0.117	0.237	0.243	0.235
0.05	GLS-AR(1)	BSP	0.094	0.086	0.086	0.100	0.086	0.117	0.155	0.228	0.229
0.05	iGLS-AR(1)	BSP	0.094	0.086	0.086	0.100	0.086	0.117	0.155	0.228	0.229
0.05	iGLS-AR(p)	BSP	0.093	0.086	0.086	0.101	0.086	0.117	0.138	0.203	0.216
0.05	GSAM	BSP	0.007	0.006	0.008	0.011	0.012	0.013	0.170	0.168	0.171
0.05	iGSAM-AR(1)	BSP	0.007	0.008	0.008	0.008	0.015	0.014	0.049	0.108	0.199
0.05	GKAM		0.021	0.017	0.021	0.024	0.023	0.023	0.204	0.191	0.203
0.05	iGKAM-AR(1)		0.021	0.017	0.021	0.023	0.025	0.023	0.065	0.118	0.204
0.10	LM	PC	0.095	0.088	0.089	0.103	0.101	0.099	0.359	0.370	0.375
0.10	GLS-AR(1)	PC	0.095	0.088	0.089	0.103	0.101	0.099	0.196	0.323	0.363
0.10	iGLS-AR(1)	PC	0.095	0.088	0.089	0.103	0.101	0.099	0.196	0.324	0.364
0.10	iGLS-AR(p)	PC	0.096	0.088	0.089	0.105	0.101	0.099	0.197	0.303	0.347
0.10	GSAM	PC	0.014	0.013	0.014	0.021	0.023	0.024	0.288	0.287	0.313
0.10	iGSAM-AR(1)	PC	0.014	0.011	0.017	0.022	0.021	0.021	0.079	0.245	0.366
0.10	LM	BSP	0.094	0.087	0.090	0.106	0.100	0.102	0.356	0.374	0.377
0.10	GLS-AR(1)	BSP	0.095	0.087	0.090	0.106	0.100	0.102	0.189	0.317	0.360
0.10	iGLS-AR(1)	BSP	0.095	0.087	0.090	0.106	0.100	0.102	0.189	0.318	0.361
0.10	iGLS-AR(p)	BSP	0.095	0.087	0.090	0.106	0.100	0.102	0.187	0.302	0.349
0.10	GSAM	BSP	0.014	0.013	0.013	0.022	0.024	0.024	0.276	0.294	0.322
0.10	iGSAM-AR(1)	BSP	0.014	0.013	0.017	0.023	0.023	0.019	0.064	0.219	0.337
0.10	GKAM		0.021	0.018	0.027	0.039	0.032	0.036	0.273	0.270	0.328
0.10	iGKAM-AR(1)		0.021	0.018	0.027	0.032	0.033	0.038	0.074	0.239	0.335

Tabla A.8: Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(1). *delta* = 0,15.

			$AR(2)$, $\delta = 1$								
			$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$			$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$			$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$		
snr	**Modelo**	**Bases**	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10
0.05	LM	PC	0.070	0.063	0.079	0.075	0.069	0.070	0.070	0.067	0.074
0.05	GLS-AR(2)	PC	0.017	0.042	0.065	0.020	0.040	0.058	0.057	0.066	0.073
0.05	iGLS-AR(2)	PC	0.017	0.041	0.065	0.020	0.040	0.058	0.057	0.066	0.073
0.05	iGLS-AR(p)	PC	0.017	0.042	0.065	0.020	0.040	0.058	0.057	0.066	0.073
0.05	GSAM	PC	0.071	0.065	0.078	0.075	0.070	0.071	0.070	0.067	0.076
0.05	iGSAM-AR(2)	PC	0.020	0.050	0.077	0.024	0.059	0.085	0.062	0.074	0.075
0.05	LM	BSP	0.071	0.065	0.079	0.076	0.070	0.072	0.072	0.069	0.076
0.05	GLS-AR(2)	BSP	0.017	0.042	0.065	0.020	0.039	0.058	0.058	0.068	0.075
0.05	iGLS-AR(2)	BSP	0.017	0.042	0.065	0.020	0.039	0.058	0.058	0.068	0.075
0.05	iGLS-AR(p)	BSP	0.017	0.042	0.064	0.020	0.040	0.059	0.058	0.068	0.075
0.05	GSAM	BSP	0.071	0.065	0.079	0.075	0.071	0.073	0.075	0.069	0.077
0.05	iGSAM-AR(2)	BSP	0.023	0.050	0.078	0.025	0.060	0.085	0.070	0.072	0.077
0.05	GKAM		0.338	0.320	0.334	0.394	0.307	0.355	0.378	0.439	0.393
0.05	iGKAM-AR(2)		0.340	0.425	0.345	0.423	0.355	0.380	0.382	0.451	0.425
0.10	LM	PC	0.127	0.138	0.138	0.140	0.144	0.158	0.135	0.145	0.131
0.10	GLS-AR(2)	PC	0.034	0.088	0.116	0.040	0.086	0.126	0.122	0.147	0.129
0.10	iGLS-AR(2)	PC	0.034	0.088	0.116	0.040	0.086	0.126	0.122	0.147	0.129
0.10	iGLS-AR(p)	PC	0.035	0.089	0.117	0.039	0.085	0.125	0.122	0.148	0.129
0.10	GSAM	PC	0.128	0.139	0.141	0.142	0.145	0.157	0.138	0.152	0.136
0.10	iGSAM-AR(2)	PC	0.035	0.108	0.162	0.059	0.089	0.172	0.105	0.161	0.159
0.10	LM	BSP	0.130	0.137	0.137	0.141	0.149	0.161	0.138	0.144	0.133
0.10	GLS-AR(2)	BSP	0.034	0.088	0.112	0.040	0.087	0.127	0.122	0.146	0.132
0.10	iGLS-AR(2)	BSP	0.034	0.088	0.112	0.040	0.087	0.127	0.122	0.146	0.132
0.10	iGLS-AR(p)	BSP	0.035	0.089	0.114	0.040	0.088	0.127	0.122	0.147	0.132
0.10	GSAM	BSP	0.131	0.139	0.140	0.142	0.147	0.165	0.141	0.151	0.138
0.10	iGSAM-AR(2)	BSP	0.037	0.121	0.171	0.071	0.095	0.169	0.110	0.155	0.159
0.10	GKAM		0.497	0.345	0.455	0.398	0.358	0.481	0.431	0.375	0.414
0.10	iGKAM-AR(2)		0.501	0.377	0.498	0.355	0.386	0.573	0.479	0.433	0.468

Tabla A.9: Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(2). $\delta = 1$.

			$AR(2)$, $\delta = 0{,}15$								
			$\phi_1 = 0{,}7$ $\phi_2 = 0{,}2$			$\phi_1 = 0{,}5$ $\phi_2 = 0{,}4$			$\phi_1 = 0{,}3$ $\phi_2 = 0{,}2$		
snr	**Modelo**	**Bases**	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10	t+ 1	t+ 5	t+ 10
0.05	LM	PC	0.098	0.087	0.086	0.092	0.082	0.109	0.079	0.093	0.081
0.05	GLS-AR(2)	PC	0.096	0.086	0.086	0.093	0.082	0.109	0.080	0.093	0.081
0.05	iGLS-AR(2)	PC	0.096	0.086	0.086	0.093	0.082	0.109	0.080	0.093	0.081
0.05	iGLS-AR(p)	PC	0.096	0.086	0.086	0.092	0.082	0.109	0.079	0.093	0.081
0.05	GSAM	PC	0.007	0.006	0.008	0.007	0.007	0.007	0.007	0.007	0.008
0.05	iGSAM-AR(2)	PC	0.002	0.005	0.007	0.003	0.006	0.008	0.009	0.008	0.009
0.05	LM	BSP	0.097	0.087	0.087	0.094	0.082	0.110	0.081	0.091	0.082
0.05	GLS-AR(2)	BSP	0.095	0.087	0.087	0.096	0.082	0.110	0.081	0.091	0.082
0.05	iGLS-AR(2)	BSP	0.095	0.087	0.087	0.096	0.082	0.110	0.081	0.091	0.082
0.05	iGLS-AR(p)	BSP	0.095	0.087	0.087	0.095	0.082	0.110	0.080	0.091	0.082
0.05	GSAM	BSP	0.007	0.006	0.008	0.008	0.008	0.008	0.007	0.007	0.008
0.05	iGSAM-AR(2)	BSP	0.004	0.006	0.008	0.003	0.007	0.010	0.008	0.008	0.011
0.05	GKAM		0.019	0.015	0.020	0.020	0.017	0.016	0.024	0.021	0.023
0.05	iGKAM-AR(2)		0.018	0.019	0.022	0.020	0.020	0.019	0.024	0.022	0.023
0.10	LM	PC	0.093	0.086	0.091	0.093	0.093	0.090	0.096	0.096	0.090
0.10	GLS-AR(2)	PC	0.094	0.086	0.091	0.094	0.094	0.090	0.096	0.096	0.090
0.10	iGLS-AR(2)	PC	0.094	0.086	0.091	0.094	0.094	0.090	0.096	0.096	0.090
0.10	iGLS-AR(p)	PC	0.095	0.086	0.091	0.094	0.093	0.090	0.096	0.096	0.090
0.10	GSAM	PC	0.012	0.013	0.013	0.012	0.013	0.014	0.014	0.015	0.014
0.10	iGSAM-AR(2)	PC	0.006	0.012	0.016	0.005	0.008	0.013	0.010	0.015	0.014
0.10	LM	BSP	0.093	0.086	0.093	0.095	0.093	0.093	0.096	0.099	0.091
0.10	GLS-AR(2)	BSP	0.094	0.085	0.093	0.095	0.093	0.093	0.097	0.098	0.091
0.10	iGLS-AR(2)	BSP	0.094	0.085	0.093	0.095	0.093	0.093	0.097	0.098	0.091
0.10	iGLS-AR(p)	BSP	0.094	0.086	0.093	0.096	0.092	0.093	0.096	0.099	0.091
0.10	GSAM	BSP	0.013	0.013	0.014	0.012	0.014	0.014	0.014	0.016	0.014
0.10	iGSAM-AR(2)	BSP	0.007	0.013	0.016	0.008	0.009	0.013	0.011	0.015	0.014
0.10	GKAM		0.021	0.022	0.026	0.025	0.024	0.029	0.026	0.025	0.021
0.10	iGKAM-AR(2)		0.021	0.023	0.031	0.017	0.024	0.034	0.027	0.028	0.021

Tabla A.10: Error cuadrático medio de las predicciones a horizonte 1, 5 y 10. Modelo (a), $n_B = 500$. AR(2). $\delta = 0{,}15$.

A.2. Resumen de los modelos

A.2.1. Modelos con ε generado por proceso ARMA(1,1) y $\delta = 1$

```
summary(res_lm.pc)
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-2.15989 -0.49978 -0.04725  0.49928  2.23531
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10684    0.04928   2.168   0.0311 *
x.PC1       -1.50699    0.07683 -19.614   <2e-16 ***
x.PC2       -3.08635    0.22617 -13.646   <2e-16 ***
x.PC3       -4.27303    0.35802 -11.935   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

```
Residual standard error: 0.7792 on 246 degrees of freedom
Multiple R-squared:  0.7441,  Adjusted R-squared:  0.7409
F-statistic: 238.4 on 3 and 246 DF,  p-value: < 2.2e-16
```

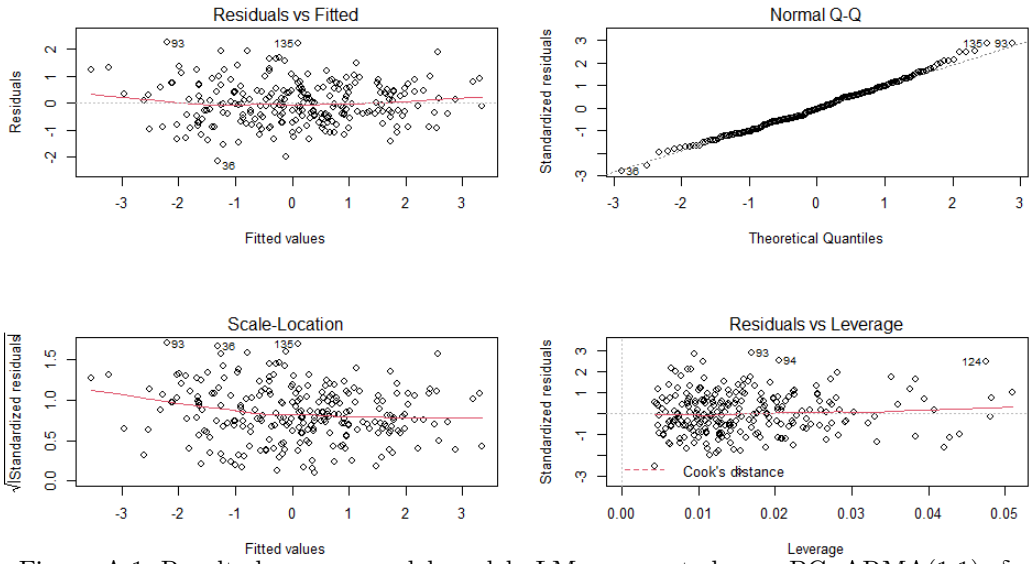


Figura A.1: Resultados resumen del modelo LM representado por PC. ARMA(1,1). $\delta = 1$

```
> summary(res_gls.pc)
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  446.155 470.6923 -216.0775
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
      Phi1     Theta1
 0.7465880 -0.1015861
Coefficients:
                Value  Std.Error    t-value p-value
(Intercept)  0.104096 0.12566023   0.828395  0.4082
x.PC1       -1.420746 0.04625874 -30.713030  0.0000
x.PC2       -2.932424 0.14048970 -20.872879  0.0000
x.PC3       -3.671562 0.21507636 -17.070968  0.0000
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1  0.000
x.PC2  0.000 -0.030
x.PC3 -0.005  0.068 -0.038
Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.78917386 -0.69735189 -0.05489716  0.68898214  2.86684215
Residual standard error: 0.7893607
Degrees of freedom: 250 total; 246 residual
```

```
> res_igls.pc
Call:
list("fregre.pc")

Coefficients:
```

```
(Intercept)        x.PC1        x.PC2        x.PC3
     0.1059      -1.4086      -2.9390      -3.6263
```

```
> summary(res_lm.bsp)
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-2.1999  -0.5396  -0.0842   0.4930   2.2166
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.10684    0.04925   2.169   0.0310 *
x.bspl4.1    -9.00882    4.76767  -1.890   0.0600 .
x.bspl4.2    27.02550    4.43593   6.092  4.3e-09 ***
x.bspl4.3   -14.52732    5.77085  -2.517   0.0125 *
x.bspl4.4     0.57382    4.21232   0.136   0.8918
x.bspl4.5     5.19667    2.83145   1.835   0.0677 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 0.7787 on 244 degrees of freedom
Multiple R-squared:  0.7464,  Adjusted R-squared:  0.7412
F-statistic: 143.6 on 5 and 244 DF,  p-value: < 2.2e-16
```

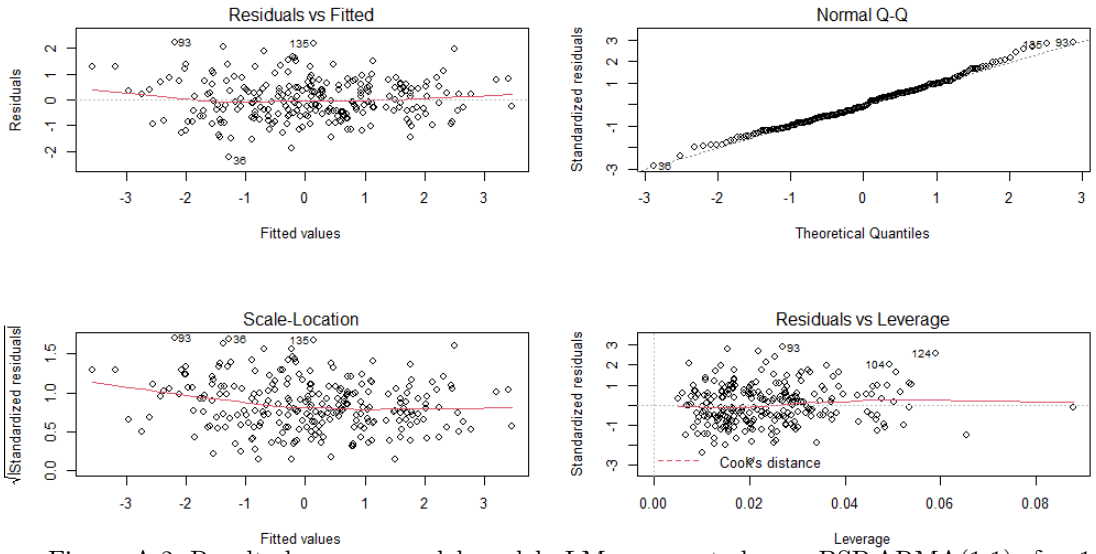


Figura A.2: Resultados resumen del modelo LM representado por BSP.ARMA(1,1). $\delta = 1$

```
> summary(res_gls.bsp)
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  430.0935 461.568 -206.0467
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
       Phi1       Theta1
 0.74472529 -0.09192754
Coefficients:
                Value Std.Error   t-value p-value
(Intercept)  0.103026  0.125390  0.821647  0.4121
x.bspl4.1   -4.483239  2.746557 -1.632312  0.1039
```

```
x.bspl4.2   21.400016  2.662887  8.036398  0.0000
x.bspl4.3   -9.649922  3.493478 -2.762268  0.0062
x.bspl4.4   -0.931555  2.543566 -0.366240  0.7145
x.bspl4.5    4.916255  1.699514  2.892742  0.0042
 Correlation:
          (Intr) x.b4.1 x.b4.2 x.b4.3 x.b4.4
x.bspl4.1 -0.001
x.bspl4.2  0.001 -0.824
x.bspl4.3  0.001  0.629 -0.935
x.bspl4.4 -0.002 -0.487  0.815 -0.953
x.bspl4.5  0.003  0.387 -0.676  0.820 -0.941
Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.83145326 -0.68800556 -0.05232443  0.64048656  2.77432368
Residual standard error: 0.7883615
Degrees of freedom: 250 total; 244 residual
```

```
> res7
Call:
list("fregre.basis")
Coefficients:
(Intercept)    x.bspl4.1    x.bspl4.2    x.bspl4.3    x.bspl4.4    x.bspl4.5
     0.1048      -4.1012      20.8042      -8.8890      -1.3945       5.0965
```

```
> summary(res_gsam.pc)
Family: gaussian
Link function: identity
Formula:
y ~ +s(x.PC1, k = -1) + s(x.PC2, k = -1) + s(x.PC3, k = -1)
<environment: 0x000001923e22e710>
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10684    0.04834    2.21    0.028 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
           edf Ref.df     F p-value
s(x.PC1) 1.000  1.000 398.39  <2e-16 ***
s(x.PC2) 3.532  4.459  45.49  <2e-16 ***
s(x.PC3) 1.000  1.000 151.54  <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.751   Deviance explained = 75.6%
GCV = 0.59979  Scale est. = 0.58412    n = 250
```

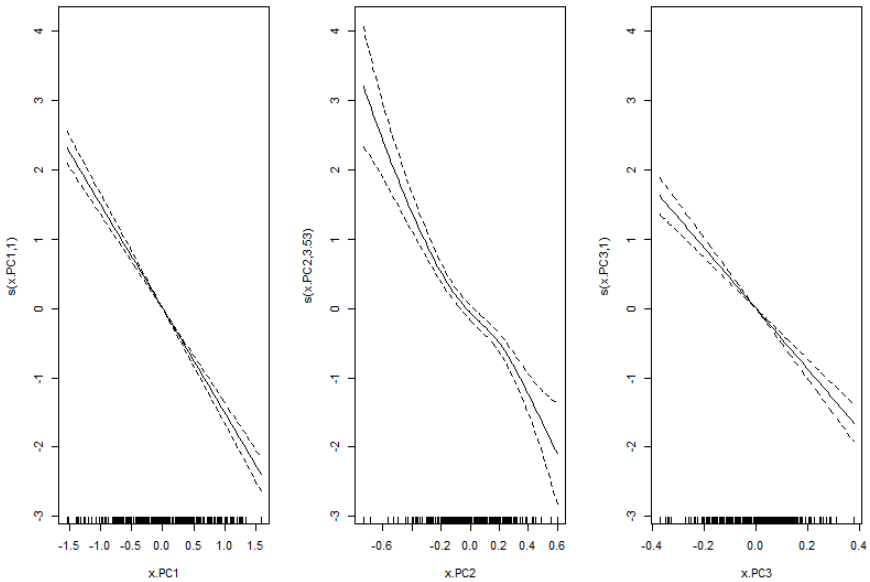


Figura A.3: Comportamiento de las PC del modelo GSAM. ARMA(1,1). $\delta = 1$

```
> summary(res_gsam.bsp)
Family: gaussian
Link function: identity
Formula:
y ~ +s(x.bspl4.1, k = -1) + s(x.bspl4.2, k = -1) + s(x.bspl4.3,
    k = -1) + s(x.bspl4.4, k = -1) + s(x.bspl4.5, k = -1)
<environment: 0x000001923fee9868>
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10684    0.04816   2.219   0.0275 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
              edf Ref.df      F p-value
s(x.bspl4.1) 7.38  8.356  2.310  0.0246 *
s(x.bspl4.2) 1.00  1.000 32.078  <2e-16 ***
s(x.bspl4.3) 1.00  1.000  6.020  0.0149 *
s(x.bspl4.4) 1.00  1.000  0.020  0.8865
s(x.bspl4.5) 1.00  1.000  3.664  0.0568 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.753   Deviance explained = 76.4%
GCV = 0.61004  Scale est. = 0.57983    n = 250
```

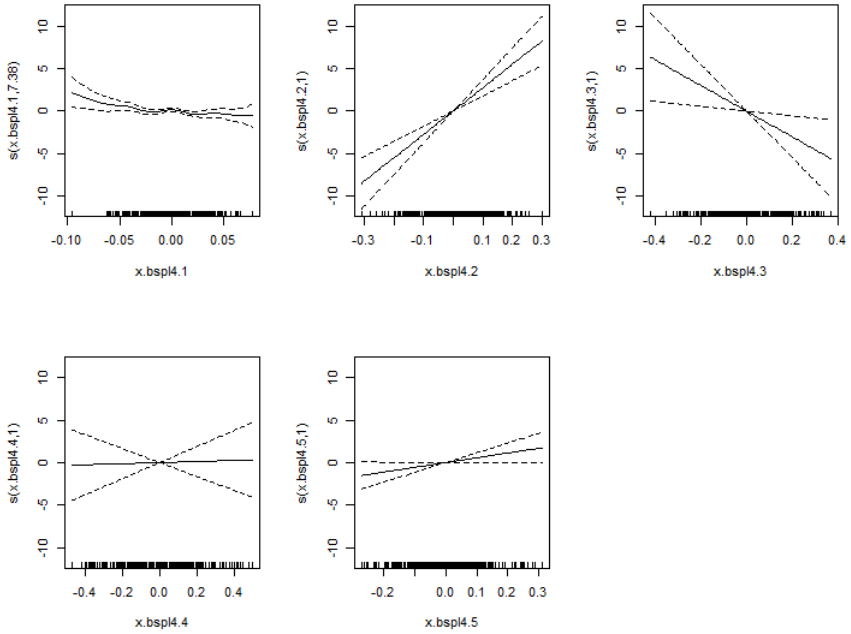


Figura A.4: Comportamiento de las BSP del modelo GSAM. ARMA(1,1). $\delta = 1$

```
> summary(res_gkam)
 *** Summary Functional Data Regression with backfiting algorithm ***
Family: gaussian
Link function: identity
alpha= 0.153   n=  250
Algorithm converged? Yes  Number  of iterations  2
****     ****     ****     ****     ****     ****
         h cor(f(X),eta)  edf
f(x) 0.271              1 13.9
****     ****     ****     ****     ****     ****
edf: Equivalent degrees of freedom
Residual deviance= 231.788  Null deviance= 583.492
AIC=  722.378 Deviance explained= 60.3 %
R-sq.= 0.603  R-sq.(adj)= 0.579
Names of possible influence curves: 46 76 88 113 115 121 122 124 164
```

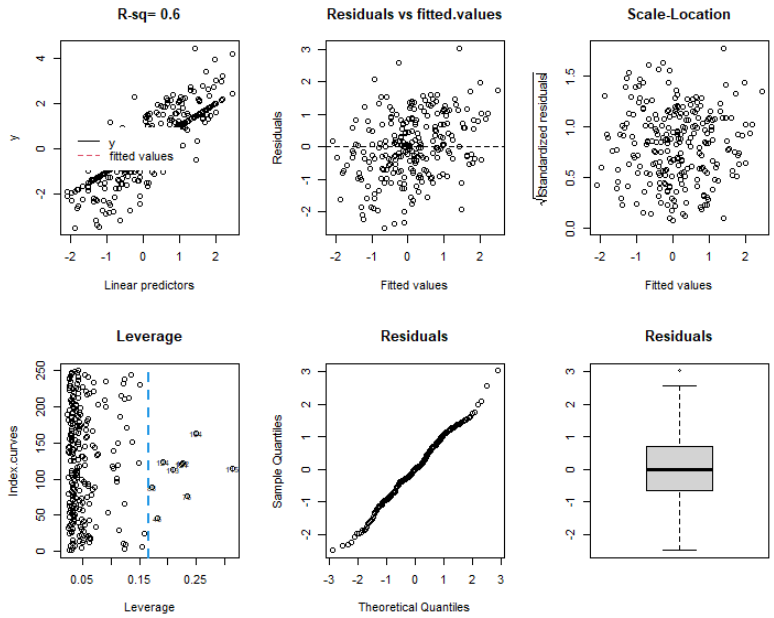


Figura A.5: Resultados resumen del modelo GKAM. ARMA(1,1). $\delta = 1$.

A.2.2. Modelos con ε generado por proceso ARMA(1,1) y $\delta = 0{,}70$

```
summary(res_lm.pc)
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
      Min        1Q    Median        3Q       Max
-1.52229 -0.39049 -0.03436  0.34884  1.61058
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26795    0.03539   7.571 7.45e-13 ***
x.PC1       -1.05782    0.05518 -19.170  < 2e-16 ***
x.PC2       -2.15434    0.16244 -13.263  < 2e-16 ***
x.PC3       -3.02223    0.25713 -11.754  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 0.5596 on 246 degrees of freedom
Multiple R-squared:  0.7352,    Adjusted R-squared:  0.732
F-statistic: 227.7 on 3 and 246 DF,  p-value: < 2.2e-16
```

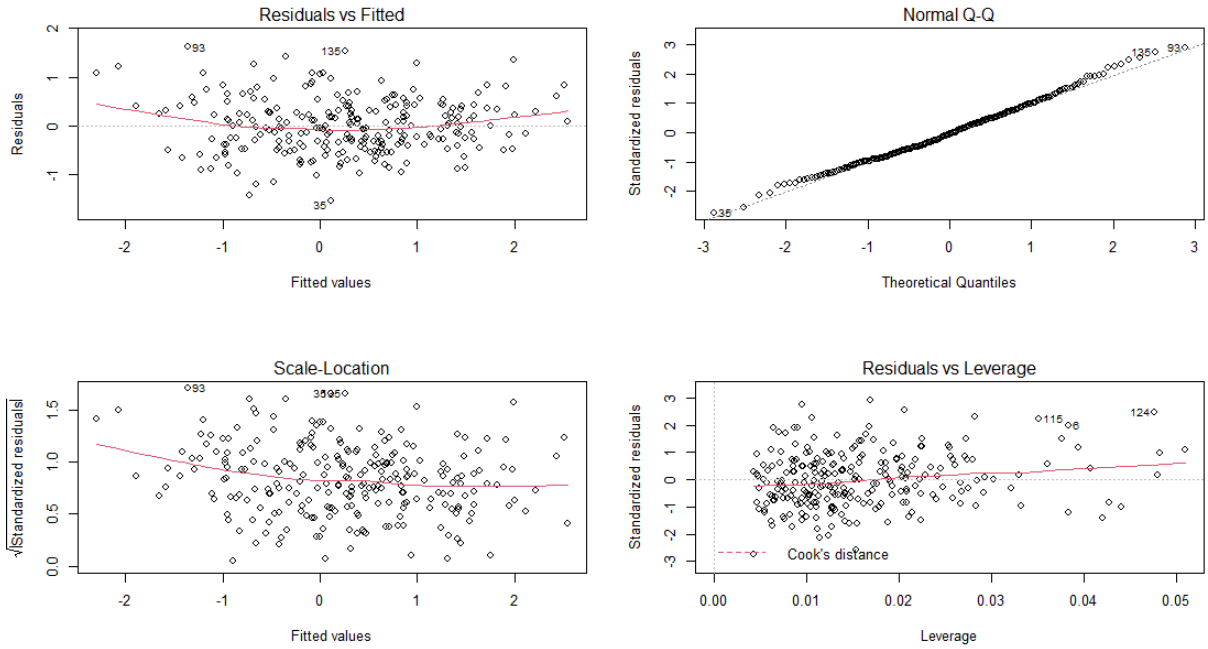


Figura A.6: Resultados resumen del modelo LM representado por PC. ARMA(1,1). $\delta = 0{,}7$.

```
> summary(res_gls.pc)
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  284.0677 308.605 -135.0339
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
      Phi1      Theta1
 0.7655765 -0.1436800
Coefficients:
                  Value  Std.Error    t-value p-value
(Intercept)  0.2655112 0.09305899   2.853149  0.0047
x.PC1       -1.0000839 0.03365820 -29.712940  0.0000
x.PC2       -2.0265291 0.10175543 -19.915685  0.0000
x.PC3       -2.5964833 0.15651318 -16.589550  0.0000
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1  0.000
x.PC2  0.000 -0.031
x.PC3 -0.005  0.067 -0.036
Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.68786767 -0.67977532 -0.03005017  0.68469790  2.75312034
Residual standard error: 0.5671286
Degrees of freedom: 250 total; 246 residual
```

```
> res_igls.pc
Call:
list("fregre.pc")
Coefficients:
(Intercept)        x.PC1        x.PC2        x.PC3
     0.2676      -0.9865      -2.0153      -2.5551
```

```
> summary(res_lm.bsp)
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
      Min        1Q    Median        3Q       Max
-1.45029 -0.38812 -0.02839  0.36325  1.59710
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.26795    0.03542   7.565 7.88e-13 ***
x.bspl4.1    -6.80719    3.42866  -1.985   0.0482 *
x.bspl4.2    19.09941    3.19009   5.987 7.59e-09 ***
x.bspl4.3   -10.10480    4.15009  -2.435   0.0156 *
x.bspl4.4     0.17053    3.02928   0.056   0.9552
x.bspl4.5     3.88806    2.03623   1.909   0.0574 .
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 0.56 on 244 degrees of freedom
Multiple R-squared:  0.7369,    Adjusted R-squared:  0.7316
F-statistic: 136.7 on 5 and 244 DF,  p-value: < 2.2e-16
```

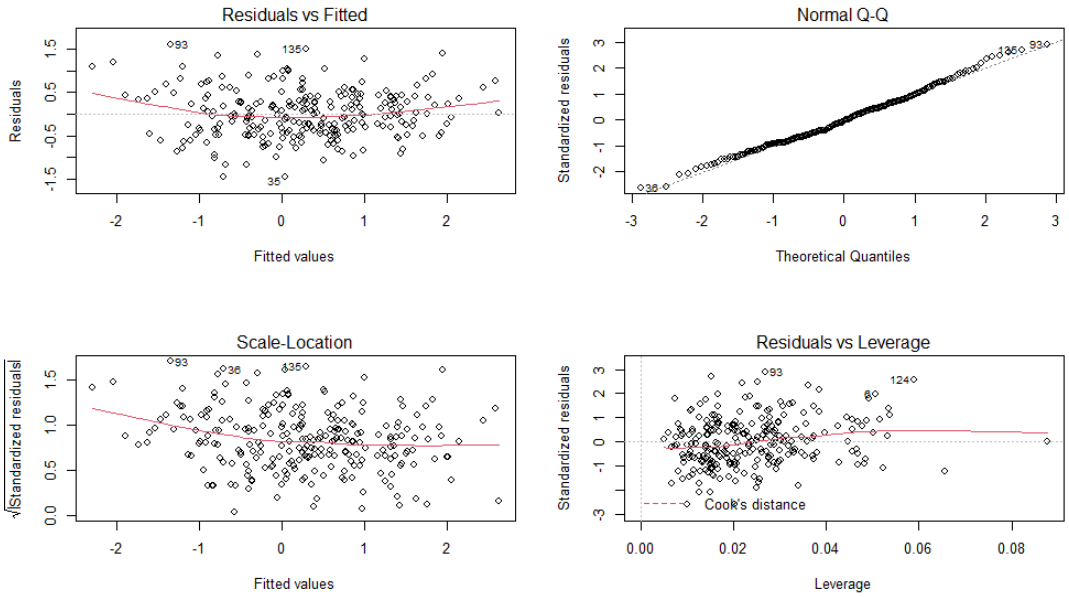


Figura A.7: Resultados resumen del modelo LM representado por BSP. ARMA(1,1). $\delta = 0{,}7$

```
> summary(res_gls.bsp)
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  272.1418 303.6163 -127.0709
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
      Phi1      Theta1
 0.7639815 -0.1403396
Coefficients:
                Value Std.Error    t-value p-value
(Intercept)  0.264803 0.0928292  2.852586  0.0047
x.bspl4.1   -3.313190 2.0166908 -1.642884  0.1017
x.bspl4.2   14.835558 1.9540523  7.592201  0.0000
```

```
x.bspl4.3    -6.397434 2.5616626 -2.497376  0.0132
x.bspl4.4    -1.041967 1.8640198 -0.558989  0.5767
x.bspl4.5     3.817041 1.2452599  3.065256  0.0024
 Correlation:
          (Intr) x.b4.1 x.b4.2 x.b4.3 x.b4.4
x.bspl4.1 -0.001
x.bspl4.2  0.001 -0.824
x.bspl4.3  0.000  0.630 -0.935
x.bspl4.4 -0.002 -0.488  0.815 -0.952
x.bspl4.5  0.003  0.386 -0.675  0.820 -0.941
Standardized residuals:
         Min           Q1          Med           Q3          Max
-2.585530826 -0.695516851 -0.006003346  0.660336106  2.781890278
Residual standard error: 0.5672723
Degrees of freedom: 250 total; 244 residual
```

```
> res_igls.bsp
Call:
list("fregre.basis")

Coefficients:
(Intercept)    x.bspl4.1    x.bspl4.2    x.bspl4.3    x.bspl4.4    x.bspl4.5
      0.267       -2.597       13.884       -5.267       -1.761        4.184
```

```
> summary(res_gsam.pc)
Family: gaussian
Link function: identity
Formula:
y ~ +s(x.PC1, k = -1) + s(x.PC2, k = -1) + s(x.PC3, k = -1)
<environment: 0x000001924c87c9e8>
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26795    0.03414   7.849 1.34e-13 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
           edf Ref.df     F p-value
s(x.PC1) 2.186  2.757 145.00  <2e-16 ***
s(x.PC2) 3.483  4.398  45.92  <2e-16 ***
s(x.PC3) 1.000  1.000 149.94  <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.751   Deviance explained = 75.7%
GCV = 0.30054  Scale est. = 0.29132    n = 250
```

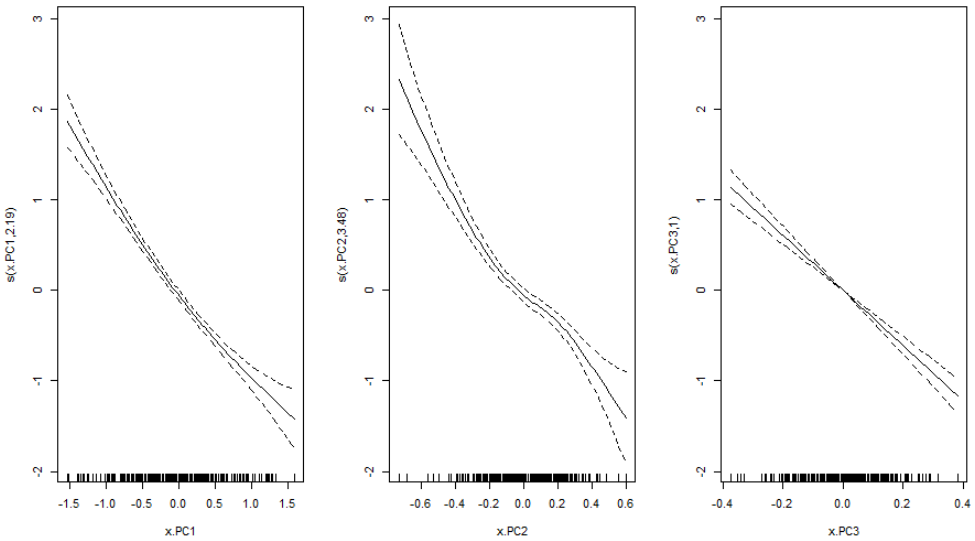


Figura A.8: Comportamiento de las PC del modelo GSAM. ARMA(1,1). $\delta = 0{,}7$

```
> summary(res_gsam.bsp)
Family: gaussian
Link function: identity
Formula:
y ~ +s(x.bspl4.1, k = -1) + s(x.bspl4.2, k = -1) + s(x.bspl4.3,
    k = -1) + s(x.bspl4.4, k = -1) + s(x.bspl4.5, k = -1)
<environment: 0x000001924f046978>
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2680     0.0341   7.858 1.37e-13 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
               edf Ref.df      F p-value
s(x.bspl4.1) 7.326  8.316  2.634  0.0104 *
s(x.bspl4.2) 1.000  1.000 31.087  <2e-16 ***
s(x.bspl4.3) 1.000  1.000  5.618  0.0186 *
s(x.bspl4.4) 1.000  1.000  0.004  0.9504
s(x.bspl4.5) 2.008  2.544  3.049  0.0406 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.751   Deviance explained = 76.3%
GCV = 0.30709  Scale est. = 0.29071    n = 250
```

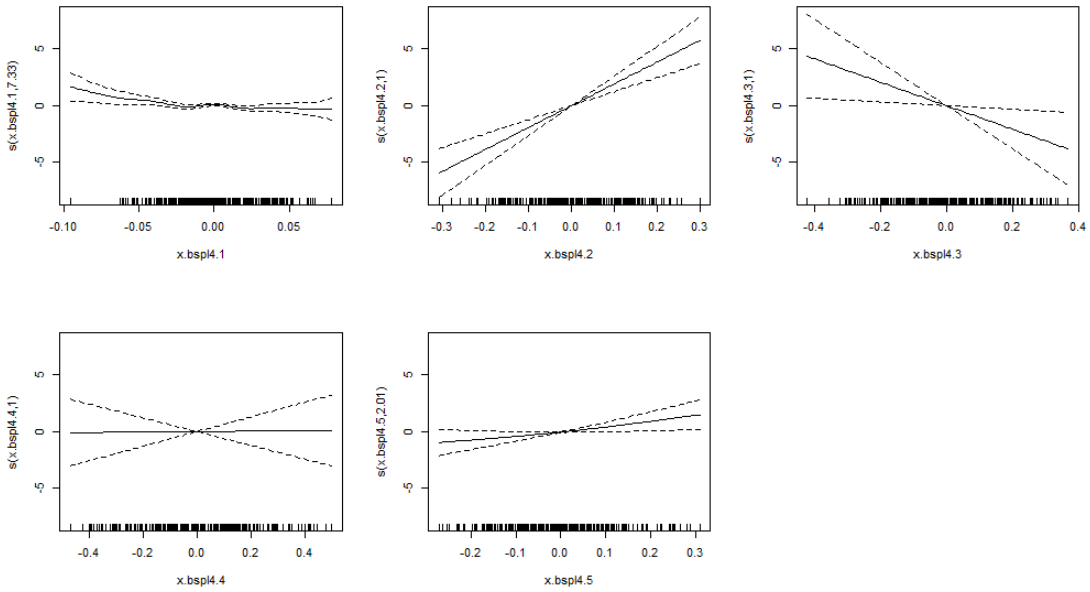


Figura A.9: Comportamiento de las BSP del modelo GSAM. ARMA(1,1). $\delta = 0{,}7$

```
> summary(res_gkam)
 *** Summary Functional Data Regression with backfiting algorithm ***
Family: gaussian
Link function: identity
alpha= 0.321    n=  250
Algorithm converged? Yes  Number  of iterations  2
****     ****     ****     ****     ****     ****
         h cor(f(X),eta)  edf
f(x) 0.271              1 13.9
****     ****     ****     ****     ****     ****
edf: Equivalent degrees of freedom
Residual deviance= 115.008  Null deviance= 290.897
AIC=  547.174 Deviance explained= 60.5 %
R-sq.= 0.605  R-sq.(adj)= 0.581
Names of possible influence curves: 46 76 88 113 115 121 122 124 164
```

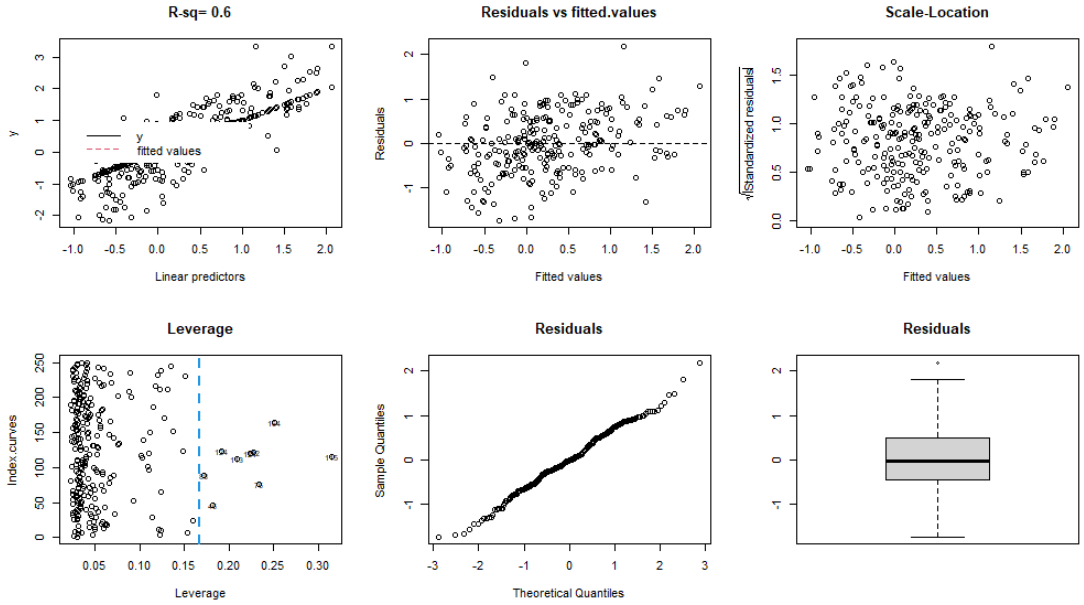


Figura A.10: Resultados resumen del modelo GKAM. ARMA(1,1). $\delta = 0{,}7$

A.2.3. Modelos con ε generado por proceso AR(1) y $\delta = 1$

```
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-3.9760  -1.1529  -0.1884   1.1055   5.3364
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8116     0.1202  -6.752 1.04e-10 ***
x.PC1         1.4030     0.1841   7.621 5.43e-13 ***
x.PC2         2.7253     0.5453   4.998 1.10e-06 ***
x.PC3         3.5679     1.0137   3.520 0.000515 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 1.9 on 246 degrees of freedom
Multiple R-squared:  0.2803,  Adjusted R-squared:  0.2716
F-statistic: 31.94 on 3 and 246 DF,  p-value: < 2.2e-16
```

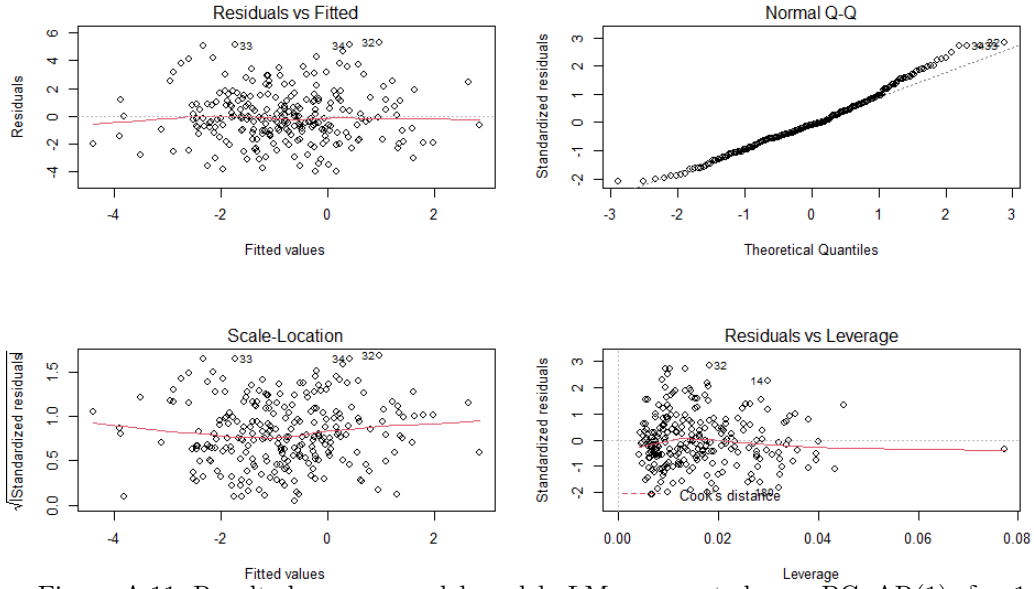


Figura A.11: Resultados resumen del modelo LM representado por PC. AR(1). $\delta = 1$.

```
> summary(res_gls.pc)
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  646.5505 667.5825 -317.2753
Correlation Structure: AR(1)
 Formula: ~ind
 Parameter estimate(s):
      Phi
0.8955645
Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -0.817286 0.5016122 -1.629319  0.1045
x.PC1        1.353891 0.0594211 22.784693  0.0000
x.PC2        2.746543 0.2020782 13.591481  0.0000
x.PC3        4.019182 0.3338892 12.037473  0.0000
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1  0.000
x.PC2  0.004 -0.055
x.PC3  0.003  0.054 -0.014
Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.0586993 -0.5962592 -0.0924313  0.5948318  2.7894436
Residual standard error: 1.924396
Degrees of freedom: 250 total; 246 residual
```

```
> res_igls.pc
Call:
list("fregre.pc")
Coefficients:
(Intercept)        x.PC1        x.PC2        x.PC3
     -0.817        1.354        2.747        4.019
```

```
> res_igls.pcp
Call:
```

```
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
    -0.8171        1.3539        2.7465        4.0190
```

```
> summary(res_lm.bsp)
Call:
lm(formula = pf, data = XX, x = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-4.0104  -1.1619  -0.2357   1.1079   5.5686
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.8116     0.1199  -6.768 9.64e-11 ***
x.bspl4.1    -20.5131    12.3990  -1.654 0.099329 .
x.bspl4.2     37.1886    10.9968   3.382 0.000839 ***
x.bspl4.3    -31.3805    13.5270  -2.320 0.021175 *
x.bspl4.4     15.7275     9.6990   1.622 0.106189
x.bspl4.5     -5.7575     6.7578  -0.852 0.395067
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 1.896 on 244 degrees of freedom
Multiple R-squared:  0.2895,    Adjusted R-squared:  0.275
F-statistic: 19.89 on 5 and 244 DF,  p-value: < 2.2e-16
```

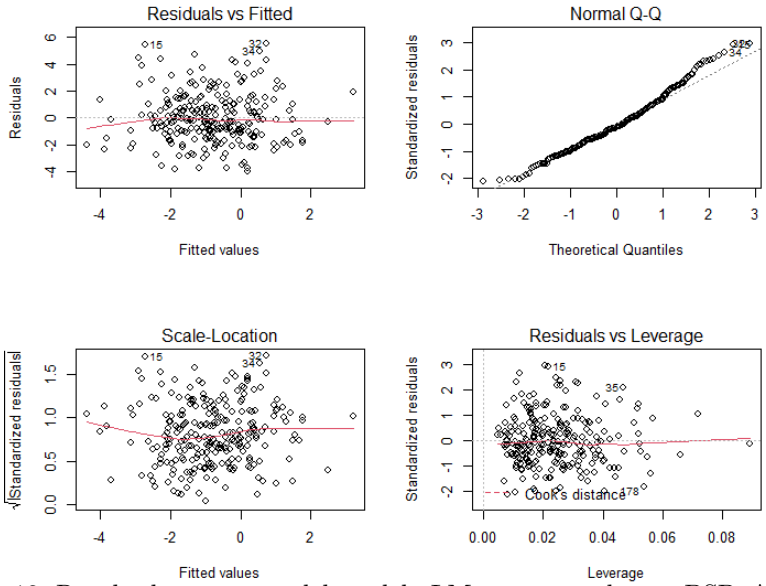


Figura A.12: Resultados resumen del modelo LM representado por BSP. AR(1). $\delta = 1$.

```
> summary(res_gls.bsp)
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC       BIC    logLik
  634.5995  662.5769 -309.2998
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
```

```
0.8940023
Coefficients:
                Value Std.Error   t-value p-value
(Intercept)  -0.821418  0.497190 -1.652119  0.0998
x.bspl4.1    -8.426613  4.118364 -2.046107  0.0418
x.bspl4.2    25.203266  3.741146  6.736776  0.0000
x.bspl4.3   -14.506736  4.706188 -3.082481  0.0023
x.bspl4.4     1.499356  3.454434  0.434038  0.6646
x.bspl4.5     4.330366  2.398588  1.805381  0.0722
 Correlation:
          (Intr) x.b4.1 x.b4.2 x.b4.3 x.b4.4
x.bspl4.1  0.001
x.bspl4.2 -0.001 -0.860
x.bspl4.3  0.003  0.701 -0.937
x.bspl4.4 -0.004 -0.554  0.790 -0.937
x.bspl4.5  0.005  0.426 -0.610  0.766 -0.927
Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.0979775 -0.5960423 -0.1004069  0.5794466  2.8555836
Residual standard error: 1.921405
Degrees of freedom: 250 total; 244 residual
```

```
> res_igls.bsp
Call:
list("fregre.basis")
Coefficients:
(Intercept)    x.bspl4.1    x.bspl4.2    x.bspl4.3    x.bspl4.4    x.bspl4.5
    -0.8209      -8.4348      25.2118     -14.5190       1.5097       4.3230
```

```
> res_igls.bspp
Call:
list("fregre.basis")
Coefficients:
(Intercept)    x.bspl4.1    x.bspl4.2    x.bspl4.3    x.bspl4.4    x.bspl4.5
    -0.8212      -8.4294      25.2062     -14.5109       1.5029       4.3279
```

```
> summary(res_gsam.pc)
Family: gaussian
Link function: identity
Formula:
y ~ +s(x.PC1, k = -1) + s(x.PC2, k = -1) + s(x.PC3, k = -1)
<environment: 0x0000019249a91c38>
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8116     0.1202  -6.752 1.04e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
          edf Ref.df     F  p-value
s(x.PC1)    1      1 58.09  < 2e-16 ***
s(x.PC2)    1      1 24.98 1.31e-06 ***
s(x.PC3)    1      1 12.39 0.000515 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.272   Deviance explained =   28%
GCV = 3.6704  Scale est. = 3.6116     n = 250
```

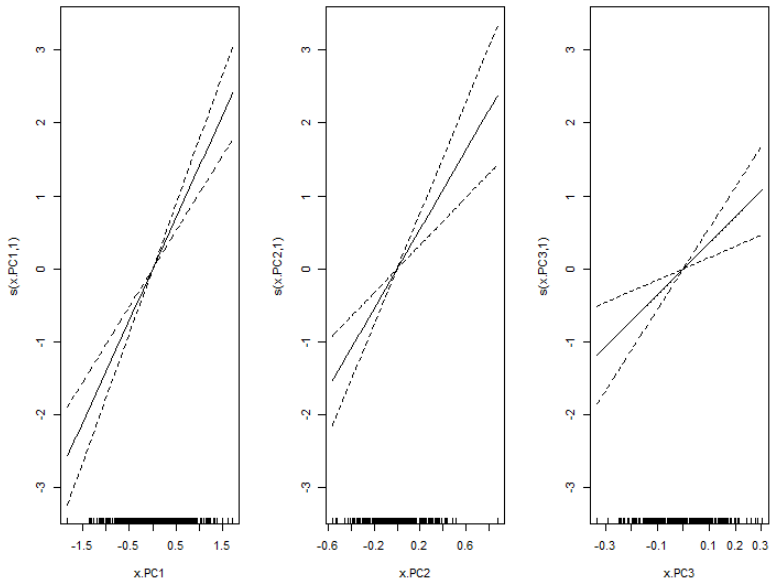


Figura A.13: Comportamiento de las PC del modelo GSAM. AR(1). $\delta = 1$.

```
> summary(res_gsam.bsp)
Family: gaussian
Link function: identity
Formula:
y ~ +s(x.bspl4.1, k = -1) + s(x.bspl4.2, k = -1) + s(x.bspl4.3,
    k = -1) + s(x.bspl4.4, k = -1) + s(x.bspl4.5, k = -1)
<environment: 0x000001925077b6b8>
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8116     0.1165  -6.967 3.35e-11 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
               edf Ref.df     F p-value
s(x.bspl4.1) 7.014  8.023 1.445 0.17362
s(x.bspl4.2) 1.000  1.000 9.927 0.00184 **
s(x.bspl4.3) 1.000  1.000 4.511 0.03473 *
s(x.bspl4.4) 2.403  3.219 1.236 0.28778
s(x.bspl4.5) 6.345  7.448 1.367 0.22581
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.316   Deviance explained = 36.5%
GCV =  3.668   Scale est. = 3.3927     n = 250
```

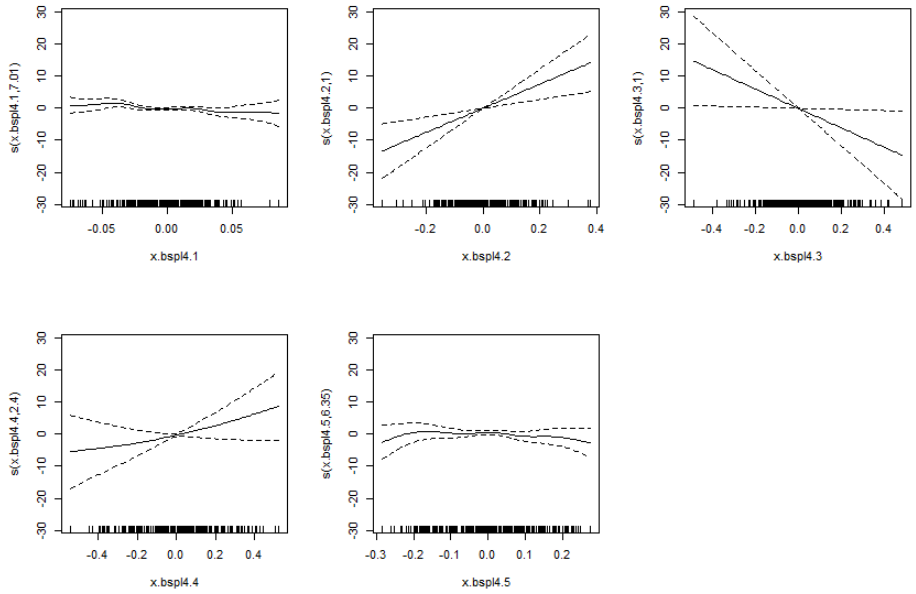


Figura A.14: Comportamiento de las BSP del modelo GSAM. AR(1). $\delta = 1$.

```
*** Summary Functional Data Regression with backfiting algorithm ***
Family: gaussian
Link function: identity
alpha= -0.831   n=  250
Algorithm converged? Yes  Number  of iterations  2
****     ****     ****     ****     ****     ****
         h cor(f(X),eta) edf
f(x) 0.268              1  15
****     ****     ****     ****     ****     ****
edf: Equivalent degrees of freedom
Residual deviance= 884.206  Null deviance= 1234.573
AIC=  1059.25 Deviance explained= 28.4 %
R-sq.= 0.284  R-sq.(adj)= 0.238
Names of possible influence curves: 5 17 21 84 86 149 220
```

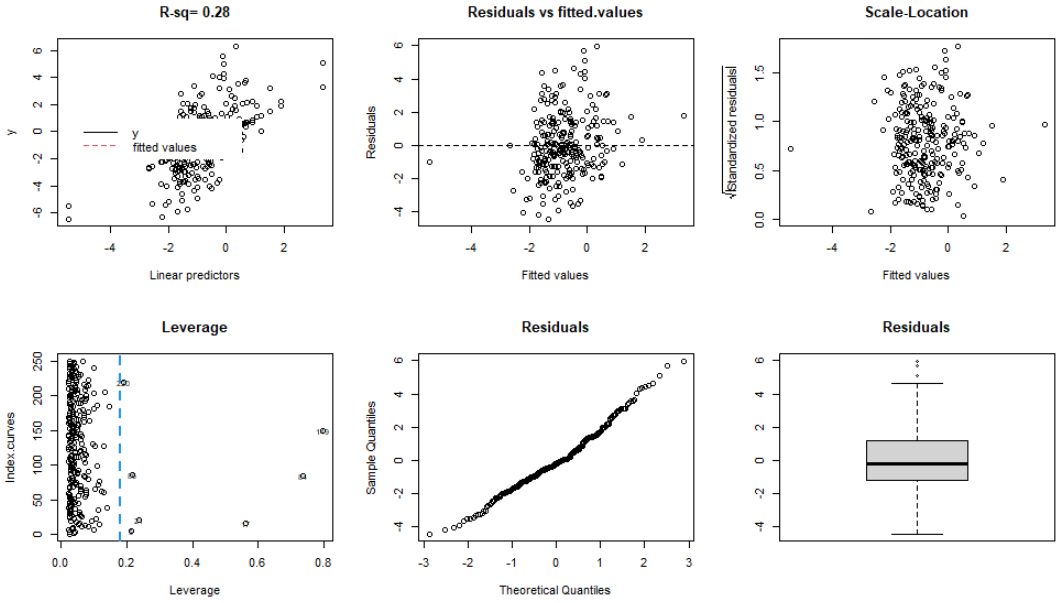


Figura A.15: Resultados resumen del modelo GKAM. AR(1). $\delta = 1$.

A.3. Gráficos

A.3.1. Comparativo de los β estimados

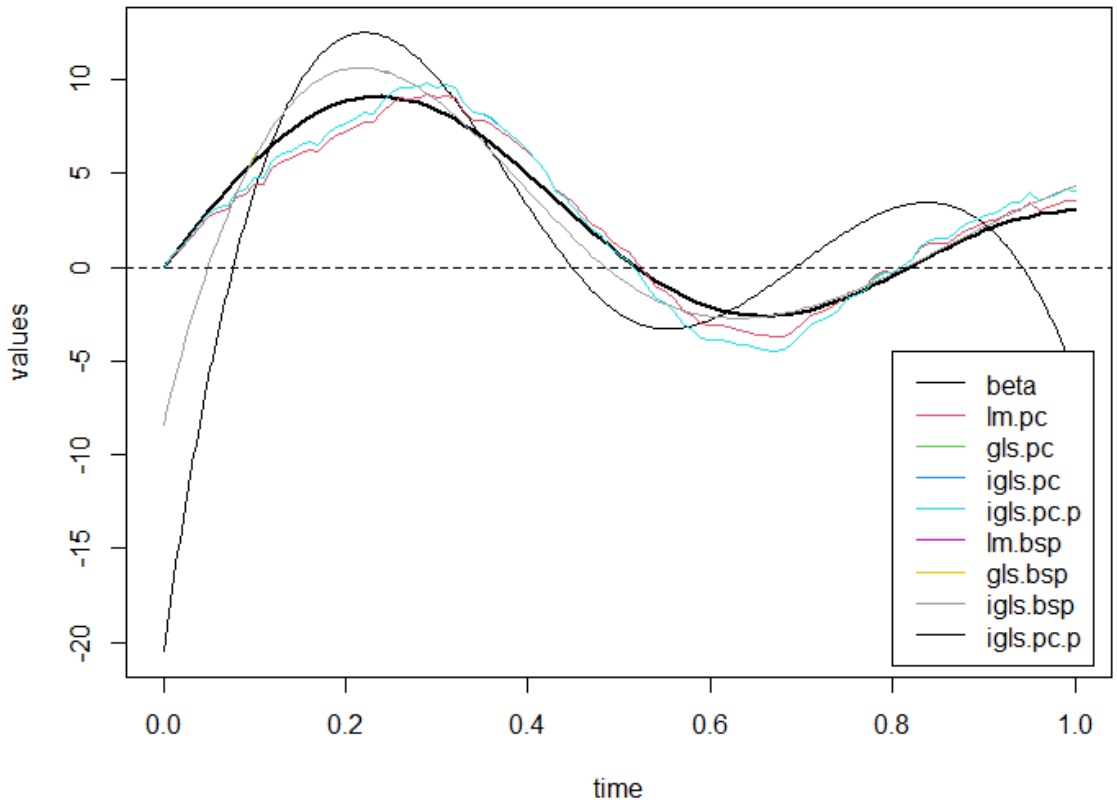


Figura A.16: Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. AR(1). $\delta = 1$.

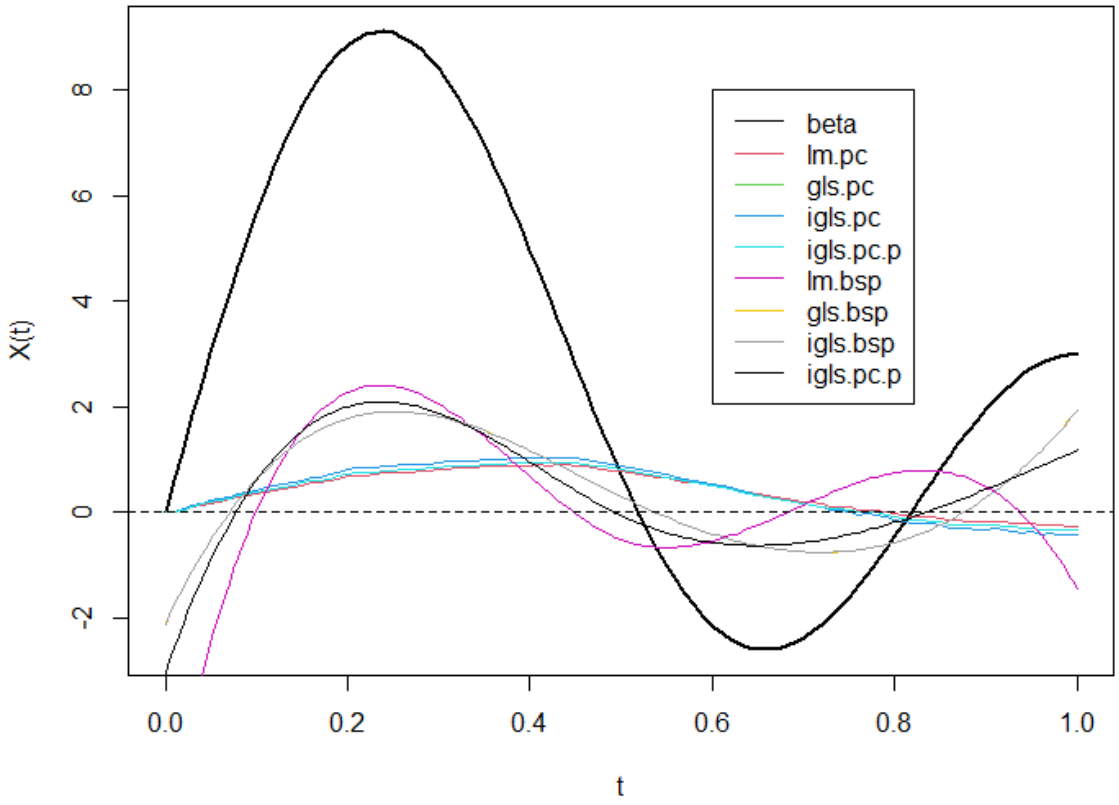


Figura A.17: Comparativo del β vs ajuste $\hat{\beta}$ de los modelos LM, GLS e iGLS. AR(1). $\delta = 0{,}15$.

A.3.2. Ajuste de modelo

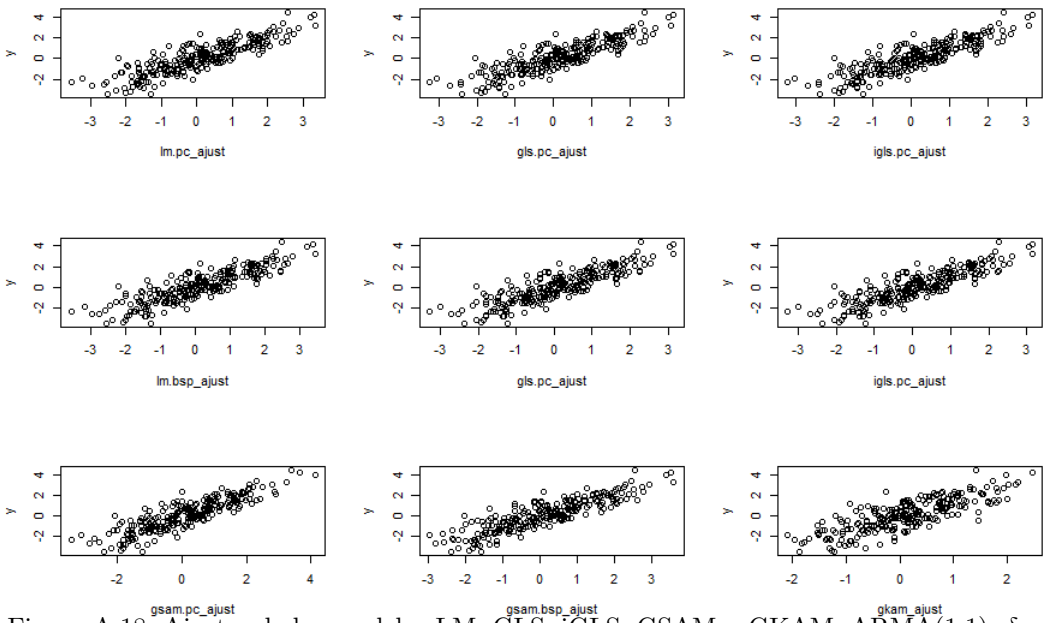


Figura A.18: Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. ARMA(1,1). $\delta = 1$.

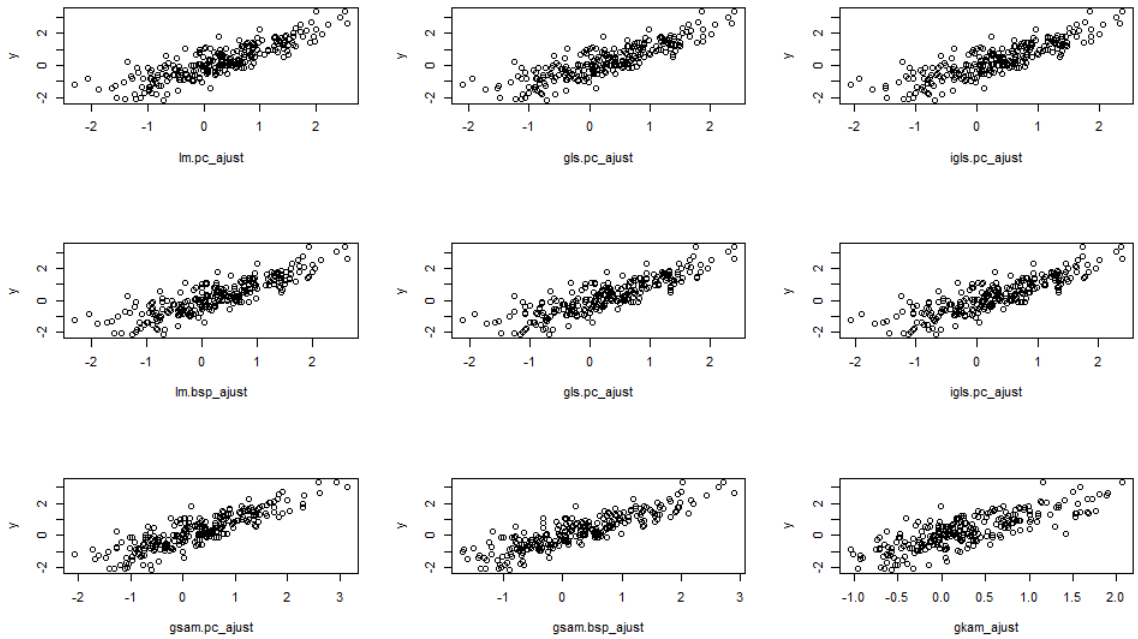


Figura A.19: Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. ARMA(1,1). $\delta = 0{,}7$.

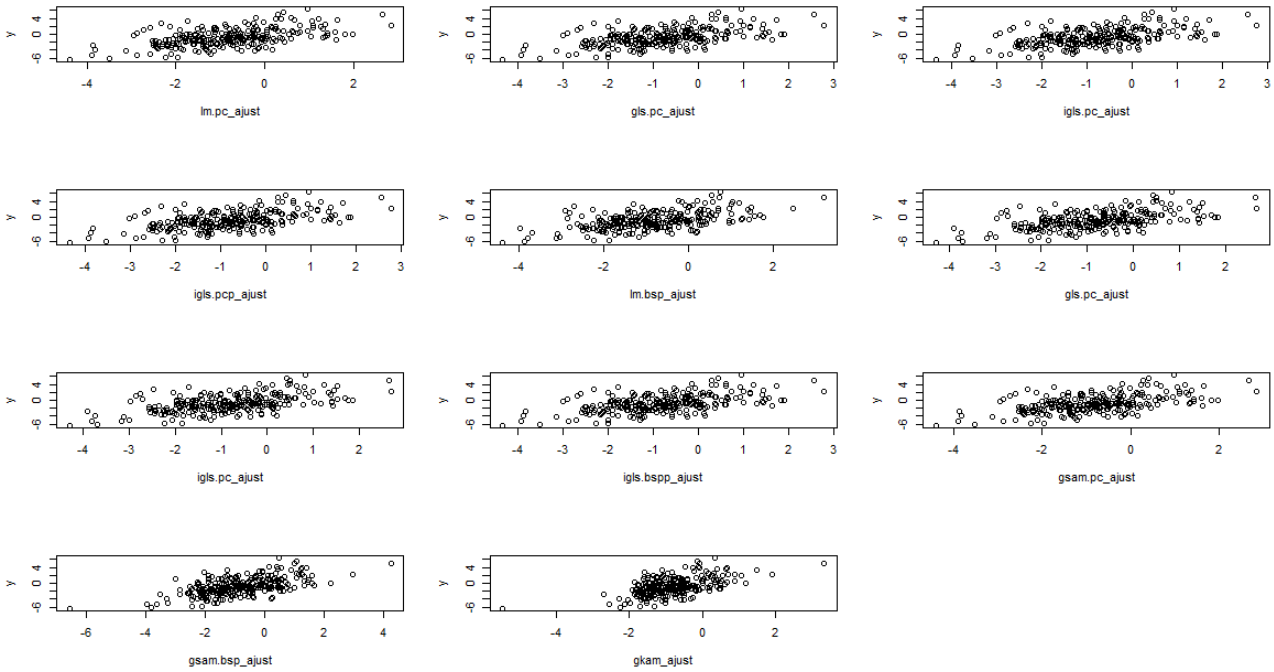


Figura A.20: Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. AR(1). $\delta = 1$.

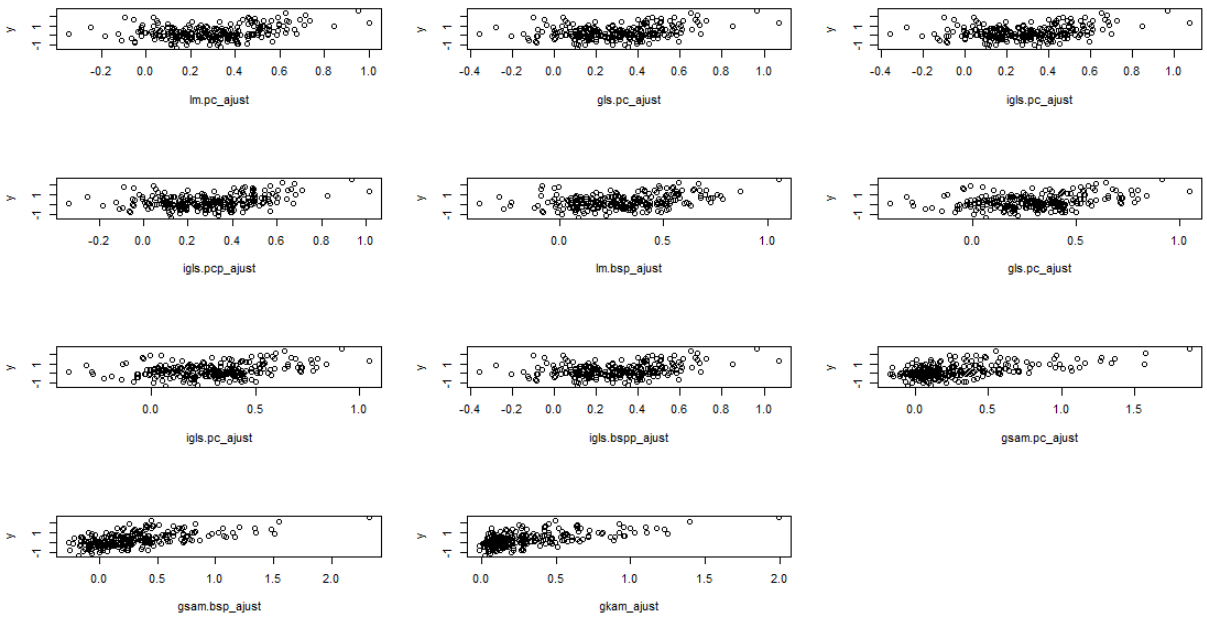


Figura A.21: Ajustes de los modelos LM, GLS, iGLS, GSAM y GKAM. AR(1). $\delta = 0{,}15$.

A.3.3. Predicciones

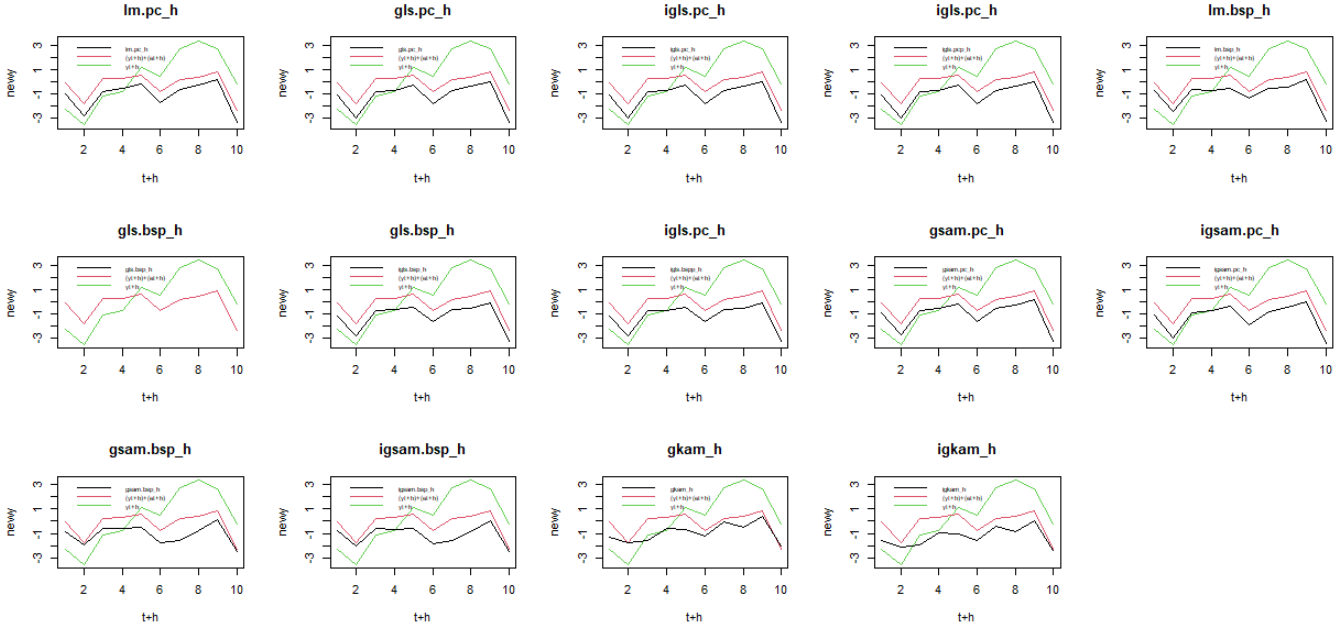


Figura A.22: Predicciones de los modelos a horizonte $h = 10$. AR(1). $\delta = 1$.

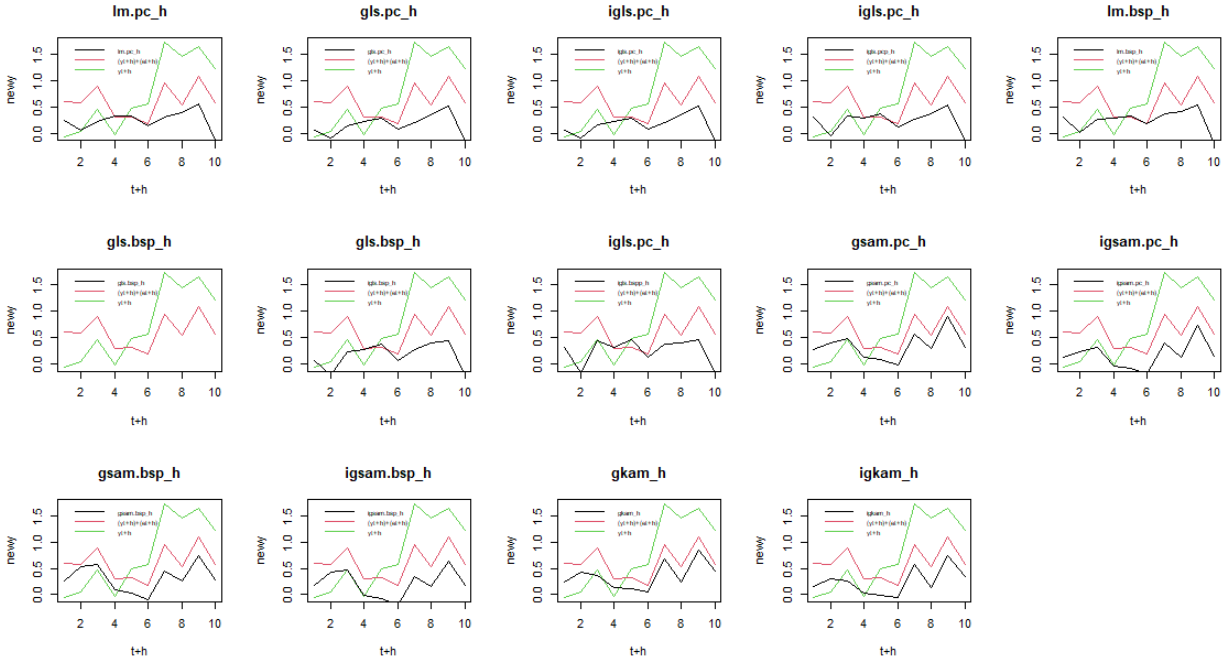


Figura A.23: Predicciones de los modelos a horizonte $h = 10$. AR(1). $\delta = 0{,}15$.

Apéndice B

Resultados de la aplicación

B.1. Resultados de los modelos FRM para estimar la incidencia acumulada de contagio del Covid-19.

En esta sección del apéndice se encuentran los resultados de la salidas y gráficos de manera parcial para las CCAA de Cataluña, Galicia y Madrid.

B.1.1. Salida de modelos para la comunidad de Cataluña.

Resultados del modelo FLM.

```
summary(mod.lin.c[[1]])
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-44.803   -4.997   -0.661    4.023   46.376
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.831e+02  4.950e-01   571.9   <2e-16 ***
x.PC1       3.629e-01  2.609e-03   139.1   <2e-16 ***
x.PC2       5.362e-01  3.272e-03   163.9   <2e-16 ***
x.PC3       8.605e-01  2.030e-02    42.4   <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 10.31 on 430 degrees of freedom
Multiple R-squared:  0.9975,   Adjusted R-squared:  0.9975
F-statistic: 5.683e+04 on 3 and 430 DF,  p-value: < 2.2e-16
```

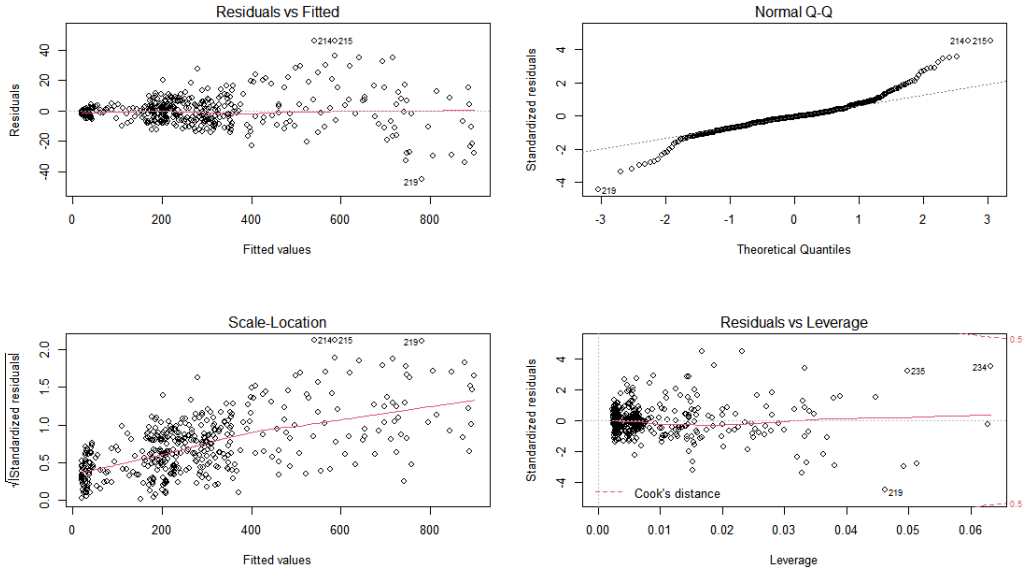


Figura B.1: Resultados gráficos del modelo FLM para la comunidad de Cataluña. Función *fregre.lm*.

Resultados del modelo SAM.

```
summary(mod.sam.c[[1]])
Family: gaussian
Link function: identity
Formula:
H1 ~ +s(x.PC1, k = -1) + s(x.PC2, k = -1) + s(x.PC3, k = -1)
<environment: 0x0000022bb049a458>
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 283.1037     0.4681   604.8   <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
           edf Ref.df       F p-value
s(x.PC1) 7.730  8.502 1163.12  <2e-16 ***
s(x.PC2) 5.907  7.008 3805.06  <2e-16 ***
s(x.PC3) 8.668  8.925   91.23  <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.998   Deviance explained = 99.8%
GCV = 100.49  Scale est. = 95.094     n = 434
```

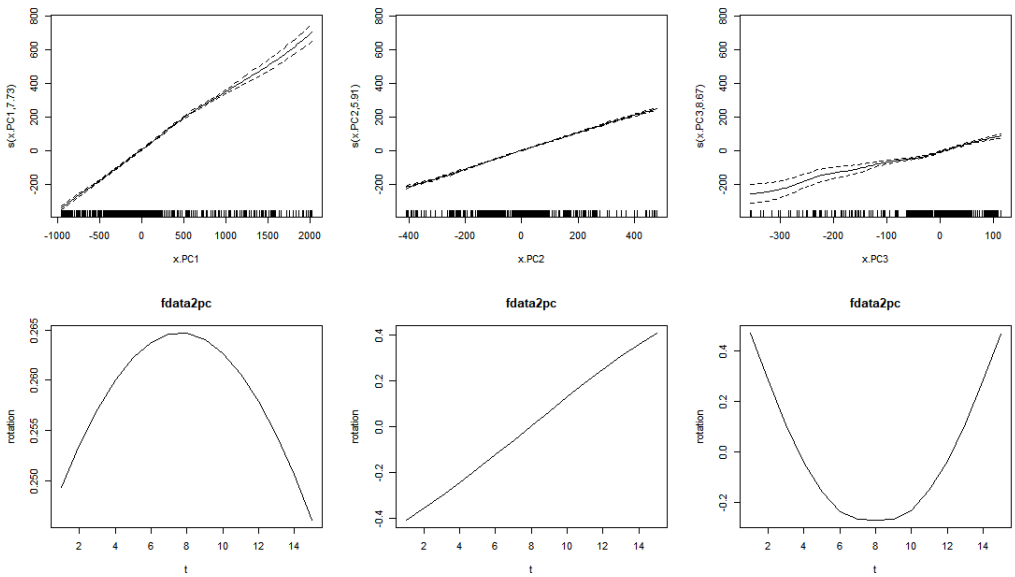


Figura B.2: Resultados gráficos del modelo SAM para la comunidad de Cataluña. Función *fregre.gsam*.

Resultados del modelo GLS.

- Modelo GLS-AR(1)

```
summary(mod.gls.c[[1]])
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  3073.576 3097.959 -1530.788
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.687933
Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 283.08816 1.2348851 229.24252       0
x.PC1         0.34538 0.0057248  60.33087       0
x.PC2         0.52826 0.0078774  67.06044       0
x.PC3         0.72094 0.0436711  16.50847       0
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1 -0.004
x.PC2 -0.004 -0.012
x.PC3 -0.007  0.955 -0.013
Standardized residuals:
       Min         Q1        Med         Q3        Max
-3.5645666 -0.4722616 -0.1274158  0.3538167  5.0576414
Residual standard error: 11.11705
Degrees of freedom: 434 total; 430 residual
```

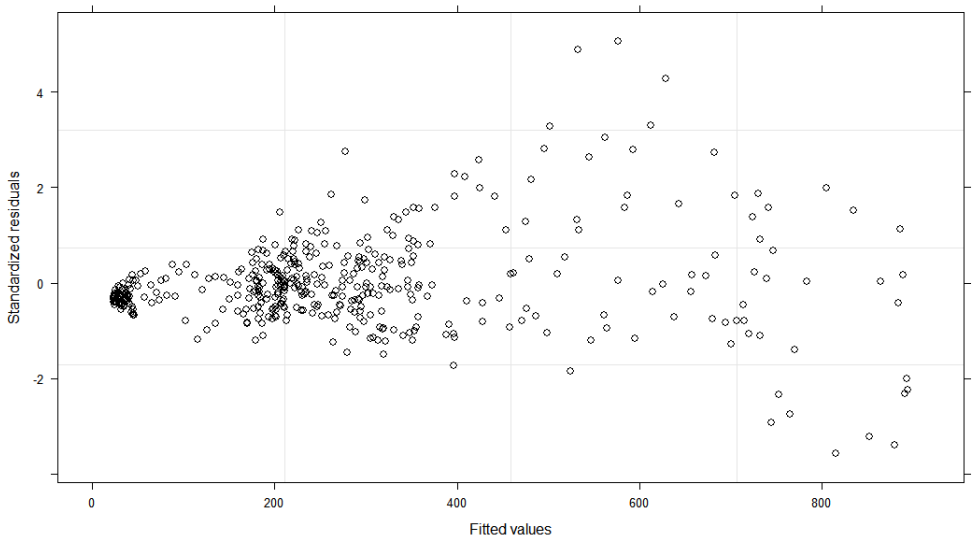


Figura B.3: Resultados gráficos del modelo GLS-AR(1) para la comunidad de Cataluña. Función *fregre.gls*.

- Modelo GLS-ARMA(1,1)

```
summary(mod.gls.1.1.c[[1]])
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  2991.37 3019.816 -1488.685
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
     Phi1     Theta1
0.2933865 0.6843714
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 283.09968 0.8331398 339.7985       0
x.PC1         0.35779 0.0043021  83.1667       0
x.PC2         0.53470 0.0054823  97.5313       0
x.PC3         0.81935 0.0333632  24.5586       0
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1 -0.002
x.PC2 -0.001 -0.011
x.PC3 -0.003  0.964 -0.010
Standardized residuals:
        Min          Q1         Med          Q3         Max
-3.92812295 -0.46729815 -0.09033135  0.35505121  4.68801221
Residual standard error: 10.43474
Degrees of freedom: 434 total; 430 residual
```

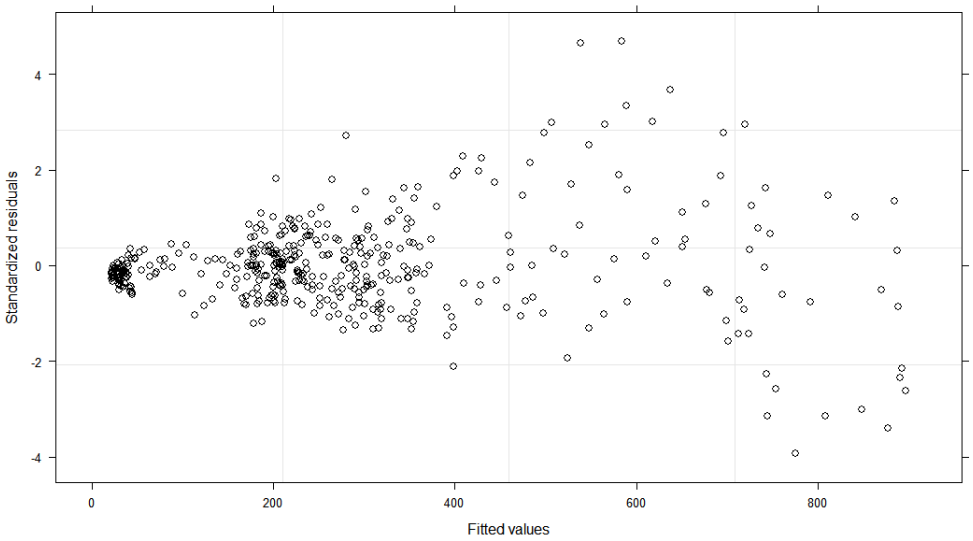


Figura B.4: Resultados gráficos del modelo GLS-ARMA(1,1) para la comunidad de Cataluña. Función *fregre.gls*.

Resultados del modelo iGLS.

- Modelo iGLS-AR(1)

```
mod.igls.1.c[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   283.0853        0.3492        0.5303        0.7514
> mod.igls.1.c[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
          ar1
       0.6451
s.e.   0.0365
sigma^2 estimated as 64.52:  log likelihood = -1520.33,  aic = 3044.66
```

- Modelo iGLS-AR(P)

```
mod.igls.p.c[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   283.1018        0.3619        0.5363        0.8520
> mod.igls.p.c[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(aa$order, 0, 0), include.mean = FALSE, transform.pars =
    TRUE,
    method = "CSS")
Coefficients:
          ar1       ar2
```

```
      0.8548  -0.3907
s.e.  0.0441   0.0441
sigma^2 estimated as 55.55:  part log likelihood = -1487.56
```

- Modelo iGLS-ARMA(1,1)

```
mod.igls.1.1.c[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)        x.PC1        x.PC2        x.PC3
   285.0376       0.3676       0.5209       0.8968
> mod.igls.1.1.c[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
         ar1     ma1
      0.9342  0.3657
s.e.  0.0176  0.0634
sigma^2 estimated as 63.4:  log likelihood = -1517.65,  aic = 3041.3
```

B.1.2. Salida de modelos para la comunidad de Galicia.

Resultados FLM

```
summary(mod.lin.g[[1]])
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-25.1989  -2.4984  -0.3948   1.9365  24.8955
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.638e+02  3.094e-01  529.51   <2e-16 ***
x.PC1       3.644e-01  1.626e-03  224.16   <2e-16 ***
x.PC2       5.342e-01  2.231e-03  239.51   <2e-16 ***
x.PC3       8.890e-01  1.271e-02   69.93   <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 6.446 on 430 degrees of freedom
Multiple R-squared:  0.9988,  Adjusted R-squared:  0.9988
F-statistic: 1.164e+05 on 3 and 430 DF,  p-value: < 2.2e-16
```

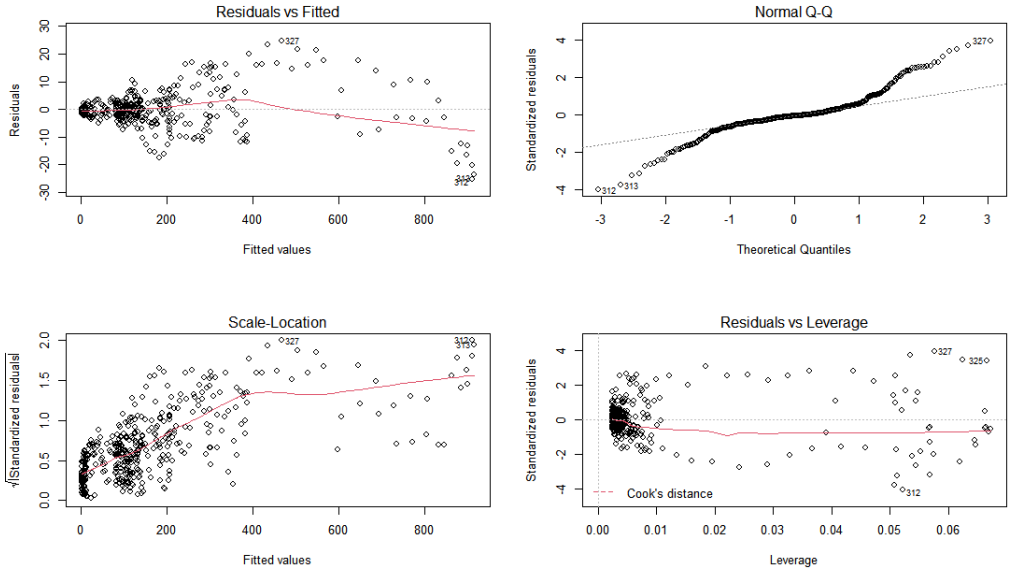


Figura B.5: Resultados gráficos del modelo FLM para la comunidad de Galicia. Función *fregre.lm*.

Resultados del modelo SAM.

```
summary(mod.sam.g[[1]])
Family: gaussian
Link function: identity
Formula:
H1 ~ +s(x.PC1, k = -1) + s(x.PC2, k = -1) + s(x.PC3, k = -1)
<environment: 0x0000022bad3745d0>
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 163.8364     0.2318   706.7   <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
           edf Ref.df      F p-value
s(x.PC1) 8.991  8.998 2153.5  <2e-16 ***
s(x.PC2) 9.000  9.000 9805.7  <2e-16 ***
s(x.PC3) 8.374  8.896  200.8  <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.999   Deviance explained = 99.9%
GCV = 24.897  Scale est. = 23.328     n = 434
```

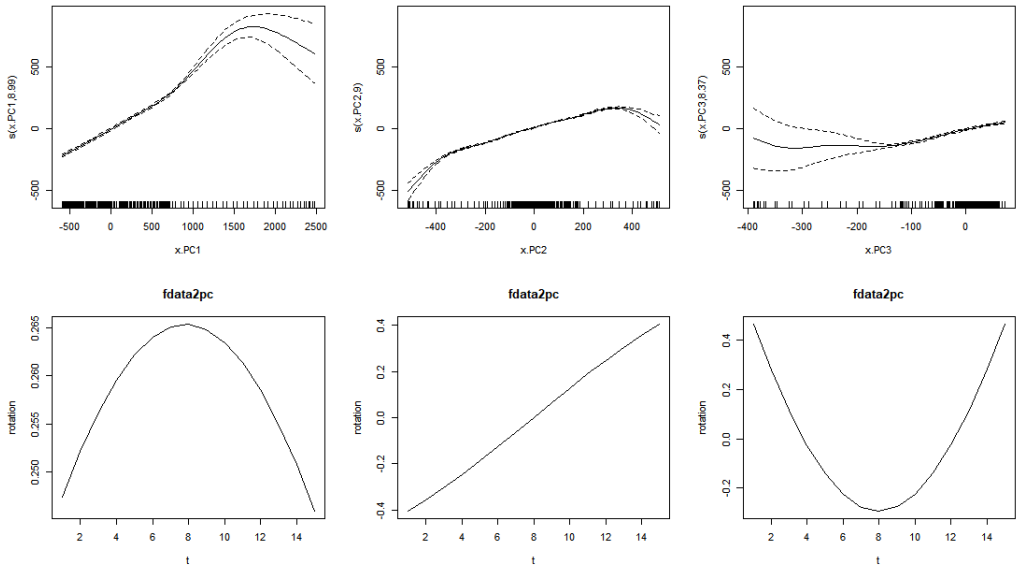


Figura B.6: Resultados gráficos del modelo SAM para la comunidad de Galicia. Función *fregre.gsam*.

Resultados del modelo GLS

- GLS-AR(1)

```
summary(mod.gls.g[[1]])
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  2216.718 2241.101 -1102.359
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.9785129
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 163.20131  6.129887 26.62387       0
x.PC1         0.32377  0.008073 40.10508       0
x.PC2         0.49117  0.009523 51.57630       0
x.PC3         0.66471  0.040243 16.51753       0
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1  0.031
x.PC2 -0.012  0.007
x.PC3  0.015  0.877  0.005
Standardized residuals:
       Min         Q1        Med         Q3        Max
-0.8828536 -0.5726020 -0.1957088  0.2449665  4.2802594
Residual standard error: 14.62506
Degrees of freedom: 434 total; 430 residual
```

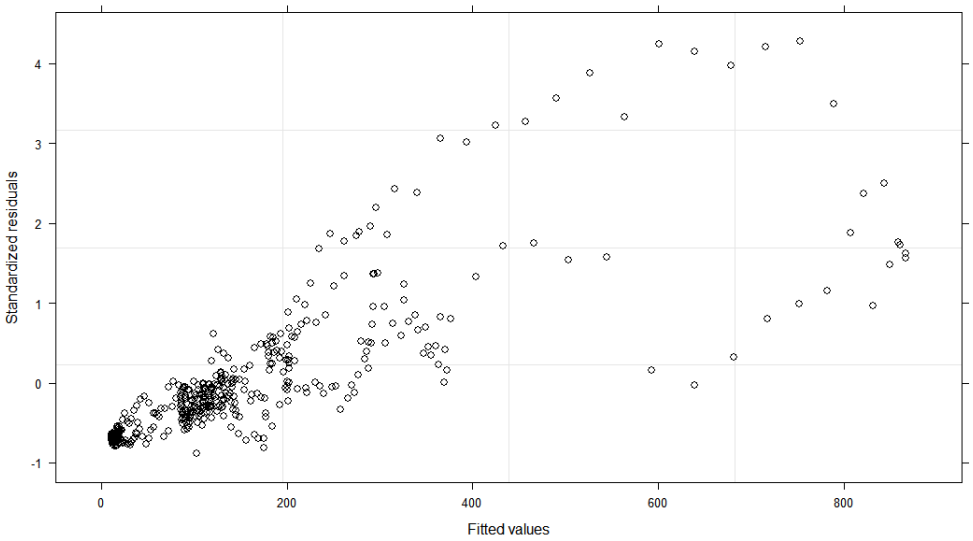


Figura B.7: Resultados gráficos del modelo GLS-AR(1) para la comunidad de Galicia. Función *fregre.gls*.

- GLS-ARMA(1,1)

```
summary(mod.gls.1.1.g[[1]])
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  2215.67 2244.116 -1100.835
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
      Phi1     Theta1
0.97508327 0.08460939
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 163.31378  5.784913 28.23099       0
x.PC1         0.32261  0.008493 37.98502       0
x.PC2         0.49096  0.010233 47.97849       0
x.PC3         0.65479  0.042888 15.26757       0
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1  0.030
x.PC2 -0.012  0.007
x.PC3  0.014  0.876  0.004
Standardized residuals:
       Min         Q1        Med         Q3        Max
-0.8942805 -0.5789641 -0.2053262  0.2440050  4.3056577
Residual standard error: 14.6693
Degrees of freedom: 434 total; 430 residual
```

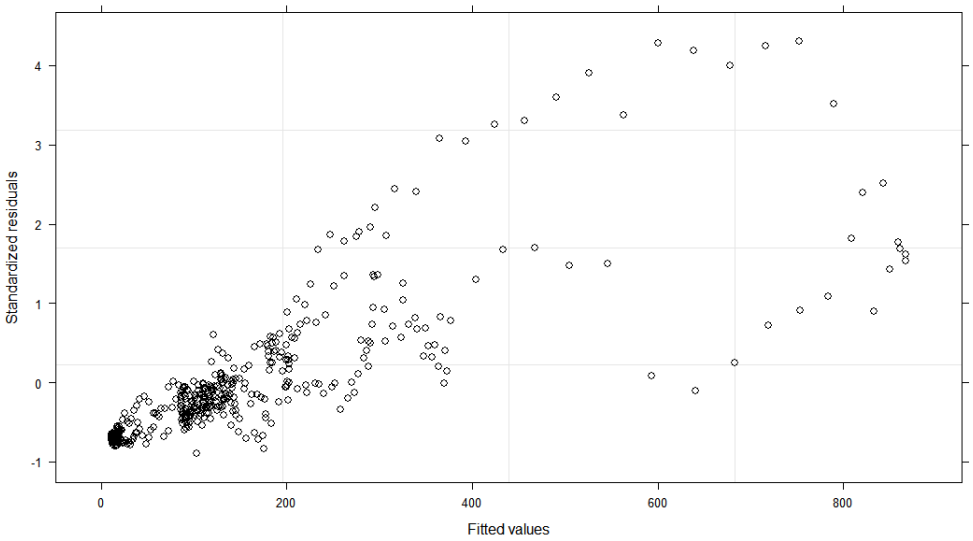


Figura B.8: Resultados gráficos del modelo GLS-ARMA(1,1) para la comunidad de Galicia. Función *fregre.gls*.

Resultados del modelo iGLS

- iGLS-AR(1)

```
mod.igls.1.g[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   163.7599        0.3417        0.5041        0.7342
> mod.igls.1.g[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
          ar1
       0.9279
s.e.   0.0174
sigma^2 estimated as 9.077:  log likelihood = -1095.44,  aic = 2194.89
```

- iGLS-AR(P)

```
mod.igls.p.g[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   163.7599        0.3417        0.5041        0.7342
> mod.igls.p.g[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
          ar1
       0.9279
s.e.   0.0174
```

```
sigma^2 estimated as 9.077:  log likelihood = -1095.44,  aic = 2194.89
```

- iGLS-ARMA(1,1)

```
mod.igls.1.1.g[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   164.2567        0.3314        0.4944        0.6958
> mod.igls.1.1.g[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
          ar1     ma1
      0.9573  0.0821
s.e.  0.0137  0.0485
sigma^2 estimated as 8.95:  log likelihood = -1092.74,  aic = 2191.48
```

Salida de modelos para la comunidad de Madrid.

Resultados FLM

```
summary(mod.lin.m[[1]])
Call:
lm(formula = pf, data = XX, x = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-63.326  -6.423  -1.358   5.355  90.802
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.755e+02  7.455e-01  503.67   <2e-16 ***
x.PC1       3.643e-01  3.105e-03  117.33   <2e-16 ***
x.PC2       5.287e-01  3.832e-03  137.95   <2e-16 ***
x.PC3       8.329e-01  2.409e-02   34.58   <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 15.53 on 430 degrees of freedom
Multiple R-squared:  0.9971,  Adjusted R-squared:  0.997
F-statistic: 4.845e+04 on 3 and 430 DF,  p-value: < 2.2e-16
```

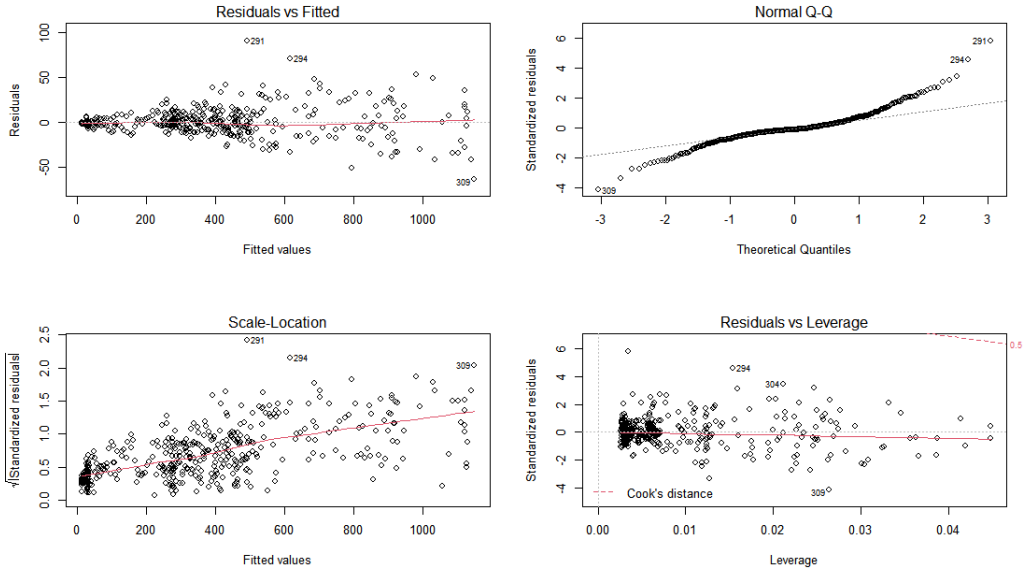


Figura B.9: Resultados gráficos del modelo FLM para la comunidad de Madrid. Función *fregre.lm*.

Resultados del modelo SAM.

```
summary(mod.sam.m[[1]])
Family: gaussian
Link function: identity
Formula:
H1 ~ +s(x.PC1, k = -1) + s(x.PC2, k = -1) + s(x.PC3, k = -1)
<environment: 0x0000022baff69db0>
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 375.4631     0.6809   551.5   <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Approximate significance of smooth terms:
           edf Ref.df       F p-value
s(x.PC1) 8.011  8.734 1060.42  <2e-16 ***
s(x.PC2) 7.837  8.638 2487.86  <2e-16 ***
s(x.PC3) 8.356  8.871   96.33  <2e-16 ***
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
R-sq.(adj) =  0.998   Deviance explained = 99.8%
GCV = 213.59  Scale est. = 201.19     n = 434
```

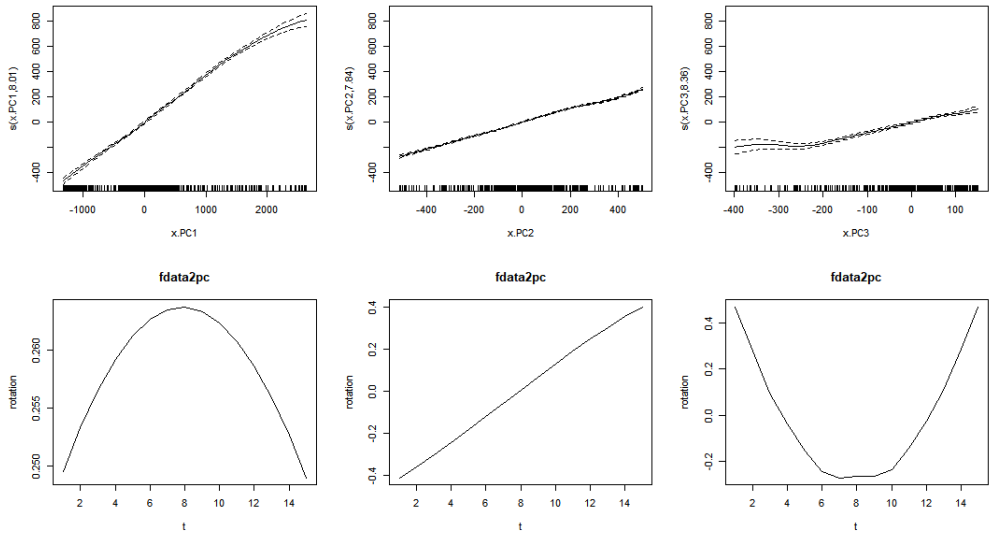


Figura B.10: Resultados gráficos del modelo SAM para la comunidad de Madrid. Función *fregre.gsam*.

Resultados del modelo GLS

- GLS-AR(1)

```
summary(mod.gls.m[[1]])
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  3490.856 3515.238 -1739.428
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
     Phi
0.59573
Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 375.4171 1.5450496 242.98063       0
x.PC1         0.3485 0.0058753  59.31880       0
x.PC2         0.5226 0.0077575  67.37232       0
x.PC3         0.7071 0.0451393  15.66570       0
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1 -0.002
x.PC2 -0.006 -0.031
x.PC3 -0.003  0.966 -0.031
Standardized residuals:
       Min         Q1        Med         Q3        Max
-3.4562925 -0.3767072 -0.1626930  0.3506447  5.7453552
Residual standard error: 16.25548
Degrees of freedom: 434 total; 430 residual
```

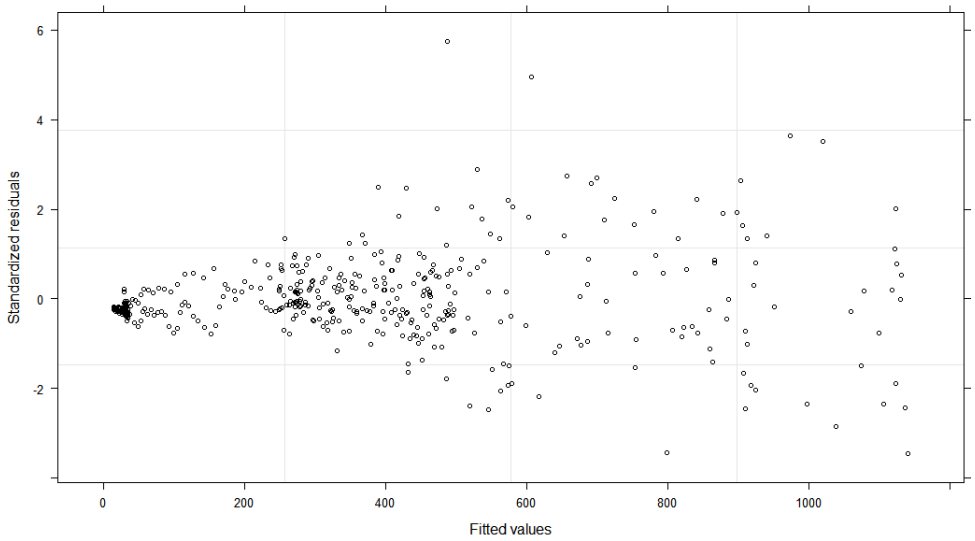


Figura B.11: Resultados gráficos del modelo GLS-AR(1) para la comunidad de Madrid. Función *fregre.gls*.

- GLS-ARMA(1,1)

```
summary(mod.gls.1.1.m[[1]])
Generalized least squares fit by REML
  Model: pf
  Data: XX
       AIC      BIC    logLik
  3367.876 3396.323 -1676.938
Correlation Structure: ARMA(1,1)
 Formula: ~1
 Parameter estimate(s):
     Phi1    Theta1
0.1042618 0.8553558
Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 375.4528 1.1181309 335.7860       0
x.PC1         0.3596 0.0045929  78.3043       0
x.PC2         0.5274 0.0057237  92.1467       0
x.PC3         0.7954 0.0355759  22.3587       0
 Correlation:
      (Intr) x.PC1  x.PC2
x.PC1 -0.001
x.PC2 -0.003 -0.028
x.PC3 -0.002  0.970 -0.027
Standardized residuals:
        Min          Q1         Med          Q3         Max
-3.9304857 -0.4062744 -0.1187129  0.3310361  5.8485879
Residual standard error: 15.64789
Degrees of freedom: 434 total; 430 residual
```

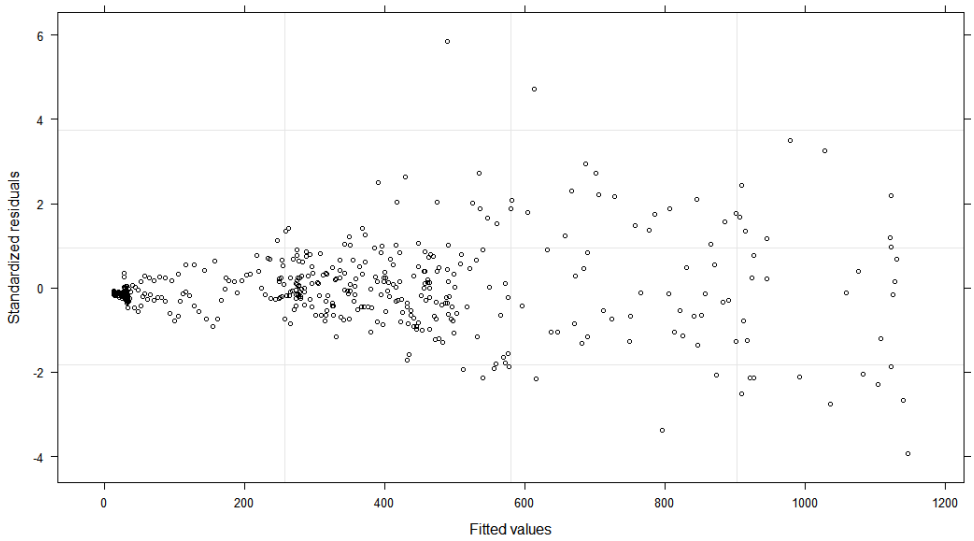


Figura B.12: Resultados gráficos del modelo GLS-ARMA(1,1) para la comunidad de Madrid. Función *fregre.gls*.

Resultados del modelo iGLS

- iGLS-AR(1)

```
mod.igls.1.m[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   375.4198        0.3511        0.5237        0.7275
> mod.igls.1.m[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
          ar1
      0.5626
s.e.  0.0396
sigma^2 estimated as 168.9:  log likelihood = -1729.09,  aic = 3462.19
```

- iGLS-ARMA(1,1)

```
mod.igls.p.m[[1]]
Call:
list("fregre.pc")
Coefficients:
(Intercept)         x.PC1         x.PC2         x.PC3
   375.4198        0.3511        0.5237        0.7275
> mod.igls.p.m[[1]][25]
$corStruct
$corStruct$ar
Call:
arima(x = x, order = c(p, d, q), include.mean = FALSE, transform.pars = TRUE)
Coefficients:
          ar1
      0.5626
s.e.  0.0396
```

```
sigma^2 estimated as 168.9:  log likelihood = -1729.09,  aic = 3462.19
```

B.2. Tablas de error a $H = 1, ..., 7$.

	$\varepsilon = y_{r+h} - \hat{y}_{r+h}$						
MODELO	H_1 "04/06/2021"	H_2 "05/06/2021"	H_3 "06/06/2021"	H_4 "07/06/2021"	H_5 "08/06/2021"	H_6 "09/06/2021"	H_7 "10/06/2021"
LM	-5	-11	-13	-13	-21	-28	-44
LM-AR(1)	-3	-4	-2	-6	-8	-10	-24
LM-ARMA(1,1)	-4	0	-2	-9	-3	-9	-23
SAM	-3	-8	-9	-6	-10	-15	-30
SAM-AR(1)	-2	-2	-1	-6	-8	-9	-22
SAM-ARMA(1,1)	-3	1	0	-8	-3	-9	-22
KAM	-2	-5	-4	1	-3	-7	-19
KAM-AR(1)	-7	-10	-13	-18	-22	-28	-46
KAM-ARMA(1,1)	-2	4	5	-2	4	0	-12
GLS-AR(1)	-4	-16	-16	-21	-25	-25	-38
GLS-ARMA(1,1)	-4	-15	-15	-23	-22	-26	-36
iGLS-AR(1)	-4	-16	-16	-21	-25	-25	-38
iGLS-AR(P)	-2	-2	-1	-18	-14	-15	-26
iGLS-ARMA(1,1)	8	3	0	26	-18	310	23

Tabla B.1: Error predicciones de los modelos FRM para la comunidad de Cataluña. Período 04/06/2020 al 10/06/2020.

	$\varepsilon = y_{r+h} - \hat{y}_{r+h}$						
MODELO	H_1 "04/06/2021"	H_2 "05/06/2021"	H_3 "06/06/2021"	H_4 "07/06/2021"	H_5 "08/06/2021"	H_6 "09/06/2021"	H_7 "10/06/2021"
LM	-5	-9	-11	-12	-16	-21	-26
LM-AR(1)	-4	-7	-8	-10	-13	-15	-17
LM-ARMA(1,1)	-4	-7	-9	-11	-14	-14	-18
SAM	-5	-7	-7	-6	-7	-9	-11
SAM-AR(1)	-4	-7	-8	-9	-12	-13	-14
SAM-ARMA(1,1)	-4	-7	-9	-10	-13	-13	-15
KAM	-20	-25	-27	-30	-35	-41	-47
KAM-AR(1)	-17	-22	-25	-28	-33	-37	-41
KAM-ARMA(1,1)	-19	-25	-29	-34	-41	-44	-51
GLS-AR(1)	-4	-8	-9	-11	-15	-17	-19
GLS-ARMA(1,1)	-4	-8	-10	-12	-16	-16	-19
iGLS-AR(1)	-4	-8	-9	-11	-15	-17	-19
iGLS-AR(P)	-4	-8	-9	-11	-15	-17	-19
iGLS-ARMA(1,1)	-4	160	-35	18	-113	238	177

Tabla B.2: Error predicciones de los modelos FRM para la comunidad de Cataluña. Período 04/06/2020 al 10/06/2020.

	$\varepsilon = y_{r+h} - \hat{y}_{r+h}$						
MODELO	H_1 "04/06/2021"	H_2 "05/06/2021"	H_3 "06/06/2021"	H_4 "07/06/2021"	H_5 "08/06/2021"	H_6 "09/06/2021"	H_7 "10/06/2021"
LM	-8	-14	-16	-18	-27	-34	-45
LM-AR(1)	-8	-10	-9	-12	-14	-14	-19
LM-ARMA(1,1)	-9	-6	-11	-12	-13	-15	-18
SAM	-12	-24	-30	-36	-47	-55	-69
SAM-AR(1)	-12	-18	-20	-25	-30	-30	-38
SAM-ARMA(1,1)	-14	-15	-21	-26	-28	-30	-36
KAM	41	35	31	28	21	17	12
KAM-AR(1)	52	53	54	55	56	61	68
KAM-ARMA(1,1)	39	43	39	39	40	42	45
GLS-AR(1)	-9	-26	-23	-30	-36	-31	-36
GLS-ARMA(1,1)	-10	-7	-14	-31	-36	-31	-33
iGLS-AR(1)	-9	-26	-22	-30	-36	-31	-36
iGLS-AR(P)	-9	-26	-22	-30	-36	-31	-36

Tabla B.3: Error predicciones de los modelos FRM para la comunidad de Madrid. Período 04/06/2020 al 10/06/2020.

B.3. Estimación de la estructuras de dependencia de los errores

	AR(1) del ε						
MODELO	ϕ_{MOD_1}	ϕ_{MOD_2}	ϕ_{MOD_3}	ϕ_{MOD_4}	ϕ_{MOD_5}	ϕ_{MOD_6}	ϕ_{MOD_7}
LM-AR(1)	0.613	0.741	0.830	0.873	0.908	0.941	0.948
SAM-AR(1)	0.590	0.736	0.820	0.855	0.890	0.922	0.930
KAM-AR(1)	0.820	0.873	0.901	0.917	0.931	0.945	0.952
GLS-AR(1)	0.688	0.993	0.994	0.995	0.996	0.997	0.997
iGLS-AR(1)	0.645	0.988	0.990	0.991	0.993	0.994	0.994
iGLS-AR(P)	0.855	1.141	1.326	1.270	1.281	1.389	1.459

Tabla B.4: Modelo autoregresivo de los errores AR(1) de los FRM de Cataluña.

	AR(1) del ε						
MODELO	ϕ_{MOD_1}	ϕ_{MOD_2}	ϕ_{MOD_3}	ϕ_{MOD_4}	ϕ_{MOD_5}	ϕ_{MOD_6}	ϕ_{MOD_7}
LM-AR(1)	0.873	0.929	0.951	0.963	0.969	0.974	0.976
SAM-AR(1)	0.770	0.843	0.882	0.900	0.911	0.918	0.918
KAM-AR(1)	0.806	0.855	0.883	0.905	0.920	0.932	0.941
GLS-AR(1)	0.979	0.991	0.994	0.996	0.997	1.000	1.000
iGLS-AR(1)	0.928	0.986	0.990	0.993	0.994	0.995	0.995
iGLS-AR(P)	0.928	0.986	0.990	0.993	0.994	0.995	0.995

Tabla B.5: Modelo autoregresivo de los errores AR(1) de los FRM de Galicia.

	AR(1) del ε						
MODELO	ϕ_{MOD_1}	ϕ_{MOD_2}	ϕ_{MOD_3}	ϕ_{MOD_4}	ϕ_{MOD_5}	ϕ_{MOD_6}	ϕ_{MOD_7}
LM-AR(1)	0.532	0.700	0.822	0.846	0.880	0.942	0.957
SAM-AR(1)	0.464	0.671	0.786	0.803	0.850	0.911	0.923
KAM-AR(1)	0.780	0.846	0.871	0.870	0.913	0.939	0.142
GLS-AR(1)	0.596	0.992	0.988	0.993	0.995	0.995	0.995
iGLS-AR(1)	0.563	0.987	0.982	0.988	0.991	0.990	0.991
iGLS-AR(P)	0.563	0.987	0.982	0.988	0.991	0.990 0	0.991

Tabla B.6: Modelo autoregresivo de los errores AR(1) de los FRM de Madrid.

	ARMA(1,1) del ε													
MODELO	MOD_1		MOD_2		MOD_3		MOD_4		MOD_5		MOD_6		MOD_7	
	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ
LM-ARMA(1,1)	0.259	0.701	0.536	0.740	0.721	0.581	0.759	0.706	0.820	0.659	0.896	0.567	0.913	0.699
SAM-ARMA(1,1)	0.207	0.700	0.516	0.694	0.702	0.541	0.726	0.651	0.788	0.602	0.864	0.514	0.884	0.646
KAM-ARMA(1,1)	0.820	NA	0.822	0.214	0.863	0.210	0.871	0.302	0.886	0.365	0.911	0.368	0.917	0.454
GLS-ARMA(1,1)	0.293	0.684	0.293	0.684	0.293	0.684	0.293	0.684	0.293	0.684	0.293	0.684	0.293	0.684
iGLS-ARMA(1,1)	0.934	0.366	0.593	0.727	0.728	0.598	0.792	0.780	0.850	0.694	0.996	0.825	0.979	0.678

Tabla B.7: Modelo autoregresivo de medias móviles de los errores ARMA(1,1) de los FRM de Cataluña.

	ARMA(1,1) del ε													
MODELO	MOD_1		MOD_2		MOD_3		MOD_4		MOD_5		MOD_6		MOD_7	
	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ
LM-ARMA(1,1)	0.834	0.175	0.887	0.431	0.927	0.398	0.941	0.539	0.952	0.594	0.960	0.620	0.966	0.749
SAM-ARMA(1,1)	0.656	0.304	0.760	0.350	0.804	0.484	0.830	0.527	0.845	0.585	0.858	0.652	0.854	0.623
KAM-ARMA(1,1)	0.806	NA	0.855	NA	0.883	NA	0.905	NA	0.909	0.074	0.918	0.121	0.921	0.192
GLS-ARMA(1,1)	0.975	0.085	0.975	0.085	0.975	0.085	0.975	0.085	0.975	0.085	0.975	0.085	0.975	0.085
iGLS-ARMA(1,1)	0.957	0.082	0.959	0.651	0.976	0.357	0.970	0.614	0.984	0.763	0.962	0.649	0.999	0.826

Tabla B.8: Modelo autoregresivo de medias móviles de los errores ARMA(1,1) de los FRM de Galicia.

	ARMA(1,1) del ε													
MODELO	MOD_1		MOD_2		MOD_3		MOD_4		MOD_5		MOD_6		MOD_7	
	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ	ϕ	θ
LM-ARMA(1,1)	0.874	NA	0.423	0.840	0.671	0.666	0.658	0.834	0.717	0.815	0.869	0.637	0.902	0.748
SAM-ARMA(1,1)	0.874	NA	0.351	0.800	0.580	0.676	0.555	0.812	0.656	0.802	0.828	0.528	0.836	0.749
KAM-ARMA(1,1)	0.780	NA	0.742	0.337	0.871	NA	0.870	NA	0.829	0.417	0.909	0.204	-0.546	0.729
GLS-ARMA(1,1)	0.104	0.855	0.104	0.855	0.104	0.855	0.104	0.855	0.104	0.855	0.104	0.855	0.104	0.855

Tabla B.9: Modelo autoregresivo de medias móviles de los errores ARMA(1,1) de los FRM de Madrid.

Bibliografía

[1] Ramsay J.O., Hooker G., Graves S. (2009) Functional Data Analysis with R and MATLAB. Springer New York. Estados Unidos.

[2] Ramsay J.O., Silverman B.W. (2005) Functional Data Analysis. Springer New York. Estados Unidos.

[3] Burguillo F.J., (2015) Desarrollo de un algoritmo de Mínimos Cuadrados Parciales para análisis de datos de chips de ADN usando el estadístico VIP para selección de genes y clasificación binaria. Salamanca.

[4] Hsing T., Eubank R. (2015) Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley. Estados Unidos.

[5] Ferraty F., Vieu P. (2006) Nonparametric Functional Data Analysis. Springer. Estados Unidos.

[6] Cuevas A. (2014) A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference, 147, 1-23.

[7] Cardot H., Ferraty F., Sarda P. (1998) Functional linear model. Statistics y Probability Letters 45 (1999) 11-22.

[8] Cuevas A., Febrero M., Fraiman R. (2007) Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics, 22(3):481-496.

[9] Pomanazzi M. (2009) Data depth, random simplices and multivariate dispersion. Statistics and Probability Letters 79 (2009) 1473-1479.

[10] Febrero M., Galeano P., González W. (2007) A functional analysis of NOx levels: location and scale estimation and outlier detection. Computational Statistics (2007) 22:411-427.

[11] Pons O., (2011) Functional Estimation for Density, Regression Models and Processes. World Scientific Publishing Co. Pte. Ltd. Estados Unidos.

[12] Ferraty F., (2011) Recent Advances in Functional Data Analysis and Related Topics. Springer Heidelberg Dordrecht London New York.

[13] Aneiros G., Cao R., Villar-Fernández J.M., Muñoz-San-Roque A., (2011) Functional Prediction for the Residual Demand in Electricity Spot Markets.F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 9-15.

[14] Attouch M.K., Gheriballah A., Laksaci A,. (2011) Robust Nonparametric Estimation for Functional Spatial Regression. On the Effect of Noisy Observations of the Regressor in a Functional Linear Model, pp 27-31.

[15] Bereswill M., Johannes J.,(2011) On the Effect of Noisy Observations of the Regressor in a Functional Linear Model. F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 41-47.

[16] Febrero-Bande M., González-Manteiga W., (2011) Generalized Additive Models for Functional Data. F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 91-96.

[17] Ferraty F., Goia A., Salinelli E., Vieu P., (2011) Recent Advances on Functional Additive Regression. F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 97-102.

[Ferraty-Martínez2011] erraty F., Martínez-Calvo A., Vieu P. (2011) Thresholding in Nonparametric Functional Regression with Scalar Response. Recent Advances in Functional Data Analysis and Related Topics, pp 103-109.

[18] Goia A., (2012) A functional linear model for time series prediction with exogenous variables. Statistics and Probability Letters 82 (2012) 1005?1011.

[19] McKeague I. W.,Qian M., (2011) Sparse Functional Linear Regression with Applications to Personalized Medicine. F. Ferraty (ed.), Recent Advances in Functional Data Analysis and Related Topics, pp 213-218.

[20] Kokoszka P., Reimherr M., (2017) Introduction to Functional Data Analysis. Taylor y Francis Group, LLC.

[21] Kariya T., Kurata H., (2004) Generalized Least Squares. John Wiley y Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, Inglaterra.

[22] Monsalve A., Harmath P. (2015) Introducción al análisis de series de tiempo con aplicaciones a la economía y finanzas. Ediciones IVIC. Venezuela.

[23] Mauricio J.A. (2007) Introducción al análisis de series de temporales. Universidad de Complutense de Madrid. España.

[24] Muscat J. (2014) Functional Analysis An Introduction to Metric Spaces, Hilbert Spaces, and Banach Algebras. Springer International Publishing. Switzerlan.

[25] Horváth l., Kokoszka P., (2012) Inference for Functional Data with Applications. Springer Science Business Media. New York.

[26] Bachman G., Narici L., (1966) Funcional Analysis. Department of Mathematics. Polytechnic Institute of Brooklyn. New York.

[27] Opsomer J., Wang Y., (1999) Nonparametric Regression with Correlated Error. https://www.researchgate.net