Trabajo Fin de Máster

Aplicación de análisis de datos funcionales en medios de pago

Laura Borrajo López

Máster en Técnicas Estadísticas

Curso 2016-2017

Propuesta de Trabajo Fin de Máster

Título en galego: Aplicación da análise de datos funcionais en medios de pago
Título en español: Aplicación de análisis de datos funcionales en medios de pago
English title: Functional data analysis application in payment systems
Modalidad: Modalidad B
Autora: Laura Borrajo López, Universidad de Vigo
Director: Manuel Febrero Bande, Universidad de Santiago de Compostela
Tutor: Pablo Montero Souto, ABANCA - Inteligencia de Clientes
Breve resumen del trabajo: El objetivo del proyecto que se propone es diseñar, entrenar y validar un modelo para predecir el resultado de acciones comerciales en alguno de los ámbitos de trabajo del área de Inteligencia de Clientes de ABANCA, siguiendo las distintas fases de trabajo: ▪ Análisis descriptivo de datos de clientes. ▪ Identificación de las variables explicativas sobre la respuesta a estudiar. ▪ Revisión y selección de técnicas adecuadas para modelizar la predicción de las futuras respuestas. ▪ Aplicación de los modelos en la generación de escenarios de negocio y anticipación de los resultados de negocio

Don Manuel Febrero Bande, Catedrático de la Universidad de Santiago de Compostela y don Pablo Montero Souto, Especialista de ABANCA - Inteligencia de Clientes, informan que el Trabajo Fin de Máster titulado

Aplicación de análisis de datos funcionales en medios de pago

fue realizado bajo su dirección por doña Laura Borrajo López 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 12 de enero de 2017.

El director: El tutor:

Don Manuel Febrero Bande Don Pablo Montero Souto

La autora:

Doña Laura Borrajo López

Índice general

Resumen

Resumen en español

Este trabajo responde a las necesidades de una entidad bancaria interesada en extraer provecho comercial de los pagos que realizan los usuarios de sus tarjetas, a fin de diferenciar grupos de clientes para acciones comerciales y perfilar a los destinatarios de sus campañas de marketing.

Tras acotar el problema planteado se justifica la posibilidad de aplicar el análisis de datos funcionales como técnica estadística para proporcionar respuestas orientadas, sobre todo, por su aplicación práctica. En este recorrido se sortean dificultades subyacentes al material de trabajo, que van desde la forma del dato funcional hasta las exigencias computacionales asociadas al tamaño de la base de datos.

Una primera parte descriptiva introduce las características informacionales de los datos bajo estudio y da paso a una segunda parte en donde se aborda la segmentación de grupos de clientes que serán utilizados en una posterior fase de modelización y predicción de eventos de interés en el negocio bancario. Para ello se emplea al test ANOVA con diferentes factores personales, a fin de contrastar las diferencias entre subgrupos de clientes. A continuación, se procede a formular varios ajustes de modelos de regresión para cada subgrupo.

Los resultados obtenidos validan la utilidad de la técnica empleada para proporcionar a la empresa una herramienta para clasificar a los usuarios de tarjetas en base a su perfil sociodemográfico y una estimación del gasto esperado, confirmando las sospechas previas de que algunos clientes aumentan su nivel de gasto en determinadas épocas de consumo.

English abstract

This work addresses the requirements from a bank which is interested on getting commercial outputs from their customers card payments. The goal is to establish different groups of customers for commercial actions and make the profile of the targets for marketing campaigns.

Once the problem has been established, possibilities to use data functional analysis is justified as a statistical approach to bring answers based on practical applications. Several difficulties linked to the work-in-progress are handled through this path, from the setting up of funcional data to the computational needs linked to the size of the dataset.

A first section introduces the informational features of the dataset. After that, a second section deals with the segmentation of customers' groups which would be used in a further step of model and predict important events in the banking business. For this purpose, the ANOVA test is used with different individual factors, in order to compare the differences between subgroups of customers. Next, several regression model adjustments are made for each subgroup.

The results validate how usefulness may be the used approach to help on the business problem, providing a tool to classify card users based on sociodemographics and to estimate the expected expenditure in annual seasons where the hypothesis of customers who increase their spending level was confirmed.

Prefacio

La industria europea de los medios de pago está cambiando de manera significativa. En los últimos años se ha ido abaratando el coste de las transacciones, al tiempo que ha crecido cada vez más el espacio para la aparición de nuevos intervinientes, sin olvidar la generación de nuevos derechos para los consumidores (en particular los de protección al consumidor y de la seguridad de los pagos) y la aparición de nuevos servicios complementarios a los pagos corrientes (*contact-less*, *electronic wallet*, *mobile payments*).

Con la implementación de la nueva directiva de servicios de pago (PSD2, por sus siglas en inglés), cualquier empresa registrada que posea el consentimiento de su cliente va a poder acceder a la información de sus cuentas bancarias en terceras entidades, facilitando así las transacciones dentro y fuera del sistema bancario tradicional. Con esta medida, el regulador está dispuesto a favorecer una mayor competencia con la intención de incentivar formas de pago más baratas.

Junto a los cambios en la regulación y configuración del mercado de pagos, la reducción de los márgenes de beneficio que se obtenían mediante la grabación de las comisiones y la negociación de las tasas de intercambio, está amenzando la centralidad de las entidades bancarias. A su vez, éstas tienen ante sí la oportunidad de ampliar sus líneas de negocio, pero también el desafío de retener aquellas donde tradicionalmente han sido dominantes.

Con la reducción de las barreras de entrada para los proveedores de servicios de pago, están apareciendo servicios disruptivos y formas alternativas de pago, especialmente, las protagonizadas por las *fintech*. Sus fórmulas específicamente diseñadas para la nueva era van desde la de creación de nuevas plataformas financieras adaptadas a Internet, como Paypal, hasta el aprovechamiento de posiciones dominantes para ocupar el espacio tradicionalmente reservado a los círculos financieros. Uno de los ejemplos más recientes lo protagoniza Facebook, que desde el pasado 30 de diciembre de 2016 ya puede operar en España como entidad de dinero electrónico, tras llevar al Registro Oficial de entidades del Banco de España la licencia para operar de Facebook Payments International Limited concedida por el Banco Central de Irlanda.

Los nuevos modelos de negocio de pagos que proponen estas compañías son especialmente eficientes, entre otros motivos, gracias a las utilidades que obtienen del tratamiento de los datos de micropagos y, de manera muy especial, de las transacciones generadas mediante el uso de tarjetas. Aunque se mantiene el secreto industrial sobre muchas de las formas que se emplean para monetizar estos datos, son conocidas, entre otras, las siguientes:

- Algoritmos de filtro colaborativo para para estimar las co-ocurrencias más comunes, donde los clientes que gastan dinero en un comercio son más propensos a gastar dinero en otros comercios, que han originanado los sistemas de recomendaciones personalizadas para las compras en comercios, al estilo de Amazon (Linden et al. 2003).

- Algoritmos de clasificación para la segmentación que ayudan a la creación de perfiles de gasto, útiles para departamentos de marketing, riesgo financiero y servicios digitales, así como para la venta a terceros de perfiles personales, que constituye la base de negocios de agregadores financieros (propio de marcas registradas como *Fintonic*) y de muchos mayoristas de la distribución (*retailers*) ante aseguradoras, entidades de crédito y gestores del riesgo parafinanciero (Hand y Henley 1997).

- Algoritmos de propensión al uso de financiación al consumo según importe y tipología de las compras, ajuste de series temporales para determinar el tiempo estimado para la próxima compra, y predicción de fraude, impago y recurrencia en dichos eventos, especialmente importantes en la financiación instantánea en el punto de venta.
- Algoritmos de decisión de compra venta de suelo comercial o instalación de paneles publicitarios, basados en la geolocalización de las transacciones.

En este contexto, las entidades bancarias han visto que necesitan hacer un mayor aprovechamiento del valor añadido que supone ser interventor en la intermediación de los cobros y pagos. Mientras que hasta no hace mucho se habían centrado casi exclusivamente en la seguridad de las operaciones de pago (contraseñas de autenticación y dispositivos seguros), cada vez está más presente la necesidad de dotar de inteligencia al negocio subyacente. En ello resulta fundamental explotar las capacidades analíticas para identificar a los clientes proclives a realizar pagos y financiarlos.

Este trabajo tiene su origen en la voluntad de explorar el uso de técnicas estadísticas para ampliar el conocimiento de los pagos de los clientes que existe en el departamento de marketing de una entidad financiera; en donde hasta la fecha no se disponen de herramientas estadísticas orientadas a la predicción del gasto con tarjetas de crédito. En el contexto bancario de mayor atención a este tipo de pagos, este objetivo atiende a la demanda particular de la entidad para identificar a candidatos en futuras campañas de marketing y definir públicos objetivos (*commercial targets*) para la comercialización de incentivos al uso de tarjetas entre sus clientes, a grandes rasgos, usuarios que están comenzando a sustituir el efectivo por el uso de tarjetas al ritmo que refleja el "Informe Tecnocom. Tendencias en Medios de Pago 2016", que puede consultarse en Tecnocom (2016).

Capítulo 1

Problema empresarial

En el quehacer del departamento de Inteligencia de Clientes de una entidad bancaria subyace el interés por entender los distintos patrones de uso de sus tarjetas como medio de pago. Además de demandar una herramienta que permita segmentar y clasificar a estos usuarios en función de su operativa con tarjeta, se cree que el consumo puede estar sujeto a patrones temporales, por lo que el objetivo final será predecir cómo evolucionará su patrón de gasto en determinadas fechas del año, en las que se tiene la sospecha de que las transacciones aumentan de manera significativa en respuesta a eventos del mercado de consumo, como períodos de rebajas o incentivos económicos de oferta.

En el ámbito de tarjetas, la entidad ha facilitado para el objetivo de este análisis un conjunto de aproximadamente 40 millones de operaciones correspondiente al año 2015, que constituye una muestra representativa de la realidad que se pretende estudiar. Estos registros llevan asociados una serie de variables referentes a la operación, como fecha, importe y tipología de tarjeta con la que se ha efectuado la compra; otras a nivel cliente, relativas a sus características socio-demográficas (sexo, edad, actividad contable, localidad) y detalles de su actividad (saldos medios y finales mensuales) y otras a nivel comercio, como la categoría y los datos geolocalizados de éste. Con tal cantidad de información, se puede entrever que una de las principales problemáticas de este estudio reside en los datos.

Considerando las posibilidades de estudio que se presentan ante tanta información, las dimensiones de los datos y el reto que supone la programación con un conjunto de datos de estas características, es imprescindible definir el problema de forma que la cantidad de datos no sea un impedimento a la hora de realizar el estudio. Cabe mencionar, que de forma análoga a lo que ocurre en Big Data, en la era de la información en la que vivimos, el secreto ante datos de estas característcas no reside en acumular cantidades ingentes de información, sino en gestionarla adecuadamente para producir valor; en el trabajo que nos concierne, se trata de poder establecer patrones de consumo que permitan clasificar al cliente según su comportamiento de gasto y ofrecerle, así, una atención más personalizada, por medio de la cual ambas partes resulten beneficiadas.

En la Sección 1.1, veremos una primera acotación del problema sobre la que comenzaremos a trabajar; no obstante, conforme avancemos veremos sucesivas restricciones que será necesario considerar de cara a la obtención de resultados.

De entre todas las características mencionadas acerca de las transacciones sobre las que la entidad tiene información, definiremos en la Sección 1.2 aquellas variables que caracterizan los datos que vamos a estudiar y analizaremos cuáles serán empleadas en la clasificación de clientes.

1.1. Acotación del problema

Dada la mayor presencia de la entidad en Galicia, el problema será analizado sobre las operaciones con tarjeta realizadas por los clientes bancarizados residentes en esta comunidad autónoma, centrando nuestro estudio en los clientes gallegos de entre 18 y 65 años de edad, considerando las transacciones realizadas tanto en comercios gallegos como en el resto de España durante el año 2015. Agruparemos

las diversas localidades en las que se han efectuado compras en cinco zonas, correspondientes a las cuatro provincias gallegas (A Coruña, Lugo, Ourense y Pontevedra) y una quinta englobando todas aquellas transacciones acaecidas en el resto de España.

Es importante señalar que los datos personales de los clientes han sido previamente anonimizados y las coordenadas de los comercios ligeramente desplazadas en base al cumplimiento de las normas de protección de datos y a la política de riesgo operativo de la entidad.

1.2. Tipo de datos a analizar

La extracción de los datos se realiza por medio de consultas en SQL operadas en una de las bases de datos de la entidad bancaria. En un primer momento, se intentó trabajar con la operativa total de un período anual; no obstante, el uso del *software* estadístico `R` en el tratamiento de datos hace inviable analizar tal volumen de datos debido a las limitaciones computacionales asociadas al manejo de datos en la memoria local de una máquina de capacidades básicas. Posteriormente, se realizó la extracción por trimestres, lo que no supuso una mejora significativamente suficiente en el procedimiento de estudio. Por lo que finalmente nos decantamos por seccionar la muestra de clientes por la variable *segmento_edad* en 4 grupos: **jóvenes** (18-25 años), **adultos 1** (26-45 años), **adultos 2** (46-59 años) y **sénior** (60-65 años). Esta decisión está justificada por la definición de segmentos de edad que es reconocible para la entidad bancaria, por lo que se sustenta en motivos de comprensibilidad para los responsables del negocio.

En el Cuadro 1.1 se expone la información acerca del tamaño muestral y número de operaciones de cada uno de estos subgrupos, con la que se manifiestan las dimensiones del problema al que nos enfrentamos. Como podemos observar, las transacciones no se producen de manera proporcional entre los diferentes subgrupos de edad, por lo que será interesante analizar de forma diferenciada cada uno de ellos.

SUBGRUPO	Nº DE CLIENTES	Nº DE OPERACIONES
Jóvenes	68 317	2 624 715
Adultos 1	264 668	16 808 697
Adultos 2	168 450	9 279 977
Sénior	50 792	2 215 140
TOTAL	552 227	30 928 529

Cuadro 1.1: Información del tamaño muestral de los subgrupos de clientes por *segmento_edad*.

De entre todas las variables relativas a operaciones que nos ha facilitado la entidad para este análisis, la extracción de datos va encaminada a aquellas que aporten información útil en el análisis que pretendemos llevar a cabo: fechas, importes, tipología de tarjeta (débito o crédito), datos sobre el comercio, su categoría y la localidad en la que se encuentra e información acerca del cliente (sexo, edad y actividad contable). Dado que las transacciones no se producen de manera equiespaciada, hemos optado por la creación de la variable *semana*, lo que facilitará la observación de períodos de una actividad intensiva.

En el Cuadro 1.2 se muestra toda la información disponible sobre cada operación. Teniendo en cuenta que ésta solo es una parte de la información que posee la entidad, se vislumbra la multitud de oportunidades de análisis que ofrecen los datos. De entre este conjunto de variables, emplearemos las variables *importe* y *semana*, imprescindibles en la creación del tipo de dato que vamos a tratar y, por supuesto, la variable *cliente_id*, que nos permite agrupar las transacciones de cada cliente. Decidimos, además, centrarnos en el uso de unas pocas variables categóricas que nos permitan agrupar clientes en subgrupos más pequeños con los que sea más sencillo y factible trabajar: *actividad*, *localidad*, *segmento_edad* y *sexo*, siempre y cuando la formación de subgrupos sea justificable en base a algún método.

VARIABLE	DESCRIPCIÓN
actividad	Tipo character. Clasifica a los clientes según su actividad contable en `Inactivos Estudiantes/Amas de casa/Pensionistas`, `Activos ocupados`, `Activos parados` u `Ocupaciones no bien especificadas`.
cliente_id	Tipo integer. Permite agrupar las transacciones de cada cliente, de forma que su identidad es anónima.
comercio_actividad_ds	Tipo character. Nombre de la empresa a la que pertenece el comercio en el que se realiza la transacción.
comercio_actividad_id	Tipo integer. Número identificador de la empresa a la que pertenece el comercio en cuestión.
comercio_ds	Tipo character. Nombre del comercio, propietario del TPV, donde se ha realizado la operación.
comercio_id	Tipo integer. Número mediante el que se identifica el comercio en el que se ha realizado la transacción.
comercio_sector_ds	Tipo character. Indica el nombre del sector al que pertenece la empresa intemediaria en la operación.
comercio_sector_id	Tipo integer. Identificador del sector al que pertenece la empresa en la que se realiza la operación.
edad	Tipo integer. Toma valores entre 18 y 65 años, intervalo de edad al que se restringe el análisis.
fecha_valor_mov	Tipo character. Proporciona la fecha en la que la operación ha sido realizada, necesaria para la creación de la variable *semana*.
importe	Tipo numeric. Indica el valor monetario de la transacción. Toma valor positivo si la operación es una compra y valor negativo en caso de devolución.
localidad	Tipo character. Indica la localidad en la que se encuentra el comercio y por tanto, en la que tiene lugar la transacción.
segmento_edad	Tipo character. Indica el grupo de edad al que pertenece el cliente: `jóvenes`, `adultos 1`, `adultos 2` o `sénior`.
semana	Tipo character, de creación propia. Obtenida a partir de *fecha_valor_mov*, indica la semana en la que se realiza la transacción, tomando valores de 1 a 53.
sexo	Tipo character, indicadora del sexo del cliente: `V` para el conjunto de hombres y `H` para el grupo de las mujeres.
tipo_tj	Tipo character. Hace diferencia entre el tipo de tarjeta: tarjeta de crédito (`C`) y débito (`D`).

Cuadro 1.2: Resumen de las variables asociadas a cada transacción.

En la Figura 1.1, mostramos cómo se distribuyen los clientes en función de las variables *edad* y *segmento_edad* y en la Figura 1.2, por las variables *sexo*, *localidad* y *actividad*, considerando las distintas categorías de estas variables y grupos de edad. Estos diagramas de barras no se corresponden con el conjunto de datos original, sino con el que trabajaremos a lo largo del Capítulo 3, obtenido tras la redución de dimensión por diferentes métodos que introduciremos más adelante. El uso del conjunto de clientes reducido se justifica en base a que refleja fielmente la situación original sin necesidad de tener que trabajar con un conjunto mayor de observaciones, lo que ocasionaría elevados consumos de memoria y mayores exigencias computacionales.

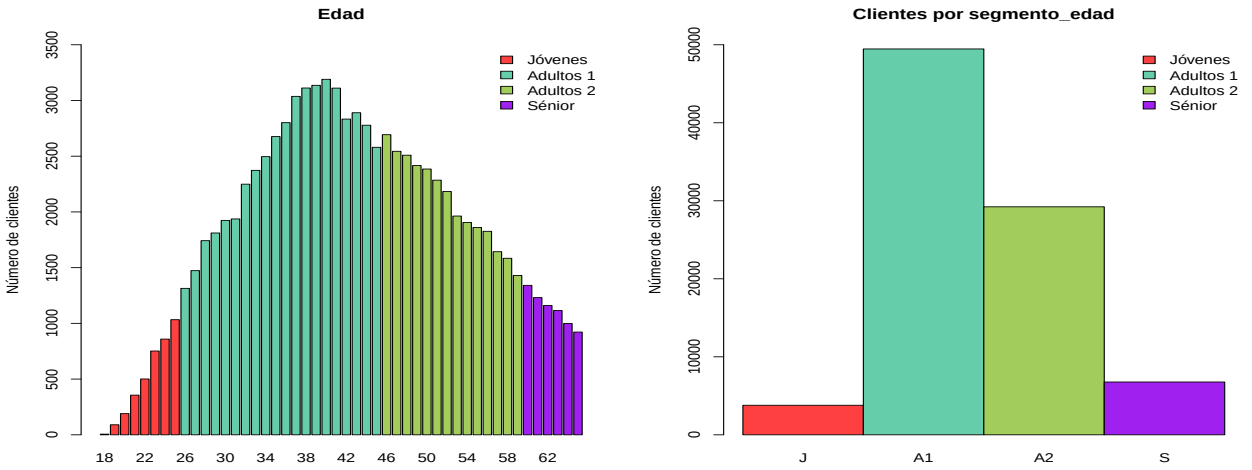

Figura 1.1: Gráfico de barras para la variable *edad* y *segmento_edad*, respectivamente.

Observando la Figura 1.1, comprobamos que el subgrupo de adultos 1 será uno de los más complejos a la hora de realizar el análisis, por tener una concentración mayor de clientes, seguido del subgrupo adultos 2. De la Figura 1.2, se puede deducir la mayor presencia de mujeres independientemente del grupo de edad considerado; en lo referente a la localidad, vemos que la mayor parte de transacciones transcurren en A Coruña y Pontevedra, en consonancia con la mayor industrialización del territorio, siendo la primera la que concentra una mayor cantidad de registros y en relación a la actividad contable, observamos que salvo el subgrupo de los jóvenes que está ligeramente dominado por los *inactivos*, el grupo mayoritario es el de los *activos ocupados*, lo cual resulta coherente con las características del problema bajo estudio, pues una de las restricciones que se empleará para la reducción del tamaño muestral será considerar a aquellos clientes cuya operativa de tarjeta tenga una cierta regularidad, lo que puede estar en parte ligado a tener unos ingresos recurentes.

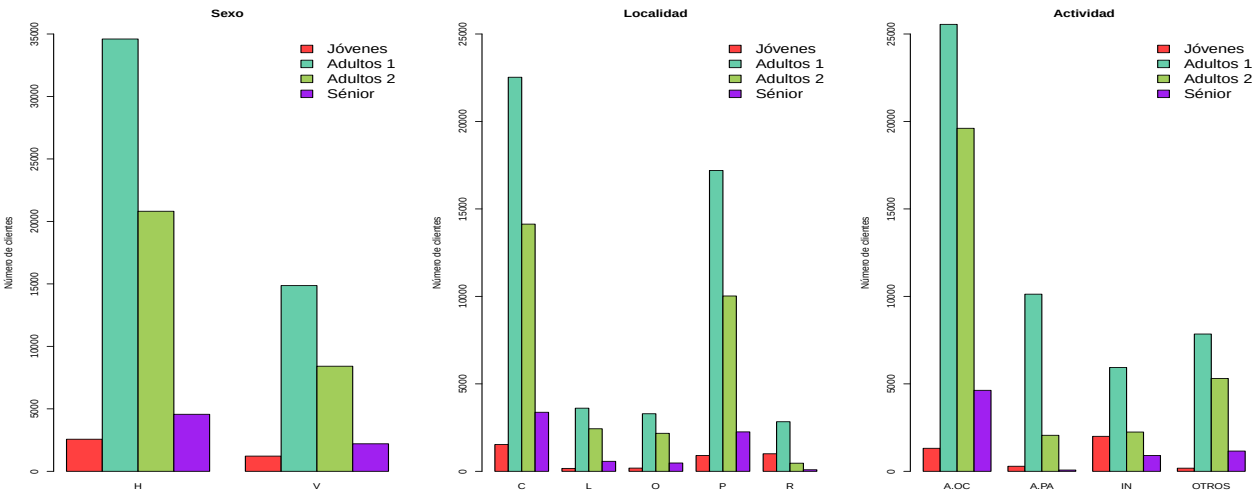

Figura 1.2: Gráfico de barras para la variable *sexo*, *localidad* y *actividad*, respectivamente, considerando categorías y grupo de edad.

Capítulo 2

Herramientas estadísticas

Para satisfacer los objetivos de estudio y predicción de este trabajo es necesario analizar los importes de gasto de los clientes a lo largo del año 2015. Al reflexionar sobre esto, viene a nuestra cabeza la idea de un proceso estocástico subyacente, por consistir estos en secuencias de datos que evolucionan con el tiempo; más concretamente, podemos pensar en series temporales, un caso particular de los procesos estocásticos en las que las observaciones son tomadas en intervalos regulares de tiempo.

El TSA (*Time Series Analysis*) mediante la metodología ARIMA o Box-Jenkins podría ser una opción a tener en cuenta para llevar a cabo nuestro estudio (véase Peña et al. 2011, Box et al. 2015). No obstante, esto implicaría enfrentarse al análisis de tantas series temporales como individuos se tienen, una idea poco razonable. Además, se tiene interés por conocer el consumo a lo largo de las 53 semanas del año y dado que las series de tiempo son eficientes cuando la dependencia es relativamente corta en el tiempo, esto nos hace pensar que el TSA no es la mejor opción para nuestro problema.

El análisis de datos funcionales, FDA (*Fuctional Data analysis*), puede ser otra opción a considerar. Las variables funcionales se caracterizan por la evolución de una variable a lo largo del tiempo, de modo que los valores que toman son funciones en lugar de vectores como en análisis multivariante clásico. La imposibilidad de medir la mayoría de estas variables continuamente en el tiempo, unida a la complejidad teórica de muchos de los métodos estadísticos disponibles para su análisis llevan a que sean más utilizadas las series temporales. Aunque existen muchas técnicas para la modelización y predicción de datos temporales discretos, la mayoría de ellas, como por ejemplo la ya comentada teoría clásica de Box-Jenkins, imponen que se verifiquen hipótesis bastante restrictivas como estacionariedad, observaciones igualmente espaciadas o pertenencia a una clase de procesos específica, por lo que el FDA puede ser una herramienta preferible ante ciertas circunstancias.

En la Sección 2.1.1 se muestran las aplicaciones del FDA en un entorno financiero, en base a poder justificar su utilización en nuestro estudio.

Como ya comentamos, la cantidad de registros en disposición para nuestro estudio oscila los 40 millones de operaciones anuales; en particular, nuestro conjunto de datos original abarca más de 30 millones de transacciones, considerando 16 variables sobre cada operación. Vemos, por tanto, que estamos ante un conjunto de datos complejos. En la Sección 2.1.2, se analizan las posibilidades del estudio con datos funcionales sobre conjuntos de estas características.

2.1. Aplicaciones de datos funcionales

En este apartado analizaremos las aplicaciones del análisis de datos funcionales para tratar el tipo de problemas al que nos enfrentamos. En la Sección 2.1.1 veremos el impacto del FDA en un contexto financiero a lo largo del tiempo, mientras que en la Sección 2.1.2 analizaremos cómo ha sido empleado en el estudio de datos complejos.

2.1.1. Datos funcionales en Finanzas

En un contexto de medios de pago, Laukaitis y Rackauskas (2002) emplean el modelo funcional autorregresivo (AR) para predecir la intensidad y número de transacciones realizadas por medio de tarjetas de crédito, tanto en cajeros automáticos como en TPV (Terminal Punto de Venta), empleando datos de alta frecuencia. Posteriormente, puede consultarse en Laukaitis (2008) cómo este mismo modelo es utilizado como predictor del flujo de caja, tratando series temporales como funciones continuas aleatorias proyectadas en el subespacio de baja dimensión.

Laukaitis y Rackauskas (2005) realizan un análisis funcional de la varianza (FANOVA) para constatar la existencia de variaciones en los procesos de flujo de caja atribuíbles a variables categóricas: la zona geográfica en el caso de redes de cajeros automáticos y la categoría del comercio (Merchant Category Code) en el caso de redes de TPV's. Al ser los soportes de cajeros y TPV's bastante caros, resulta interesante estudiar cómo afectan dichas variables de cara a tomar una decisión acerca de la inversión de los recursos lo más eficiente posible.

En Benko (2007) puede consultarse un análisis de datos funcionales con aplicación a finanzas, concretamente al estudio de las curvas de rendimientos del EURIBOR. Dado que las observaciones se tienen en un grid discreto, el dato funcional debe construirse mediante la recopilación de todas ellas generando una curva continua.

El análisis de datos financieros de alta frecuencia e intradía es realmente importante en diversas áreas como finanzas, econometría y estadística. El principal problema con el que uno se puede encontrar al analizar una serie de rendimientos intradía es que ésta sea muy ruidosa. En Miao (2013) se propone el uso del FDA como una posible solución a esta cuestión, mediante el cual las curvas ruidosas se transforman en suaves y estacionarias. En vista de esta y otras aplicaciones comentadas, observamos que el análisis de datos funcionales en este contexto está estrechamente relacionado con el estudio de series temporales financieras.

En Dablemont et al. (2007) se aplica el FDA al modelado y predicción en series de tiempo financieras. Tras este análisis se comprueba como el uso de datos funcionales puede ser particularmente eficaz cuando las observaciones son escasas, irregularmente espaciadas, cuando se producen en diferentes puntos de tiempo para cada curva o cuando solo se observan fragmentos de estas, en comparación a los métodos estándar, que fallan completamente en estas circunstancias.

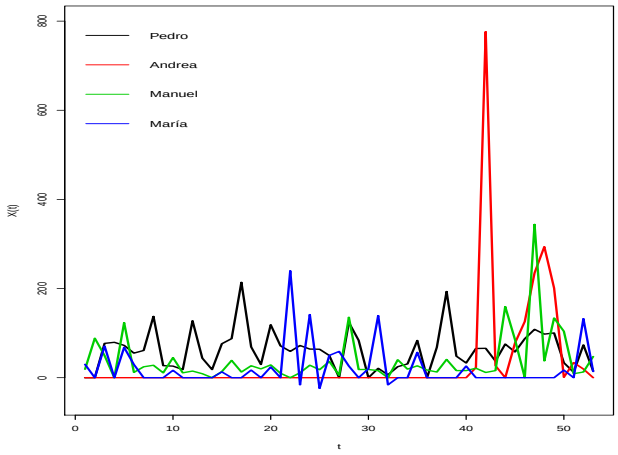

Figura 2.1: Representación de la operativa anual por semanas de 4 clientes del subgrupo de los jóvenes.

Precisamente, son éstas las características presentes en los datos descritos en la Sección 1.2, de ahí que pensemos en la aplicación del FDA al problema bajo estudio. Así, por ejemplo, dentro del grupo de clientes jóvenes, nos encontramos con diferentes situaciones: un cliente, llamémosle Pedro, con 455

operaciones anuales a lo largo de 48 semanas; Manuel, con 5 semanas de inactividad y un total de 73 operaciones anuales; Andrea, que realiza 151 operaciones en 11 semanas de movimientos o María, que a lo largo de 23 semanas de actividad realiza 52 transacciones (véase la Figura 2.1).

Además, cabe mencionar en este punto la propia limitación de los datos para inferir el comportamiento de consumo, puesto que se trata de una parte de la actividad de gasto de cada individuo; concretamente, la que se restringe al uso de las tarjetas de esta misma entidad bancaria, al margen de la operativa que puedan realizar con otros medios de pago, incluyendo el efectivo, que se presume mayoritario en los micropagos que se realizan en el territorio.

2.1.2. Análisis funcional de datos complejos

Realizando una exhaustiva revisión bibliográfica sobre la aplicación del FDA sobre Big Data o sobre conjuntos de datos complejos, vemos que por lo de ahora no existen métodos específicos para tratar conjuntos de datos funcionales grandes o complejos; si bien se pueden abordar otras opciones, como la segmentación, la redución de dimensión o la creación de subgrupos que faciliten el análisis de datos funcionales.

En Lardin-Puech et al. (2014) se tratan conjuntos de datos muy grandes de fenómenos que evolucionan con el tiempo. Para su análisis, se propone el uso de técnicas de remuestreo ante objetivos sencillos de estimación, en el que se establece un intercambio entre el tamaño de los datos y la precisión de los estimadores. Se presentan diferentes métodos de remuestreo para tratar grandes conjuntos de datos funcionales, que se pueden emplear para construir bandas de confianza y mejorar, con la ayuda de información auxiliar, la estimación de los estimadores en comparación con el muestreo aleatorio simple sin reemplazamiento, ilustrando dichos procedimientos sobre un conjunto de curvas de consumo de electricidad.

Cao (2015) revisa qué se considera grande para obtener la mención de Big Data, calificación que puede adquirir una muestra de datos por tener tamaño muestral grande, por ser la dimensión de la estructura de datos que se recoge para cada individuo grande o por la combinación de ambas circunstancias. Frecuentemente, los datos viven en un espacio de dimensión infinita; al menos, en teoría. Un ejemplo de esta situación es el tratado en Kavousian et al. (2012) sobre los datos de consumo eléctrico medidos en intervalos de 10 minutos, situación en la que tiene sentido considerar cada dato como una función.

En Chen et al. (2015) se consideran los datos funcionales como parte del Big Data por constituir realizaciones suavizadas de los procesos estocásticos subyacentes, los cuales son objetos de dimensión esencialmente infinita. Se pone de manifiesto la necesidad de reducir la dimensión al trabajar con este tipo de datos (de gran dimensión o infinito-dimensionales) y se advierte acerca del carácter preliminar del FDA en este contexto, bajo la esperanza de que en un futuro tenga un impacto importante en la resolución de este tipo de situaciones.

2.2. Análisis del problema con datos funcionales

La herramienta escogida para llevar a cabo este trabajo es el análisis de datos funcionales. Se trata de poner a prueba las posibilidades del FDA como vía exploratoria para abordar unos datos que nunca antes han sido tratados desde esta perspectiva en el trabajo de la entidad y que, si bien es cierto que se podrían abordar problemas similares desde otros enfoques, resulta de interés como aportación al recorrido que podría tener un trabajo continuado en este ámbito.

Las principales referencias consultadas sobre este tema son Ferraty y Vieu (2006), Horváth y Kokoszka (2012), Ramsay y Dalzell (1991) y Ramsay y Silverman (2005), en las que nos apoyaremos a lo largo del estudio del caso.

En Ferraty y Vieu (2006) se expone que una variable aleatoria $\mathcal{X}$ es una variable funcional si toma valores en un espacio de dimensión infinita, o espacio funcional, $\mathcal{F}$; mientras que un conjunto funcional $\{\mathcal{X}_i\}_{i=1}^n$ consiste en la observación de n variables funcionales idénticamente distribuídas a $\mathcal{X}$. El dato funcional se corresponde con cada observación χ de $\mathcal{X}$.

Cuando $\mathcal{X}$ describe una curva aleatoria podemos denotar la variable funcional en función de sus observaciones

$$\mathcal{X} = \{\mathcal{X}(t); t \in T\}$$

donde T denota el conjunto de puntos en los que la curva es observada.

Para analizar las curvas de gasto de clientes, es imprescindible generar el dato funcional, por lo que debemos acumular las operaciones que cada cliente realiza en un período de tiempo determinado, período que coincidirá con los puntos de discretización temporal; en nuestro caso, semanas, teniendo así un total de 53 puntos de discretización correspondientes a cada una de las semanas del año. La construcción de la variable *semana* se hace de manera intencional para dar forma funcional a los datos, puesto que no se dispone de número de observaciones suficientes para cada *timestamp*, como posible punto de discretización máximo de las operaciones. De esta forma, cada curva representa la operativa de tarjeta de un cliente a lo largo del año 2015, donde los valores positivos se corresponden a gastos (compras) que ha efectuado el cliente y los valores negativos son relativos a abonos (devoluciones).

Un cliente intermediario, es decir, el comercio que posee un TPV, genera series cuyos datos son variables de Poisson en un grid continuo, mientras que un cliente finalista genera las mismas variables pero en un grid discreto con ceros en los puntos de discretización. En este sentido, el tipo de dato que genera un TPV o un cajero está más próximo a la filosofía de un dato funcional que el dato generado por los plásticos (denominación que adoptan las tarjetas en el sector financiero) de los clientes finalistas, porque el régimen de operatividad es discontinuo. No obstante, puesto que el interés de la entidad reside en conocer mejor a sus clientes finalistas, se pueden adaptar estos datos teniendo en cuenta las consideraciones anteriormente citadas y otras restricciones que veremos más adelante, de forma que sea posible expresarlos con apariencia de dato funcional.

2.2.1. Limitaciones de la aplicación de técnicas FDA

Independientemente de las técnicas utilizadas en la resolución del problema que se plantea, la principal limitación que surge en el análisis parte del volumen de datos que debemos analizar, ya que el conjunto de transacciones totales de los clientes gallegos de entre 18 y 65 años durante el 2015 equivalen a más de 5 GB de información.

Las técnicas de análisis de datos funcionales también tienen ciertas limitaciones debido al carácter complejo de los datos. En algunos de estos casos, la solución viene de emplear una técnica análoga más eficiente computacionalmente hablando, este es el caso de algunos modelos de regresión o del cálculo de ciertas profundidades; en otras ocasiones, será necesaria una mayor segmentación de los datos para aplicar dichas técnicas, como ocurre con la suavización o el test ANOVA sobre un conjunto grande de curvas. Estas medidas serán puestas en práctica a lo largo del Capítulo 3.

El *machine learning*, también conocido como aprendizaje automático o aprendizaje de máquina, puede otorgarnos ciertas pautas a seguir en el tratamiento de grandes volúmenes de datos o cuando estos son complejos. Consiste, básicamente, en un proceso de inducción del conocimiento en la máquina, que se complementa a la perfección con el Big Data, tal y como se deduce de la siguiente afirmación de Lohr (2012):

"The wealth of new data accelerates advances in computing, a virtuous circle of Big Data. Machine-learning algorithms, for example, learn on data, and the more data, the more the machines learn."

La idea sería trasladar las etapas del *machine learning* a nuestro estudio, consistentes en segmentación, reducción de la dimensionalidad y extracción de subgrupos, con lo que se vería significativamente simplificado el análisis, planteando una primera aproximación a la forma que podrían presentar los algoritmos de predicción, de modo que pudieran ser retroalimentados por nuevos conjuntos de datos.

2.2.2. Técnicas de clasificación ANOVA

Con el objetivo de clasificar a nuestros clientes, utilizaremos el test ANOVA para justificar la segmentación del grupo total de clientes en diferentes subgrupos en base a diferentes variables.

Dadas k muestras independientes de datos funcionales, el test consiste en contrastar la hipótesis nula de igualdad de sus respectivas funciones medias. Se trata de un ajuste similar al del modelo anova clásico, con la salvedad de que las muestras son datos funcionales. Puede ser visto como la versión asintótica del anova F-test. En Cuevas et al. (2004) puede consultarse la validez asintótica de este método, así como un estudio de simulación y una aplicación a datos reales de cardiología experimental, en los que se pone en práctica el ANOVA sobre datos funcionales.

Se aplicará tanto el ANOVA *One Factor*, considerando una sola variable factor que permita hacer grupos, como el ANOVA *Multiway*, aplicado sobre más de un factor y considerando todas las interacciones posibles, que puede consultarse en Cuesta-Albertos y Febrero-Bande (2010).

Este análisis de segmentación y clasificación de clientes se desarrollará en la Sección 3.3, considerando las variables factor: *segmento_edad*, *sexo*, *localidad* y *actividad*, tanto por separado como conjuntamente.

2.2.3. Modelos de regresión

En este apartado se realiza una revisión de los principales modelos existentes en regresión funcional con respuesta escalar. Se trata de una exploración descriptiva, en la que profundizaremos en el aspecto práctico a lo largo de la Sección 3.4, donde ajustaremos algunos de estos modelos de regresión, aquellos que sean más convenientes dadas las características de nuestros datos. Mediante el análisis de regresión, se analizará el gasto producido en cuatro fechas importantes en lo que a consumo se refiere: San Valentín, Semana Santa, Black Friday y Navidad, tras el cual se abordará el objetivo final de predicción.

- **Modelos lineales**

  Sea $\mathcal{X} \in \mathcal{L}_2(T)$ e $y \in \mathbb{R}$. Suponiendo que $\mathbb{E}(\mathcal{X}(t)) = 0$, $\forall t \in [0,T]$ y que $\mathbb{E}(y) = 0$, el modelo de regresión lineal funcional, FLM (*Functional Linear Model*), se define del siguiente modo:

  $$y = \langle \mathcal{X}, \beta \rangle + \epsilon = \int_T \mathcal{X}(t)\beta(t)dt + \epsilon$$

  donde $\beta \in \mathcal{L}_2(T)$, ϵ es el término correspondiente al error y T denota el intervalo de definición del dato funcional.

  Nos centraremos en la regresión desde el punto de vista de la estimación, en particular en la del parámetro β. Una forma de estimar β consiste en representar el parámetro, y opcionalmente $\mathcal{X}_i$, en una base de $\mathcal{L}_2(T)$ del siguiente modo:

  $$\beta(t) = \sum_k \beta_k \theta_k(t), \qquad \mathcal{X}_i(t) = \sum_k c_{i,k}\psi_k(t).$$

  A continuación, proponemos diferentes mecanismos para la estimación del parámetro β.

  - Representación en bases:

    Parte de la representación de β y $\mathcal{X}$ en bases de $\mathcal{L}_2$, utilizando B-splines, Wavelets o bases de Fourier (véase Ramsay y Silverman 2002, 2005; Cardot 2000; Cardot et al. 2003; Antoniadis y Sapatinas 2003).

  - Componentes principales:

    En el modelo FPC (*Functional Principal Components*), las componentes principales de $\mathcal{X}$ son combinaciones lineales dadas por las autofunciones del operador de covarianza de $\mathcal{X}$. Al igual que en el modelo clásico de componentes principales, el proceso $\mathcal{X}$ y el conjunto de sus autofunciones abarcan el mismo espacio lineal, por lo que las componentes principales constituyen una base ortonormal de $\mathcal{L}_2$.

En Cardot et al. (1999) se busca evitar un sobreajuste y minimizar la suma residual de cuadrados de los estimadores involucrados. En Hall et al. (2006) se estudia el comportamiento teórico de los $\hat{\beta}$ respecto a la selección de componentes principales y en Cardot et al. (2007) se propone la regresión *Ridge* consistente en modificar la estimación de β con el fin de resolver su estabilidad cuando se añaden al modelo algunos términos correspondientes a autovalores pequeños.

- Mínimos cuadrados parciales:

  La idea del FPLS (*Functional Partial Least Squares*) consiste en construir un conjunto de variables ortogonales de forma similar al modelo FPC pero maximizando, en este caso, la covarianza entre $\mathcal{X}$ e y, en vez de buscar las direcciones de máxima variabilidad de $\mathcal{X}$. Puede consultarse Preda y Saporta (2005), donde se presentan los fundamentos teóricos del método para datos funcionales y Krämer et al. (2008), donde se aborda el cálculo de los grados de libertad asociados al ajuste por mínimos cuadrados parciales.

- **Modelos no lineales y semilineales**

  - No lineales:

    Se considera $(\mathcal{X}, y)$ un par de variables aleatorias con $y \in \mathbb{R}$ y $\mathcal{X} \in \mathbb{E}$, donde $\mathbb{E}$ es un espacio semimétrico. En este modelo, la fórmula de la esperanza condicionada

    $$m(\mathcal{X}) = \mathbb{E}(Y|X = \mathcal{X})$$

    se estuctura como un estimador tipo *kernel* (véase Ferraty y Vieu 2006).

  - Semilineales:

    Se considera $(\mathcal{X}, \mathbf{Z}, y)$ con $y \in \mathbb{R}$ (respuesta), $\mathcal{X} \in \mathbb{E}$ (funcional) y $\mathbf{Z} \in \mathbb{R}^p$ (covariables), definiendo el siguiente modelo (véase Aneiros-Pérez y Vieu 2006):

    $$y = \mathbf{Z}\beta + m(\mathcal{X}) + \epsilon$$

- **Modelos generalizados**

  - Modelo lineal generalizado:

    Los FGLM (*Functional Generalized Linear Model*) suponen una extensión de los modelos funcionales lineales a situaciones más generales en que la respuesta y pertenece a la familia exponencial. En este contexto se tiene que

    $$\mathbb{E}(y) = \mu = g^{-1}(\eta)$$

    donde g denota la función link y η el predictor lineal.

    Para estimar η, se proyecta $\mathcal{X}$ y β sobre un número finito de elementos de una base funcional. Para ello se pueden emplear bases fijas: B-splines, Wavelets, Fourier (véase James 2002) o recurrir a los modelos funcionales de componentes principales (puede consultarse en Cardot y Sarda 2005; Escabias et al. 2004, 2005; Müller y Stadtmüller 2005) o al FPLS (véase Preda and Saporta 2005; Escabias et al. 2007).

    La estimación de β se obtiene aplicando iterativamente el método IRLS (*Iterated Reweighted Least Squares*) hasta alcanzar el β que minimiza la desviación (*deviance*).

- Modelo aditivo generalizado:

  La suposición de un efecto estrictamente lineal en el predictor puede no ser apropiada, bien porque algunos efectos pueden tener una forma desconocida o porque las interacciones entre las covariables pueden adoptar una forma compleja. Por ello, un modelo FGAM (*Functional Generalized Additive Model*) consiste en escribir la esperanza condicionada general como un predictor adictivo de las variables funcionales.

  - Modelo aditivo espectral generalizado:

    El FGSAM (*Functional Generalized Spectral Additive Model*) se basa en la obtención de ese predictor a partir de los *scores* de componentes principales, como puede verse en en Müller y Yao (2008).

  - Modelo aditivo Kernel generalizado:

    El FGKAM (*Functional Generalized Kernel Additive Model*) emplea una representación tipo *kernel* de la contribución de las variables funcionales a la respuesta (véase Febrero-Bande y González-Manteiga (2013)).

Capítulo 3

Estudio del caso

El análisis de los datos se desarrollará en diferentes etapas. En primer lugar, en la Sección 3.1 se detallarán las medidas adoptadas en la creación del dato funcional y las técnicas aplicadas en la reducción del tamaño muestral. A continuación, en la Sección 3.2 se realizará un análisis exploratorio de los datos. En la Sección 3.3 se expondrá la creación de subgrupos de clientes motivada por el uso del test ANOVA, sobre los cuales se llevará a cabo el análisis de regresión, en la Sección 3.4, que posibilitará la predicción del comportamiento futuro de los clientes existentes y de las nuevas incorporaciones a la entidad en el ámbito de medios de pago, en determinadas fechas señaladas. Para finalizar, veremos en la Sección 3.5, un resumen acerca de las exigencias computacionales de este trabajo.

Para la resolución del problema que nos plantea la entidad emplearemos el entorno de programación `R`. Concretamente, nos apoyaremos en el paquete `fda.usc` implementado por Febrero-Bande y Oviedo (2012). A lo largo de este capítulo se mostrarán algunos de los resultados más relevantes obtenidos durante el análisis.

Mencionar que los resultados gráficos expuestos en este capítulo se corresponderán, en su mayoría, a muestras de tamaño muy inferior al analizado realmente, debido a la gran cantidad de memoria que sería necesaria para almacenar las gráficas del conjunto total de datos funcionales que vamos a tratar. No obstante, recalcar que el estudio ha sido realizado teniendo en cuenta la totalidad del conjunto de clientes definido a continuación.

3.1. Preparación de datos

En este apartado se describirán los pasos seguidos en la obtención del dato funcional que queremos analizar, entre ellos el importante reto que ha supuesto tener que lidiar con la gran dimensionalidad de los datos.

Como ya comentamos anteriormente, dado que la cantidad de registros de operaciones realizadas con tarjeta disponibles para este trabajo oscila en torno a los 40 millones de observaciones anuales, el problema será analizado sobre las transacciones realizadas por los clientes bancarizados residentes en Galicia, centrando el estudio en los clientes de entre 18 y 65 años de edad, considerando las operaciones realizadas en comercios gallegos y en el resto de España durante el 2015. Debido a problemas de consumo de memoria en la lectura de la operativa total, los datos serán analizados sobre la agrupación en base a la variable *segmento_edad*, segmentación que vendrá justificada más adelante.

3.1.1. Trabajos preliminares a la aplicación de técnicas estadísticas

El principal inconveniente a la hora de trabajar con estos datos es su gran tamaño muestral, pues como pudimos ver en el Cuadro 1.1 contamos con un gran número de operaciones y clientes en cada subgrupo de edad. Analizando esta situación inicial, ya simplificada por la segmentación de la variable *edad*, queda patente, por medio del Cuadro 3.1 de tiempos de consumo correspondientes a la carga de

datos y a la creación de la variable *semana* y matriz de gasto, la dificultad de analizar estos conjuntos de datos con el *software* estadístico `R`, pese a que la lectura de datos se ha realizado con la función `fread` del paquete `data.table`, específico para tratar grandes conjuntos de datos.

SUBGRUPO	Nº DE OPERACIONES	TIEMPO LECTURA	TIEMPO SEMANA Y GASTO
Jóvenes	2 624 715	0.45	2.10
Adultos 1	16 808 697	8.75	72.43
Adultos 2	9 279 977	2.88	16.65
Sénior	2 215 140	0.33	1.57

Cuadro 3.1: Tiempos de consumo (en minutos) correspondientes a la carga de datos y la creación de la variable *semana* y matriz de gasto para cada subgrupo de edad.

En la selección del período en el que evaluar la curva, nos hemos decantado por crear la variable *semana*, con el objetivo de simplificar la composición del dato funcional. De este modo, se agrupan los importes por cliente y semana, de forma que cada curva del dato funcional se corresponde con la operativa de tarjeta de un cliente a lo largo del año 2015, siendo los puntos de discretización las correspondientes 53 semanas. De esta forma, pasamos a analizar tantas curvas como número de clientes son estudiados, pero evaluadas en tan solo 53 puntos, medida que permite reducir significativamente la dimensión del problema, si lo comparamos con los 365 puntos de evaluación que resultarían de considerar el gasto diario.

A continuación, mostramos una serie de exclusiones y pasos a seguir en la composición del dato funcional para solventar el problema del tamaño muestral, entre otros, como son la presencia de clientes que no aportan la información necesaria para realizar un análisis completo de su operativa anual, bien por tener escasos movimientos anuales de tarjeta o bien porque que han sido dados de baja o de alta a lo largo del año 2015; la varianza en los datos y la presencia de *outliers*, que no reflejan el comportamiento global de la clientela de la entidad.

- **Reducción del tamaño muestral:**

  Una vez obtenida la matriz de importe de gasto de cliente por semana, decidimos reducir la muestra a aquellos usuarios que han realizado operaciones con su tarjeta a lo largo de 43 o más semanas en el año 2015, ya que observamos que muchos clientes hacen un uso poco regular de estas, bien sea porque son clientes dados de alta o de baja en ese año o porque el número de semanas en el que han operado es muy escaso. De este modo, centramos el análisis en los usuarios regulares de las tarjetas, que pueden tener patrones de comportamiento más trazables, dejando al margen del estudio los casos de uso esporádico, que presentarían un patrón errático e introducirían ruído en la comprensión de la actividad principal que nos interesa evaluar.

  Observando el Cuadro 3.2, podemos cerciorarnos de la significativa reducción que experimenta el tamaño muestral y, en menor medida, el número de operaciones al aplicar esta medida.

SUBGRUPO	Nº DE CLIENTES	Nº DE OPERACIONES
Jóvenes	3 987	660 352
Adultos 1	52 067	9 057 855
Adultos 2	30 765	5 211 240
Sénior	7 125	1 091 733
TOTAL	93 944	16 021 180

Cuadro 3.2: Información del tamaño muestral de los subgrupos de clientes por *segmento_edad* tras eliminar los registros de clientes con más de 10 semanas de inactividad.

En la Figura 3.1 podemos visualizar la representación del dato funcional para una muestra de 100 clientes de cada grupo de edad, mediante la que se define el patrón de consumo de los clientes en cuestión. Como prueba de la difícil tarea que supone adaptar el FDA a un conjunto de datos complejos como el que poseemos, comentar que no sería posible obtener una gráfica análoga para el conjunto original de clientes, definido en el Cuadro 1.1 debido a problemas de memoria en la realización de ciertas tareas. Sí podríamos realizar la representación del conjunto definido en el Cuadro 3.2 con el inconveniente asociado de grandes necesidades de almacenamiento, por lo que a efectos visuales, nos conformamos con la representación para una pequeña muestra.

Basta observar la imagen para comprobar que la detección de observaciones atípicas será una cuestión a tratar. También se observa la necesidad de aplicar alguna tranformación que equilibre la varianza en los datos así como una suavización de los mismos.

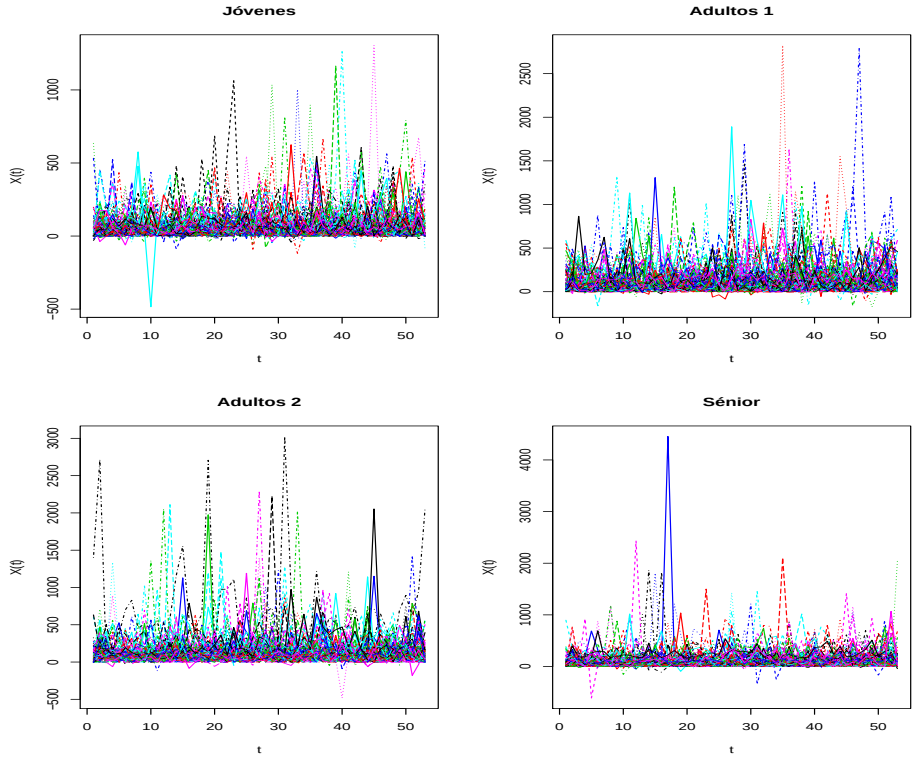

Figura 3.1: Representación gráfica de una muestra de 100 curvas de consumo a lo largo del año 2015 para cada subgrupo de edad, tras eliminar las relativas a clientes con 10 semanas de inactividad o menos.

- **Transformación logarítmica:**

  Para equilibrar la varianza en los datos, aplicaremos la transformación logarítmica sobre los importes de las transacciones. Esta modificación lleva asociados algunos inconvenientes. La imposibilidad de calcular el logaritmo de un valor negativo, cuando durante la semana se han producido más devoluciones que compras, o cero, cuando la operativa semanal es nula, hace que sea necesario aplicar el logaritmo sobre el valor absoluto del importe y sobre el resultado asignar de nuevo signo positivo si se trata de una compra y signo negativo en caso de devolución; sin embargo, este procedimiento establece nuevas trabas cuando el resultado semanal tiene un importe menor de un euro, puesto que el logaritmo de una compra por valor entre 0 y 1 toma

valor negativo (asociado a devoluciones) y el logaritmo del valor absoluto de una devolución por valor entre -1 y 0 toma valor negativo, pero al asignarle signo negativo se convierte en positivo (asociado a compras), por lo que decidimos fijar todos los importes comprendidos entre -1 y 1, ambos inclusives, a cero, dado que se trata de cantidades despreciables asociadas a compensaciones, devoluciones, reintegros o descuentos aplicables por las compras, que no reflejan el interés de los importes y las operaciones bajo estudio. De este modo, eliminamos también los falsos ceros, correspondientes a resultados semanales de importes más pequeños que el céntimo y que de aplicarle la transformación logarítmica tomarían valores extremos que gráficamente se traducirían en picos, dando la falsa apariencia de datos atípicos.

- **Suavización:**

  Lo que nos interesa observar en el dato funcional es la tendencia a largo plazo de la actividad de la tarjeta, por lo que es necesario suavizar los datos. Pueden consultarse diferentes métodos de suavización en Ramsay y Silverman (2005). En este caso, nos decantaremos por la ténica basada en la utilización de una función kernel, considerando kernel gaussiano por su mejor funcionamiento en comparación a otros y tomando como parámetro de suavización $h = 3$.

  Como ya comentamos en la Sección 2.2.1 acerca de las limitaciones del FDA en este contexto, resulta inviable, con las herramientas a nuestra disposición, realizar la suavización en el grupo adultos 1 a menos que se realice una segmentación previa en subconjuntos de clientes más pequeños.

- **Detección de atípicos:**

  En datos funcionales no existe una definición general de *outlier*. Consideraremos que una curva es una observación atípica cuando ha sido generada por un proceso diferente al resto de curvas. Un *outlier* se caracteriza por tener una profundidad funcional baja, por lo que será *outlier* en el sentido de la profundidad usada.

  La profundidad es una herramienta estadística que proporciona un orden a los datos, clasificándolos en más o menos profundos, de dentro hacia fuera. Al tratarse de tal cantidad de datos, las profundidades más adecuadas, por calcularse más ágilmente, son la de Fraiman-Muniz (FMD), analizada en Fraiman y Muniz (2001) y la de Proyecciones Aleatorias (RPD), que puede consultarse en Cuevas et al. (2007). En nuestro análisis, nos decantamos por la profundidad de Fraiman-Muniz, que puede ser vista como el promedio de una profundidad univariante a lo largo del dominio de definición del dato funcional.

  A pesar de la existencia de métodos *bootstrap* para la detección de atípicos, dadas las características de nuestros datos, no son aplicables a este contexto. En su lugar, decidimos emplear el método de recortado (*trimming*), que puede verse con más detalle en Febrero et al. (2007) y que calcula el promedio del 95 % de los datos más profundos, por ser este el valor escogido para el parámetro de recortado.

  En base a este método, tal y como veremos en el Cuadro 3.3, logramos reducir nuevamente el tamaño muestral, aunque en menor medida en esta ocasión, y deshacernos de aquellas observaciones que difieren del comportamiento general del grupo.

  En la Figura 3.2 ejemplificamos, sobre una muestra de 100 clientes del subgrupo joven, el proceso de obtención de datos que acabamos de describir. Desde la eliminación de registros de clientes con 10 o más semanas de inactividad, pasando por la transformación logarítmica y la suavización, hasta restringir la muestra al 95 % de datos más profundos. Un análisis análogo a este ha sido realizado para los subgrupos adultos 1, adultos 2 y sénior sobre la totalidad de clientes.

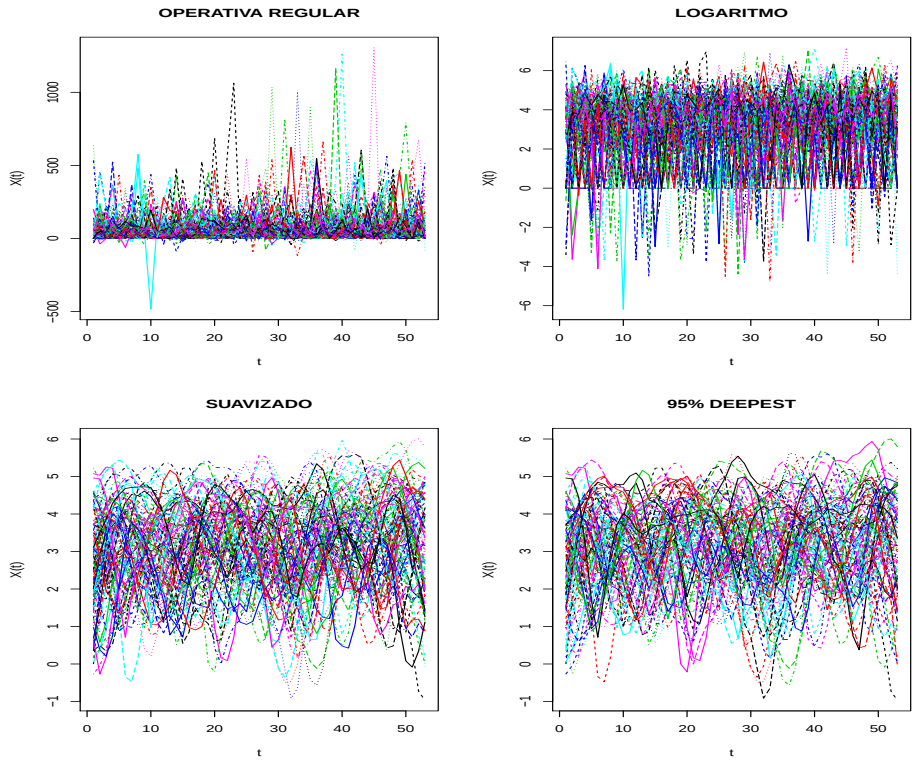

Figura 3.2: Proceso de obtención de datos, analizado sobre una muestre de 100 clientes del subgrupo de jóvenes.

3.1.2. Definición de subgrupos

Tras las exclusiones aplicadas sobre las curvas de consumo para la composición del dato funcional objeto de estudio, estamos ante el escenario descrito en el Cuadro 3.3, que aporta información acerca de los tamaños muestrales y número de operaciones de los subgrupos de edad con los que trabajaremos de ahora en adelante. Si comparamos esta información con las cifras obtenidas en el Cuadro 1.1, verificamos que el problema al que nos enfrentamos en un principio ha sido significativamente simplificado y es, al fin, abordable.

SUBGRUPO	Nº DE CLIENTES	Nº DE OPERACIONES
Jóvenes	3 787	605 977
Adultos 1	49 463	8 309 163
Adultos 2	29 226	4 790 249
Sénior	6 768	1 005 274
TOTAL	89 244	14 710 663

Cuadro 3.3: Información del tamaño muestral de los subgrupos de clientes por *segmento_edad* tras eliminar los registros de clientes con más de 10 semanas de inactividad y restringir el estudio al 95 % de los datos más profundos.

3.2. Análisis exploratorio de datos funcionales

A lo largo de esta sección se realiza una revisión de las principales herramientas en el análisis exploratorio de datos funcionales.

Sea la variable funcional $\mathcal{X}$ correspondiente a los importes de gasto semanales (transformados por logaritmos) efectuados con tarjeta a lo largo del año 2015. Tenemos que

$$\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n,$$

donde n se corresponde con el tamaño muestral al que se hace referencia en el Cuadro 3.3

$$n = n_J + n_{A1} + n_{A2} + n_S$$

siendo $n_J = 3\,787$, $n_{A1} = 49\,463$, $n_{A2} = 29\,226$ y $n_S = 6\,768$ el número de clientes analizado de cada subgrupo de edad y, por tanto, $n = 89\,244$ el número total de datos funcionales a estudiar.

El análisis exploratorio será realizado sobre una muestra de 500 clientes del subgrupo de los jóvenes, representada en la Figura 3.3. Pese a que en la Sección 3.1 fue obtenido el dato funcional que nos interesa analizar, en este punto el estudio será realizado sobre una muestra del conjunto de curvas previo a la tranformación logarítimica, suavización y detección de atípicos, puesto que una de las herramientas que veremos a continuación, el cálculo de profundidades, está estrechamente relacionada con la detección de *outliers*, por lo que analizar la profundidad de los datos sería el paso previo que nos permite justificar la reducción de dimensión al prescindir de las observaciones menos profundas.

A efectos de ejemplificar la exploración de datos, consideraremos, a lo largo de este apartado,

$$\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n \quad \text{con} \quad n = 500,$$

siendo $\mathcal{X}$ la variable funcional correspondiente a los importes de gasto semanales.

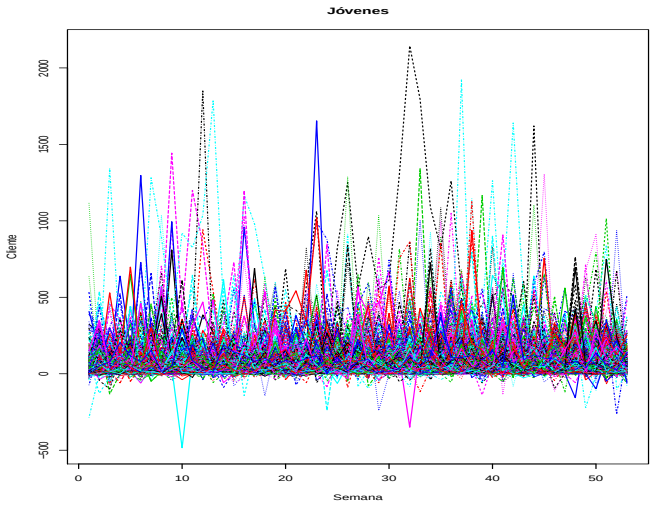

Figura 3.3: Representación de una muestra de 500 observaciones del subgrupo de jóvenes correspondientes a la operativa de gasto con tarjeta durante el año 2015.

3.2.1. Medidas de localización

En este apartado procedemos a calcular ciertas medidas de localización global que permiten obtener información acerca del comportamiento de los datos.

Dado $\mathcal{S}$ el proceso generador de datos y $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ la muestra de datos antes definida con $\mathcal{X}_i \overset{iid}{\sim} \mathcal{S}$, la media funcional se define como el centro de gravedad de los datos:

$$\min_{a \in \mathcal{F}} \sum_{\mathcal{X} \in \mathcal{S}} d(\mathcal{X}, a)^2$$

y como estamos en un espacio de Hilbert, esto es la definición de media usual

$$\bar{\mathcal{X}}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i(t).$$

La otra medida de localización a considerar es la mediana funcional, definida como el elemento minimizador del error absoluto medio

$$\min_{a \in \mathcal{F}} \sum_{\mathcal{X} \in \mathcal{S}} d(\mathcal{X}, a)$$

que en este caso, dado que en general no es posible obtener una expresión cerrada como en el caso de la media, se opta por buscar su análoga muestral como elemento que minimiza este criterio

$$\min_{a \in \mathcal{S}_n} \sum_{i=1}^{n} d(\mathcal{X}_i, a).$$

Se representa la media y la mediana muestral en la Figura 3.4. Se observa que estamos ante un conjunto de curvas que tiene un comportamiento de gasto medio muy estable, de importes bajos, en donde se hace difícil apreciar subgrupos diferenciales a simple vista, así como irregularidades apreciables entre las distintas semanas de actividad, con la excepción habitual de ciertos clientes.

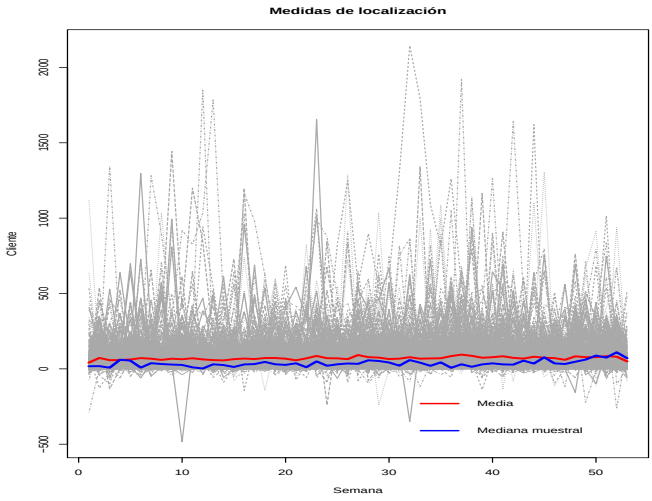

Figura 3.4: Representación de las medidas de localización para una muestra de 500 clientes del subgrupo jóvenes.

3.2.2. Cálculo de profundidades

Proseguimos el análisis exploratorio con el cálculo de profundidades en base a diferentes métodos.

Como ya comentamos, la profundidad es una herramienta estadística que proporciona información acerca del orden interno de los datos, clasificándolos en más o menos profundos, de dentro hacia fuera. Este orden se puede emplear para definir medidas de localización y también *outliers* por oposición,

como ya vimos en la Sección 3.1, en la que definimos una observación atípica como aquella asociada a una profundidad funcional baja.

Una medida de localización es encontrar el elemento que maximiza cada una de las siguientes profundidades, así como todas las posibilidades que salen de medidas robustas para eliminar un porcentaje de datos menos profundos y calcular la media con los restantes (media recortada). En función de la profundidad considerada, el concepto de qué es más o menos profundo se ve ligeramente modificado:

- Profundidad de Fraiman-Muniz (FMD)

  Sean $\mathcal{S}_n = \{\mathcal{X}_i(t)\}_{i=1}^n$ realizaciones *iid* de una variable aleatoria funcional con dominio $\mathcal{T}$ y $\mathcal{D}$ una medida de profundidad en $\mathbb{R}$. Para cada punto de discretización $t_0 \in \mathcal{T}$, se considera

  $$z_i(t_0) := \mathcal{D}(\mathcal{X}_i(t_0))$$

  la profundidad univariante del dato i en t_0 con respecto a $\{\mathcal{X}_i(t_0)\}_{i=1}^n$. La FMD (*Fraiman-Muniz Depth*) puede ser vista como el promedio de la profundidad univariante a lo largo del dominio de definición del dato funcional (véase Fraiman y Muniz 2001):

  $$FMD(\mathcal{X}_i) = \int_{\mathcal{T}} z_i(t)dt.$$

- Profundidad Modal (MD)

  Dadas $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ realizaciones *iid* de una variable aleatoria funcional, $K : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ una función kernel asimétrica y h el parámetro ventana, se define en Cuevas et al. (2007) la MD (*Modal Depth*) como una medida de cuántos puntos hay en la vecindad

  $$MD(\mathcal{X}_i) := \sum_{j=1}^{n} K\left(\frac{d(\mathcal{X}_i, \mathcal{X}_j)}{h}\right).$$

- Profundidad de Proyecciones Aleatoria (RPD)

  Sean $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ realizaciones *iid* de una variable aleatoria funcional, $h \in \mathcal{H}$ una realización del proceso de dirección independiente $\mathcal{H}$ y $P_i^h = \langle h, \mathcal{X}_i \rangle \in \mathbb{R}$ la proyección de $\mathcal{X}_i$ a lo largo de la dirección h. Se define la RPD (*Random Projection Depth*) en Cuevas et al. (2007) como

  $$RPD(\mathcal{X}_i, h) := \mathcal{D}(P_i^h)$$

  siendo $\mathcal{D}$ una medida de profundidad univariante. Aunque una proyección es suficiente, sería preferible generar una colección de proyecciones aleatorias $\{h_I\}_{I=1}^M$ y calcular la profundidad usando todas las proyecciones

  $$RPD(\mathcal{X}_i, \{h_I\}_{I=1}^M) := \frac{1}{M}\sum_{I=1}^{M} \mathcal{D}(P_i^{h_I}).$$

- Profundidad de Tukey Aleatoria (RTD)

  Cuesta-Albertos y Nieto-Reyes (2008) presentan una variante de la profundidad de proyecciones aleatorias, la RTD (*Random Tukey Depth*)

  $$RTD(\mathcal{X}_i, \{h_I\}_{I=1}^M) := \min_M \mathcal{D}(P_i^{h_I}).$$

Como ya comentamos en su momento, las profundidades más adecuadas en el tratamiento de grandes conjuntos de datos son la FMD y RPD, de ahí que se haya empleado la profundidad de Fraiman-Muniz en la detección de observaciones atípicas en la sección anterior. No obstante, en la Figura 3.5 realizamos el cálculo de profundidades sobre la muestra de 500 clientes del subgrupo de jóvenes, empleando las cuatro opciones citadas. Se observan los datos funcionales en escala de grises, siendo más profundo cuanto más oscuro, en rojo la observación más profunda y en amarillo la media recortada considerando un parámetro de recortado α=0.25. Se pueden percibir las diferencias existentes en función de la profundidad empleada. Además, algunos picos de las profundidades llevan a mantener la hipótesis de trabajo respecto de ciertos eventos que disparan el gasto. En la Figura 3.6 se representan los datos más profundos en cada caso.

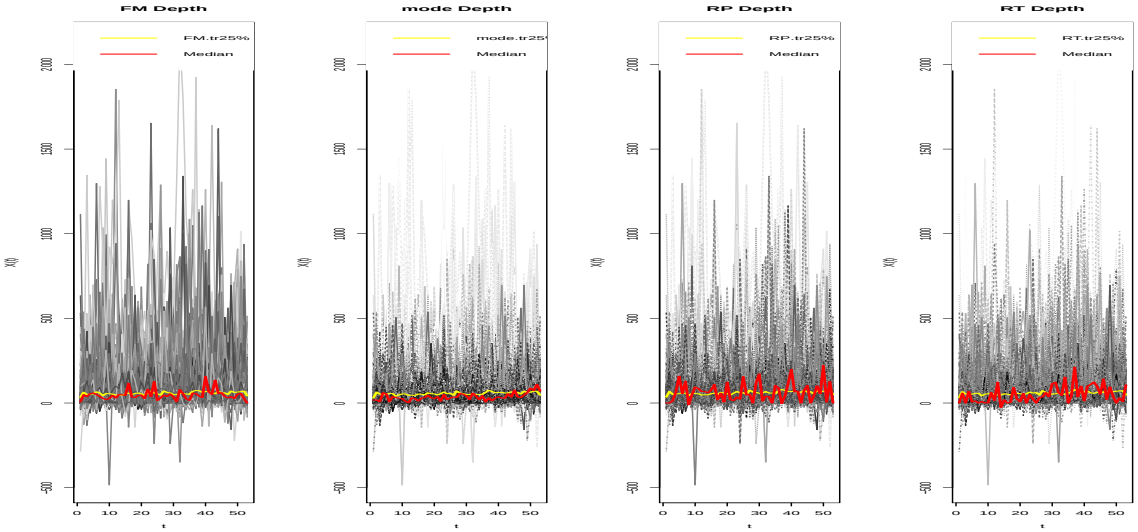

Figura 3.5: Representación gráfica del cálculo de profundidades en base a diferentes opciones utilizadas sobre una muestra de 500 clientes del subgrupo jóvenes. De izquierda a derecha: profundidad de Fraiman-Muniz, Modal, Proyecciones Aleatorias y Tukey Aleatoria.

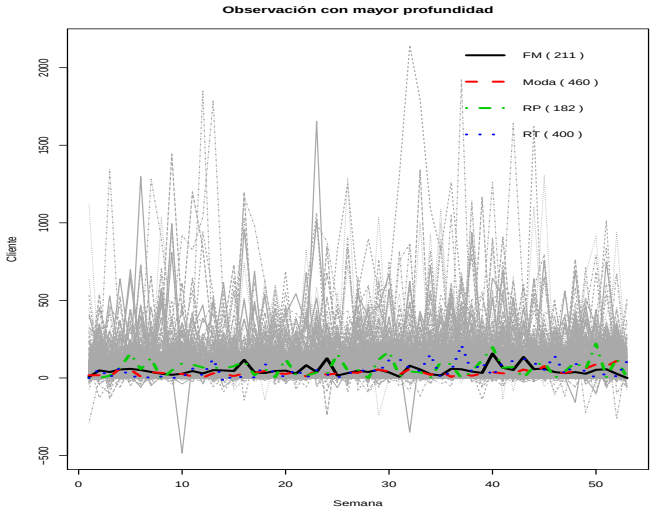

Figura 3.6: Representación de las observaciones más profundas según las profundidades de Fraiman-Munir, Modal, Proyecciones Aleatorias y Tukey aleatoria sobre una muestra de 500 clientes del subgrupo jóvenes.

En la Figura 3.7 se muestran gráficas tipo *boxplot* funcional. En color gris los datos con profundidad inferior al cuantil 5 %, desde este hasta la mediana en color azul y en rojo aquellos datos funcionales cuya profundidad es mayor que la mediana. Al igual que en la Figura 3.5, se aprecian ciertas diferencias en relación a la profundidad utilizada.

El proceso de detección de atípicos descrito en la Sección 3.1, en el que restringimos nuestro conjunto de datos al 95 % más profundo considerando la FMD, puede ser fácilmente explicado llegados a este punto, puesto que el proceso equivaldría a prescindir de las observaciones en color gris de la primera gráfica en la Figura 3.7, correspondientes al 5 % de los datos menos profundos según esta profundidad.

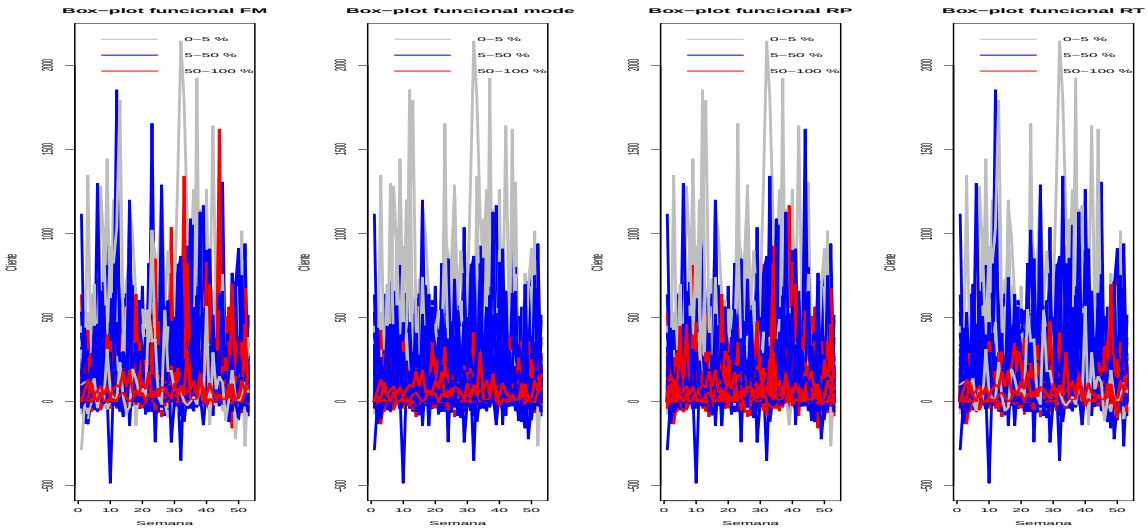

Figura 3.7: *Boxplot* funcional en base a diferentes profundidades (FMD, MD, RPD y RTD, respectivamente sobre una muestra de 500 clientes del subgrupo jóvenes.

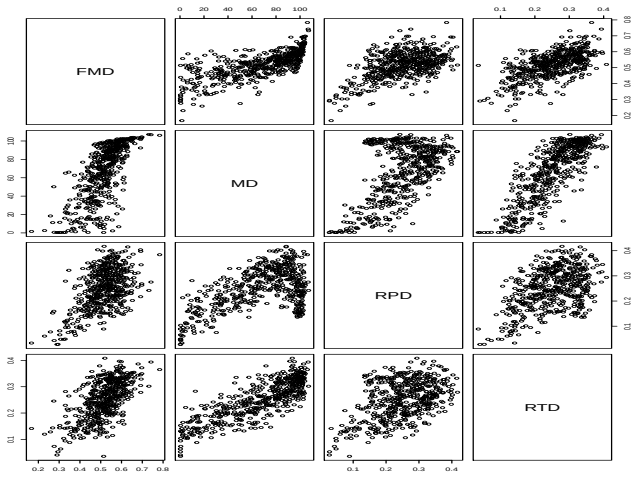

Figura 3.8: Representación comparativa de los cálculos de las profundidades.

Las similitudes y diferencias entre las distintas profundidades quedan fielmente reflejadas en la Figura 3.8, en la que se realiza una comparativa dos a dos. Cuanto más próxima a la diagonal, más semejantes serán los resultados obtenidos por los métodos en cuestión. Se aprecian significativas diferencias y una cierta similitud entre los resultados de la MD con la RTD y la RPD.

3.2.3. Transformaciones de los datos funcionales

Para finalizar con el análisis exploratorio, consideramos la posibilidad de introducir la transformación derivada, $\{G^{(1)}(\mathcal{X}_i)\}_{i=1}^n$, en nuestro estudio. Esta transformación permitirá diferenciar a los clientes que gastan homogéneamente a lo largo de las semanas, lo que se traducirá en un perfil casi constante, de aquellos que presentan ciclos de gasto seguidos de periodos de inactividad, lo que se reflejará en subidas y bajadas más o menos pronunciadas.

En la Figura 3.9, observamos en la primera gráfica la derivada de los datos y en la segunda el correspondiente cálculo de profundidades utilizando la FMD. Vemos que el perfil general, salvo pequeñas subidas y bajadas, es constante; lo cual tiene sentido, pues estamos analizando una muestra del subconjunto de clientes con 10 semanas o menos de inactividad, por lo que tienen una operativa de consumo regular. Si bien es cierto que este resultado se corresponde a una muestra de pequeño tamaño de un grupo de edad en particular; analizando las gráficas de la totalidad de clientes para cada subgrupo, sí se observarían subidas y bajadas un poco más pronunciadas, de ahí que en el análisis de regresión desarrollado en la Sección 3.4 se tenga en cuenta la derivada del dato funcional.

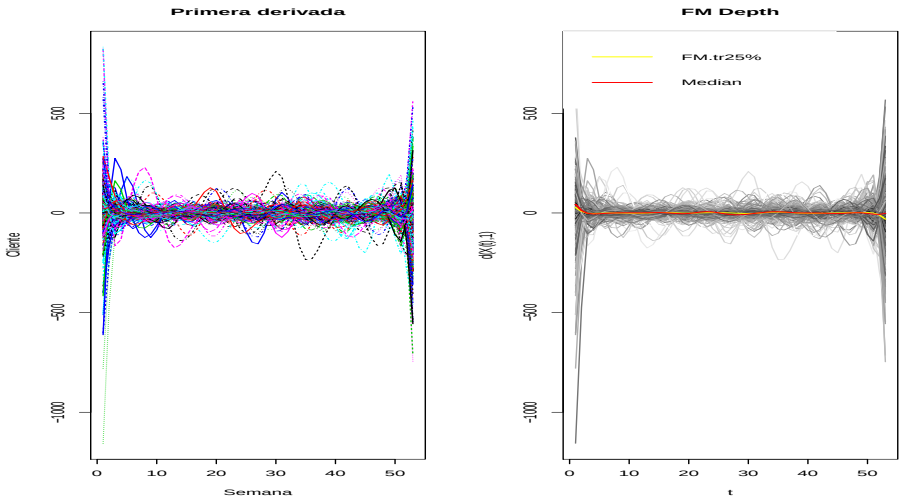

Figura 3.9: Transformación del dato funcional mediante derivada sobre una muestra de 500 clientes del subgrupo jóvenes.

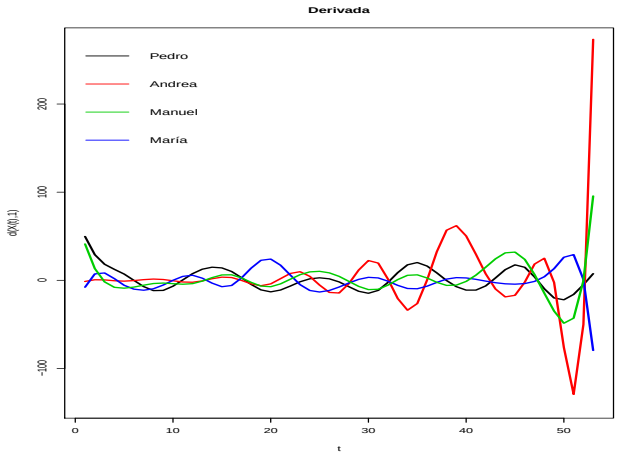

Figura 3.10: Representación de la derivada de la operativa anual de 4 clientes del subgrupo de los jóvenes.

En la Figura 3.10, en la que se observa la derivada del dato funcional correspondientes a los cuatro clientes de ejemplo, vemos que Pedro, Manuel y María tiene un perfil más constante que Andrea, pues como vimos en la Sección 2.1.1, su consumo era más homogéneo, mientras que Andrea concentraba 151 operaciones en tan solo 11 semanas de actividad.

3.3. ANOVA

En base al objetivo de clasificación de clientes, se empleará el test ANOVA para justificar la segmentación del conjunto total de clientes en subgrupos teniendo en cuenta diferentes variables.

Sea una muestra funcional clasificada por una variable factor que la divide en diferentes grupos,

$$\{\mathcal{X}_i, G_i\}_{i=1}^n \in \mathcal{F} \times \mathbb{G} = \{1, \ldots, G\}$$

donde G denota la variable discreta indicadora del grupo de pertenencia de cada observación.

Se quiere contrastar la hipótesis nula de igualdad de medias

$$H_0 : \bar{\mathcal{X}}_1 = \bar{\mathcal{X}}_2 = \cdots = \bar{\mathcal{X}}_G$$

frente a la hipótesis alternativa

$$H_1 : \exists\ k,\ j \quad \text{tal que} \quad \bar{\mathcal{X}}_k \neq \bar{\mathcal{X}}_j$$

siendo el estadístico de contraste el siguiente

$$V_n = \sum_{k<j} n_k ||\bar{\mathcal{X}}_k - \bar{\mathcal{X}}_j||^2.$$

Vemos por tanto, que el ANOVA funcional constituye una reparametrización del test ANOVA clásico multivariante, con la salvedad de que en este caso no se tiene la distribución de referencia como sí ocurre en el caso multivariante.

Se aplicará tanto el ANOVA *One Factor*, considerando cada una de las variables factor que permiten hacer grupos (*segmento_edad*, *sexo*, *localidad* y *actividad*) por separado y también se desarrollará el ANOVA *Multiway*, aplicado sobre varios factores conjuntamente y considerando todas las interacciones posibles entre ellos.

3.3.1. Análisis de comparación intergrupal

Desde el comienzo de este trabajo se ha hablado de la segmentación previa al análisis por la variable *segmento_edad*, imprescindible para que el volumen de datos no fuese un impedimento al realizar el estudio. No obstante, queda por analizar si esta agrupación de clientes está realmente justificada; es decir, si existen diferencias entre jóvenes, adultos 1, adultos 2 y sénior, lo que nos permitiría emplear la variable factor *segmento_edad* en la clasificación de clientes o si por el contrario, sería necesaria una agrupación distinta.

Para ello, se realiza el contraste de igualdad de medias sobre los subgrupos de edad; esto es, el test ANOVA considerando la variable factor *segmento_edad*, siendo la hipótesis nula a contrastar:

$$H_0 : \bar{\mathcal{X}}_J = \bar{\mathcal{X}}_{A1} = \bar{\mathcal{X}}_{A2} = \bar{\mathcal{X}}_S.$$

Dado que nos interesa analizar la totalidad de clientes descrita en el Cuadro 3.3 y que tal volumen de datos es incompatible con la función `anova.onefactor` del entorno `R`, este contraste se desarrollará con sucesivas repeticiones del test sobre muestras de 5 000 clientes. En particular, 1 000 repeticiones ANOVA sobre muestras formadas por 212 clientes del subgrupo jóvenes, 2 771 clientes adultos 1, 1 637 adultos 2 y 379 clientes sénior, de forma que la muestra tenga una composición de clientes proporcional al conjunto original.

En la Figura 3.11 se observa el resultado del test ANOVA sobre una de las muestras de 5 000 clientes. Se obtiene un *p-valor* de 0, lo que permite rechazar la hipótesis nula de igualdad de medias entre los subgrupos de edad, con los niveles de significación usuales, para esa muestra concreta de clientes. Este resultado puede ser generalizado a la totalidad de clientes, pues tras las sucesivas repeticiones del test sobre 1 000 muestras distintas de clientes, en todos los casos se obtiene un *p-valor* de 0. Analizando la primera gráfica, vemos que la media del subgrupo de los clientes más jóvenes es significativamente menor que la de los 3 grupos restantes; no obstante, no se perciben claramente las similitudes o diferencias entre estos últimos.

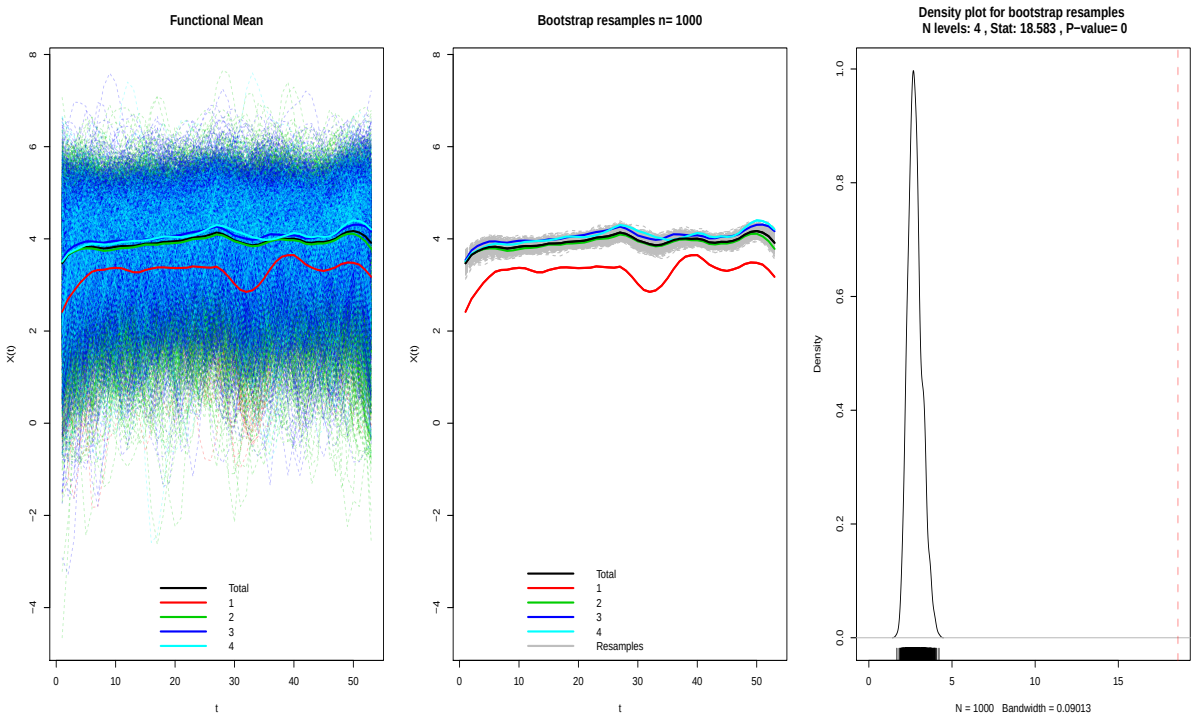

Figura 3.11: ANOVA intergrupal sobre el factor *segmento_edad*: contraste de igualdad de medias entre los subgrupos de edad jóvenes (1), adultos 1 (2), adultos 2 (3) y sénior (4), sobre una muestra de 5 000 clientes.

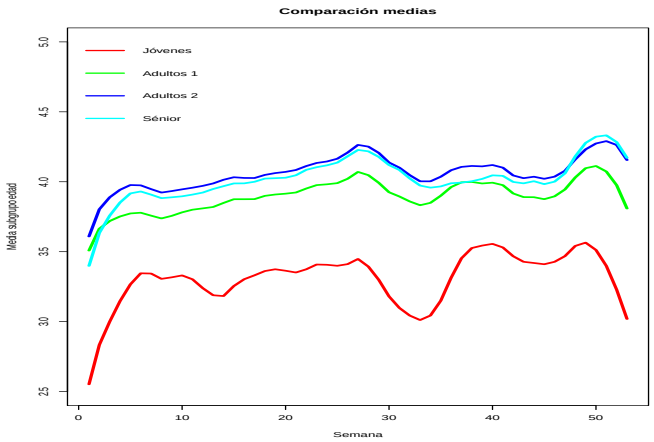

Figura 3.12: Comparación de medias de los subgrupos de edad jóvenes, adultos 1, adultos 2 y sénior.

Por ello, representamos en la Figura 3.12 la media da cada subgrupo de edad. Mediante esta gráfica podemos confirmar la notable diferencia en el gasto medio del grupo jóvenes con respecto al resto, lo cual puede estar relacionado con el hecho de que la *actividad* predominante en este grupo sea la correspondiente a *inactivos estudiantes* mientras que los *activos ocupados*, con unos ingresos fijos mensuales, dominan los 3 grupos de mayor edad.

Llegados a este punto, sería interesante analizar la igualdad de medias entre los subgrupos más semejantes, por lo que comenzaremos por eliminar el subgrupo jóvenes del contraste. Siguiendo un proceso análogo al anterior, realizamos un nuevo test, siendo la hipótesis nula a contrastar

$$H_0 : \bar{\mathcal{X}}_{A1} = \bar{\mathcal{X}}_{A2} = \bar{\mathcal{X}}_S$$

y considerando en esta ocasión 1 000 muestras de 5 000 clientes compuestas por 2 894 adultos 1, 1 710 adultos 2 y 396 sénior.

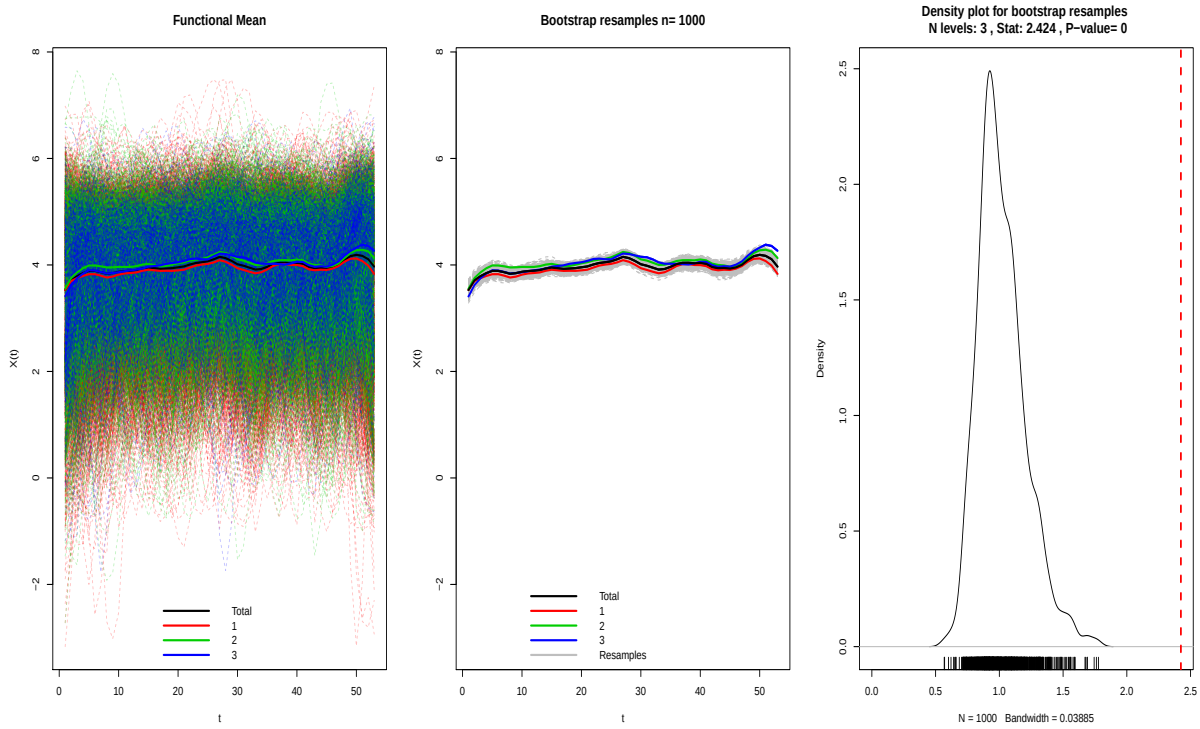

Figura 3.13: ANOVA intergrupal para el contraste de igualdad de medias entre los subgrupos de edad adultos 1 (1), adultos 2 (2) y sénior (3), sobre una muestra de 5 000 clientes.

En la Figura 3.13 se presenta el resultado del test ANOVA sobre una de las muestras de 5 000 clientes analizadas. Se obtiene un *p-valor* de 0, por lo que de nuevo podemos rechazar la hipótesis nula de igualdad de medias entre los 3 subgrupos de mayor edad, con los niveles de significación usuales, para esta muestra concreta de clientes. Tras sucesivas repeticiones del test sobre 1 000 muestras distintas de clientes, en todos los casos se obtiene un *p-valor* de 0, por lo que el resultado es el mismo para la muestra total de clientes adultos 1, adultos 2 y sénior.

En la Figura 3.12 también se refleja la gran similitud entre las medias de los subgrupos adultos 2 y sénior. Será necesario, por tanto, realizar un nuevo contraste con el fin de decidir si la similitud es tal, que ambos grupos pueden ser considerados uno a efectos de clasificación de clientes o si por el contrario, las diferencias son suficientes para clasificarlos de modo distinto. De este modo, contrastamos en este

caso la hipótesis nula

$$H_0 : \bar{\mathcal{X}}_{A2} = \bar{\mathcal{X}}_S$$

sobre muestras de clientes formadas por 4 060 adultos 2 y 940 sénior.

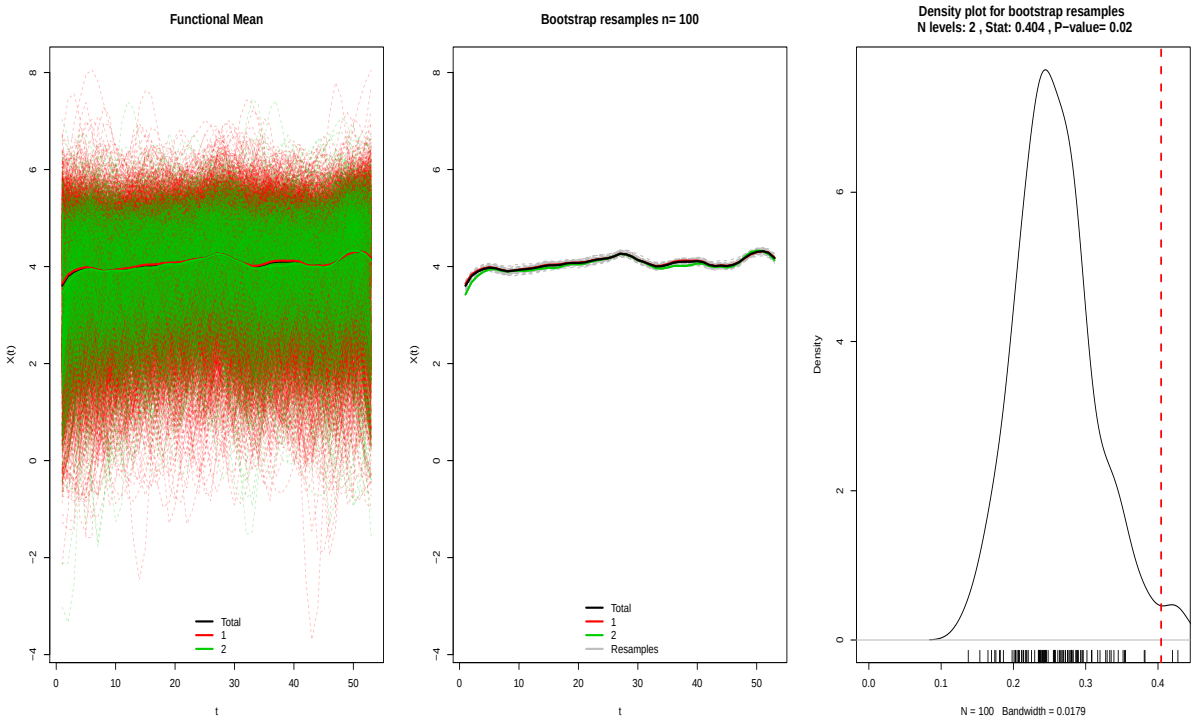

Figura 3.14: ANOVA intergrupal para el contraste de igualdad de medias entre los subgrupos de edad adultos 2 (1) y sénior (2), sobre una muestra de 5 000 clientes.

En la Figura 3.14, correspondiente al resultado gráfico para una de esas muestras, se obtiene un *p-valor* = 0.02, lo que nos permite rechazar la hipótesis nula de igualdad de medias entre los dos subgrupos de edad con un nivel de significación del 5 % y sobre esta muestra en particular. Realizando 1 000 repeticiones del test sobre muestras distintas, en 897 ocasiones se obtiene un *p-valor*<0.05 y tan solo en 103 el *p-valor* obtenido es mayor o igual que 0.05, por lo que, con un nivel de significación del 5 %, también decidimos rechazar la igualdad de medias en el caso de la totalidad de clientes adultos 2 y sénior.

Tras el análisis de comparación entre grupos, podemos concluir que existen diferencias significativas entre los subgrupos de edad, por lo que podremos proseguir con esta segmentación y emplear la variable *segmento_edad* en la clasificación de clientes. No obstante, cabe recordar que existe una decisión previa, asociada a la definición de estos grupos, que puede llevarnos a esta inferencia, por lo que podría haber diferencias entre subgrupos definidos mediante otros alternativos a la clasificación prestablecida por la entidad bancaria, con miras a la aplicación de distintas condiciones y comisiones en los servicios de tarjeta que pone a disposición de cada segmento. También pueden existir otras diferencias dentro de cada uno de estos subgrupos, tal y como veremos a continuación.

3.3.2. Análisis de comparación intragrupal

Una vez justificada la segmentación por edad, analizaremos cada subgrupo por separado. En este apartado se analizarán las variables factor *sexo*, *localidad* y *actividad*, tanto de forma individual (ANOVA *One Factor*) como conjuntamente (ANOVA *Multiway*).

En este caso, el análisis ANOVA será realizado sobre la totalidad de clientes de cada subgrupo de edad (véase Cuadro 3.3); por lo que los resultados numéricos, obtenidos a lo largo de este apartado, serán los correspondientes a realizar el contraste de igualdad de medias sobre la muestra completa, segmentada por la edad.

- ANOVA *One Factor*

  Mediante el ANOVA *One Factor* contrastaremos la igualdad de medias entre los subgrupos creados por las variables *sexo*, *localidad* y *actividad* individualmente.

  - Variable factor *sexo*:

    Dentro de cada subgrupo de edad realizaremos el contraste de igualdad de medias respecto a la variable *sexo* testeando la hipótesis nula

    $$H_0 : \bar{\mathcal{X}}_M = \bar{\mathcal{X}}_H$$

    donde M denota el subgrupo femenino y H el masculino.

    Por poner un ejemplo, mostramos en la Figura 3.15 el resultado obtenido al llevar a cabo el contraste sobre el subgrupo de adultos 1. El $p\text{-}valor = 0$ obtenido, indica que podemos rechazar la hipótesis de igualdad de medias entre el subgrupo femenino y el masculino con los niveles de significación usuales y para este subgrupo de edad; es decir, la variable factor *sexo* permitirá clasificar distintamente a mujeres y hombres del primer grupo de adultos.

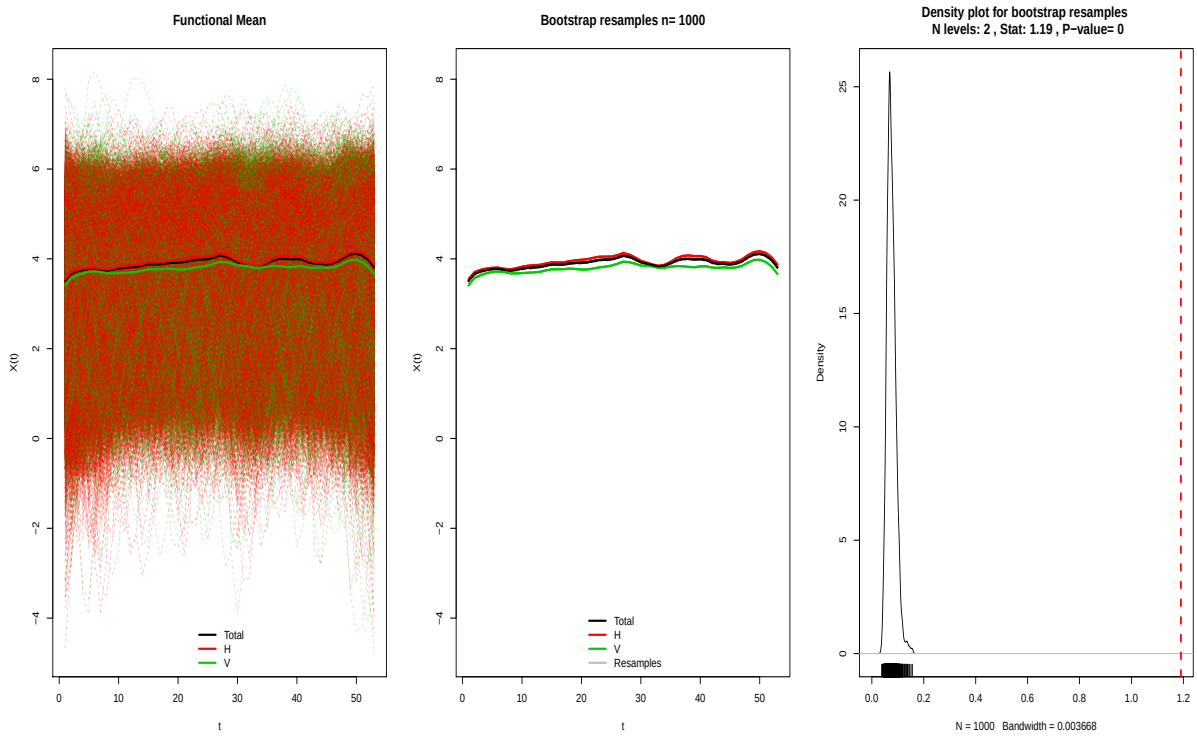

Figura 3.15: ANOVA para el contraste de igualdad de medias entre el subgrupo de mujeres (H) y el de hombres (V) para el subgrupo de edad de adultos 1.

- Variable factor *localidad*

  En este caso se realiza el test ANOVA sobre los distintos subgrupos creados a partir de la variable *localidad* testeando la hipótesis nula

  $$H_0 : \bar{\mathcal{X}}_C = \bar{\mathcal{X}}_L = \bar{\mathcal{X}}_{OU} = \bar{\mathcal{X}}_P = \bar{\mathcal{X}}_R$$

  donde los subíndices denotan los subgrupos de clientes con compras en A Coruña, Lugo, Ourense, Pontevedra y el resto de España, respectivamente.

  Mencionar que en el caso de la variable *localidad* no ocurre lo mismo que para *sexo* y *actividad*, en las que un cliente está asociado a un único factor. Un cliente, independientemente de su lugar de residencia, puede efectúar compras en cualquier zona, por lo que decidimos modificar la variable *localidad* de forma que a cada cliente se le asocie la zona en la que ha realizado un mayor número de transacciones.

  En la Figura 3.16 se presenta el resultado de este contraste sobre el subgrupo de adultos 1. Mediante el *p-valor* $= 0$ obtenido, podemos rechazar la hipótesis de igualdad de medias entre los subgrupos de clientes según la localidad con los niveles de significación usuales y para este subgrupo de edad; esto es, la variable factor *localidad* se puede emplear en la clasificación de clientes en el segundo grupo de adultos.

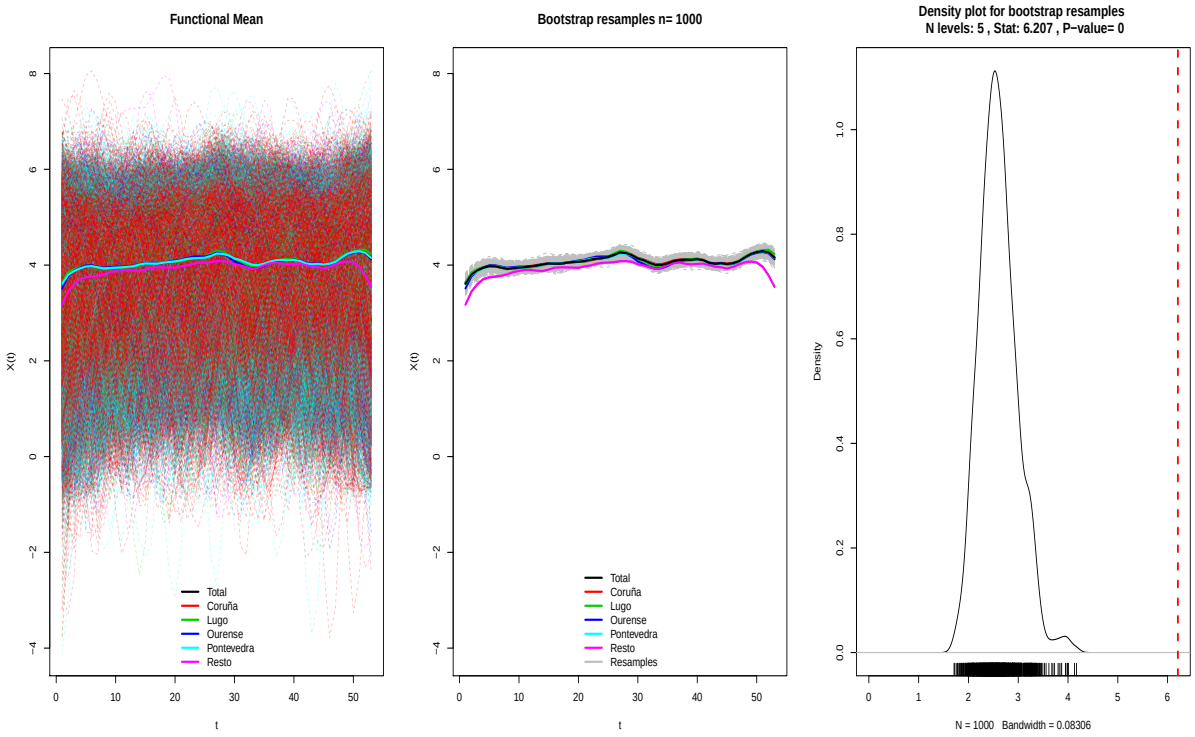

Figura 3.16: ANOVA para el contraste de igualdad de medias entre los subgrupos de clientes de A Coruña, Lugo, Ourense, Pontevedra y Resto de España para el subgrupo de edad de adultos 2.

A pesar de no que no se aprecia en esta gráfica, tras este análisis se pudo comprobar que independientemente del subgrupo de edad, el mayor gasto se realiza en las localidades de las provincias de A Coruña y Pontevedra; mientras que el interior de la comunidad autónoma presenta un menor gasto medio con tarjeta, reflejando de manera coherente las diferencias de urbanización y ruralización del territorio.

- Variable factor *actividad*

  En cuanto al contraste sobre los subgrupos creados por la variable *actividad* debemos testear la hipótesis nula

  $$H_0 : \bar{\mathcal{X}}_{OC} = \bar{\mathcal{X}}_{PA} = \bar{\mathcal{X}}_{IN} = \bar{\mathcal{X}}_{OT}$$

  donde los subíndices hacen referencia a los subgrupos de clientes Activos ocupados, Activos parados, Inactivos Estudiantes/Amas de casa/Pensionistas y Ocupaciones no bien especificadas, respectivamente.

  A modo de ejemplo, mostramos en la Figura 3.17 el resultado de aplicar el test ANOVA sobre el subgrupo sénior. El $p\text{-}valor = 0.14$ obtenido en este caso, indica que, con los niveles de significación usuales, no rechazamos la hipótesis nula de igualdad de medias entre los subgrupos por actividad para este subgrupo de edad.

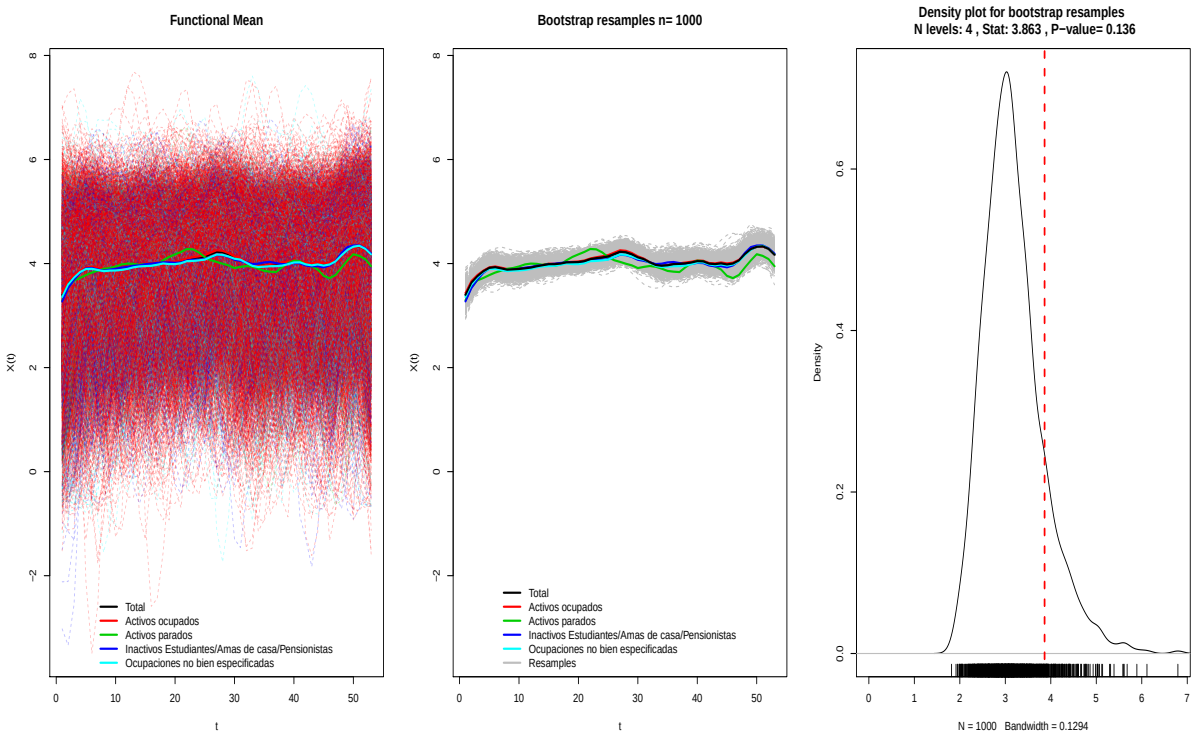

Figura 3.17: ANOVA para el contraste de igualdad de medias entre los subgrupos de Activos Ocupados, Activos Parados, Inactivos Estudiantes/Amas de casa/Pensionistas y Ocupaciones no bien especificadas para el subgrupo de edad sénior.

SEGMENTO	EDAD	SEXO	LOCALIDAD	ACTIVIDAD
Jóvenes	18-25 años	**0**	**0**	0.33
Adultos 1	26-45 años	**0**	**0**	**0**
Adultos 2	46-59 años	**0**	**0**	**0**
Sénior	60-65 años	**0**	0.08	0.14

Cuadro 3.4: *p-valores* obtenidos en el test ANOVA *One Factor* para el contraste de igualdad de medias entre los subgrupos creados por las variables *sexo*, *localidad* y *actividad*. En negrita los *p-valores* correspondientes a los contrastes significativos considerando un nivel de significación del 5 %.

En el Cuadro 3.4 se resumen los resultados obtenidos tras aplicar el test ANOVA *One factor* sobre los diferentes subgrupos de edad y para cada una de las variables factor consideradas. A la vista de los resultados y con un nivel de significación del 5 %, podemos concluir con el rechazo de la igualdad de medias en cuanto a *sexo* y *localidad* en el grupo de jóvenes, por las 3 variables en ambos grupos de adultos y tan solo por la variable *sexo* entre los clientes sénior.

No obstante, también nos interesa conocer el efecto de la interacción entre las variables factor, con el objetivo de analizar y clasificar al cliente de una forma lo más exacta posible, en función de varias características conjuntamente y no una sola.

- ANOVA *Multiway*

  Emplearemos la función `anova.RPm` de `R` para conocer el efecto que tiene la interacción entre las variables que venimos de analizar de forma individual. Esta función proporciona el *p-valor* obtenido tras realizar el contraste considerando tanto las variables separadamente como su interacción; no obstante, en el Cuadro 3.5 nos limitamos a presentar los resultados de la interacción, puesto que aunque los resultados numéricos al considerar las variables por separado no son idénticos a los del Cuadro 3.4, sí nos permiten obtener las mismas conclusiones.

SEGMENTO	SEXO:LOC	SEXO:ACT	LOC:ACT	SEXO:LOC:ACT
Jóvenes	**0.03**	0.78	0.43	0.41
Adultos 1	**0**	0.30	0.63	0.35
Adultos 2	**0**	0.76	0.28	0.54
Sénior	0.48	0.64	0.94	0.47

Cuadro 3.5: *p-valores* obtenidos en el test ANOVA *Multiway* para el contraste de igualdad de medias entre los subgrupos creados por la interacción de las variables *sexo*, *localidad* y *actividad*. En negrita los *p-valores* correspondientes a los contrastes significativos considerando un nivel de significación del 5 %.

Observamos en el Cuadro 3.5 que la única interacción significativa, considerando un nivel de significación del 5 %, es la interacción *sexo:localidad* para los subgrupos de clientes jóvenes, adultos 1 y adultos 2, por lo que la igualdad de medias entre los distintos subgrupos por *sexo* y *localidad* podrá ser rechazada para estos tres segmentos de edad.

No nos sorprende que la interacción con la variable *actividad* no sea significativa, puesto que los subgrupos de edad suelen estar dominados por una determinada actividad contable, tal y como se deduce del Cuadro 3.6, en el que observamos que 53 % de los jóvenes son inactivos estudiantes y el 52 %, 67 % y 69 % de los adultos 1, adultos 2 y sénior, respectivamente, son activos ocupados.

SEGMENTO	INACTIVOS	OCUPADOS	PARADOS	SE DESCONOCE
Jóvenes	53 %	35 %	7 %	5 %
Adultos 1	12 %	52 %	20 %	16 %
Adultos 2	8 %	67 %	7 %	18 %
Sénior	13 %	69 %	1 %	17 %

Cuadro 3.6: Tabla de porcentajes para los factores de la variable *actividad* para cada subgrupo de edad.

Un análisis análogo a este sería contrastar la igualdad de medias entre las distintas combinaciones de variables mediante el ANOVA *One Factor*. Así, vemos por ejemplo el resultado del contraste sobre los subgrupos de hombres jóvenes por localidad en la Figura 3.18, en sintonía con los resultados del Cuadro 3.5. En la Figura 3.19 mostramos una gráfica comparativa del gasto medio para las distintas agrupaciones por *sexo* y *localidad* para cada subgrupo de edad. Comprobamos que efectivamente, las menores discrepancias se obtienen en el grupo de clientes sénior. Se observa que de forma general las mujeres gastan significativamente más que los hombres de su mismo segmento de edad.

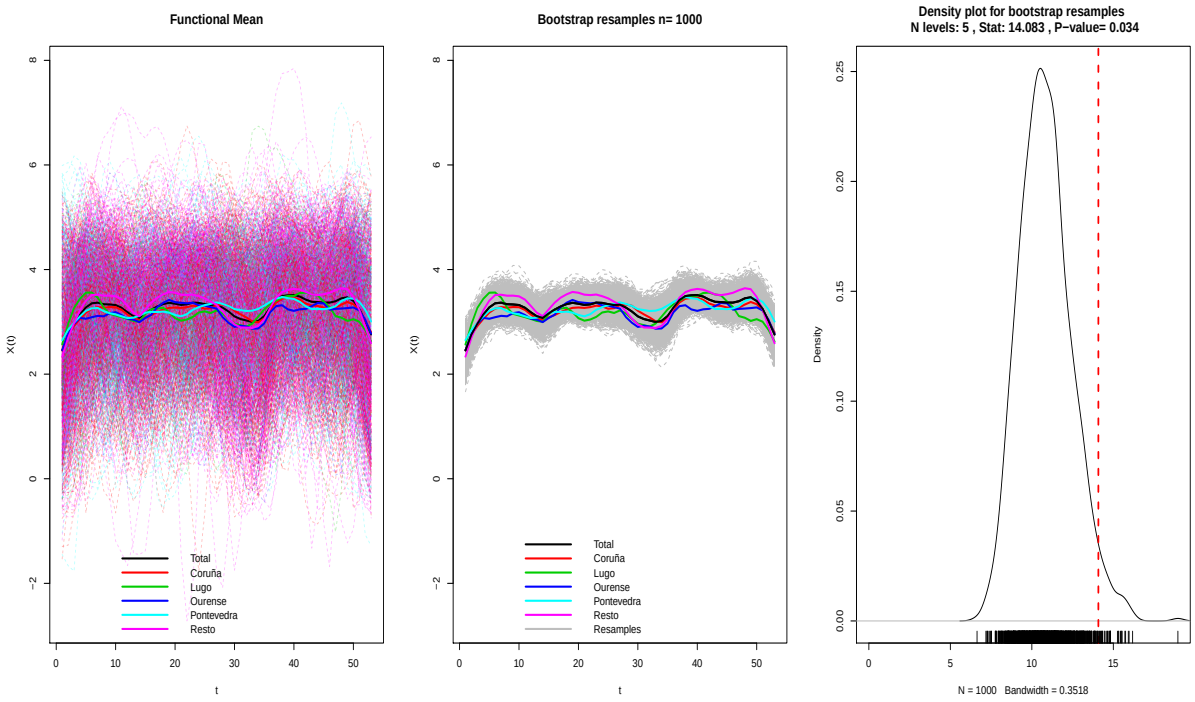

Figura 3.18: ANOVA para el contraste de igualdad de medias entre los subgrupos por localidad de los varones jóvenes.

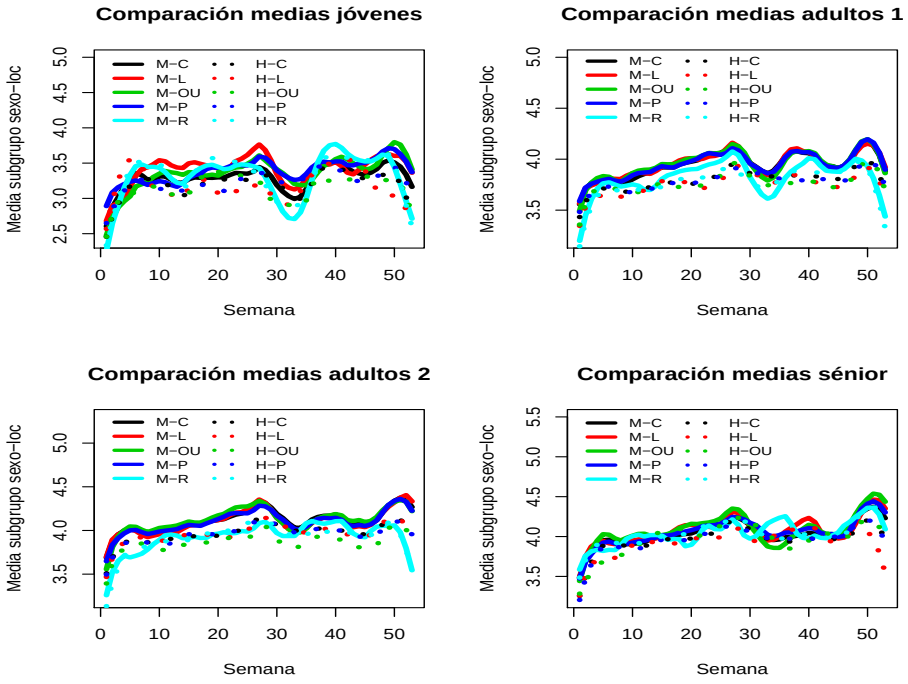

Figura 3.19: Comparación de medias de los subgrupos por las variables *sexo* y *localidad*.

3.3.3. Conclusiones ANOVA

A la vista de los resultados obtenidos a lo largo de esta sección, concluimos con la segmentación de clientes jóvenes, adultos 1 y adultos 2 por las variables *sexo* y *localidad* y de los clientes sénior por la variable *sexo*. Esto nos permite obtener una herramienta útil para la clasificación de clientes en función de sus características particulares, más concreta en el caso de los 3 primeros segmentos de edad.

En el Cuadro 3.7 se muestran los tamaños muestrales de los diferentes subgrupos formados dentro del grupo de clientes jóvenes en base a la segmentación que venimos de justificar.

JÓVENES	A Coruña	Lugo	Ourense	Pontevedra	Resto	TOTAL
Hombre	442	38	39	247	452	1 218
Mujer	1 085	130	145	655	554	2 569
TOTAL	1 527	168	184	902	1 006	3 787

Cuadro 3.7: Tamaños muestrales de los subgrupos tras la segmentación por *sexo* y *localidad* en el grupo de clientes jóvenes.

En los Cuadros 3.8 y 3.9 tenemos los tamaños muestrales de los subgrupos obtenidos tras la segmentación por *sexo* y *localidad* del grupo de adultos 1 y adultos 2, respectivamente.

ADULTOS 1	A Coruña	Lugo	Ourense	Pontevedra	Resto	TOTAL
Hombre	6 664	883	804	4 950	1 563	14 864
Mujer	15 868	2 726	2 488	12 251	1 266	34 599
TOTAL	22 532	3 609	3 292	17 201	2 829	49 463

Cuadro 3.8: Tamaños muestrales de los subgrupos tras la segmentación por las variables *sexo* y *localidad* en el grupo de clientes adultos 1.

ADULTOS 2	A Coruña	Lugo	Ourense	Pontevedra	Resto	TOTAL
Hombre	4 132	609	572	2 800	299	8 412
Mujer	10 000	1 824	1 599	7 224	167	20 814
TOTAL	14 132	2 433	2 171	10 024	466	29 266

Cuadro 3.9: Tamaños muestrales de los subgrupos tras la segmentación por las variables *sexo* y *localidad* en el grupo de clientes adultos 2.

Para finalizar, el Cuadro 3.10 muestra los subgrupos creados en el grupo de clientes sénior por la variable *sexo* y sus respectivos tamaños muestrales.

	SÉNIOR
Hombre	2 203
Mujer	4 565
TOTAL	6 768

Cuadro 3.10: Tamaños muestrales de los subgrupos tras la segmentación por la variable *sexo* en el grupo de clientes sénior.

Los subgrupos a los que se hace mención en los Cuadros 3.7, 3.8, 3.9 y 3.10 serán aquellos sobre los que se realice el estudio de regresión en la Sección 3.4.

3.4. Modelos de regresión y predicción

En la Sección 2.2.3 se realizó una revisión de los principales modelos de regresión funcional. En este apartado se llevará a cabo el ajuste de aquellos más adaptados a las características de nuestros datos.

En la Figura 3.20 se observa el consumo medio por subgrupo de edad. Se puede observar un pico de gasto próximo a San Valentín (semana 7), otro en torno a la semana 27 (primera de julio) coincidiendo con el comienzo de la temporada de verano, una subida menos pronunciada entre las semanas 36 y 40 (mes de septiembre) diferente según el subgrupo de edad y que puede guardar relación con la "vuelta al cole" y ya por último, el incremento que se produce desde la semana del Black Friday (semana 48) hasta la primera semana de diciembre (semana 49) en el caso de los jóvenes, hasta la semana 50 en el caso de adultos 1, y que se prolonga hasta la semana anterior a nochebuena (semana 51) en los subgrupos de clientes adultos 2 y sénior. En cuanto a la disminución del consumo, los descensos más importantes se producen en el grupo de los jóvenes durante Semana Santa (semana 14) y entre las semanas 28 y 36 coincidiendo con verano, lo que puede estar estrechamente ligado con que el subgrupo de jóvenes está dominado en un 53 % por estudiantes que verán reducidos sus gastos en vacaciones, sobre todo aquellos que residan fuera del domicilio familiar durante el curso académico. Del mismo modo, en el período de navidad considerado no se observa un consumo importante en el grupo de los jóvenes, aunque si vemos que durante el Black Friday tiene lugar un aumento significativo del gasto.

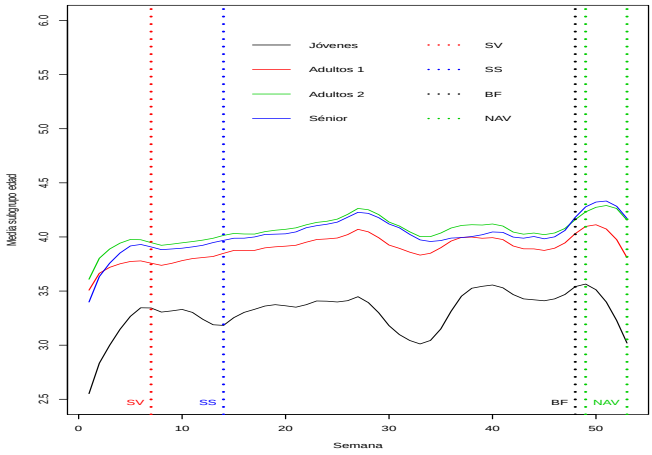

Figura 3.20: Representación del gasto medio de cada subgrupo de edad a lo largo del año 2015 y acontecimientos a analizar: San Valentín, Semana Santa, Black Friday y Navidad.

Estas oscilaciones del gasto son conocidas a nivel general, puesto que responden a fechas señaladas por los departamentos de marketing para el consumo minorista, especialmente en las industrias de distribución. Menos conocido y de mayor interés para una entidad bancaria son las funciones por las cuales se pueda tratar de predecir, con un cierto nivel de confianza, esas oscilaciones, en las cuales existe el interés de intervenir para potenciar el gasto mediante ofertas de financiación, descuentos exclusivos e incentivos para el uso de tarjetas.

Nos interesa analizar el gasto realizado en cuatro fechas importantes en lo que a consumo se refiere: San Valentín, Semana Santa, Black Friday y Navidad. En la Figura 3.20 se señalan aquellos acontecimientos sobre los que se realizará el estudio de regresión.

De este modo, la variable respuesta y se corresponderá con los importes, transformados por logaritmos, de las semanas 7, 14 y 48, en el caso de San Valentín, Semana Santa y Black Friday, respectivamente, y con el promedio de las semanas de la 49 a la 53 en el estudio de la Navidad. La variable funcional $\mathcal{X}$ coincidirá, en cada caso, con el dato funcional creado a partir de las restantes semanas y también se tendrá en cuenta la covariable correspondiente a la primera derivada de los datos, $G^{(1)}(\mathcal{X})$.

De entre los mecanismos para la estimación de β comentados en el apartado de modelos lineales de la Sección 2.2.3, nos decantamos por el modelo de componentes principales funcionales (FPC) y el de mínimos cuadrados parciales (FPLS), puesto que tras varias comprobaciones observamos un mejor ajuste por parte de estos modelos frente a la representación en bases de Fourier, Wavelets o B-splines, método menos sistemático y más caro computacionalmente. En el FPC, el número de componentes principales óptimo será elegido por el criterio de selección MSC (*Model Selection Criteria*) empleando la penalización SIC (*Schwartz Information Criterion*), mientras que en el caso del FPLS fijaremos a 4 el número de componentes a utilizar.

Finalizando con los modelos lineales, en tercer lugar se ajustará un FLM incluyendo la derivada de los datos, $G^{(1)}(\mathcal{X})$ (analizada en la Sección 3.2.3), como covariable. El modelo resulta:

$$\mathbb{E}(y|\mathcal{X}) = \sum_j \langle \mathcal{X}^j, \beta_j \rangle + \sum_k \left\langle G^{(1)}(\mathcal{X}^k), \beta_k \right\rangle + \epsilon \tag{3.1}$$

para el que emplearemos 4 componentes principales tanto en la representación de $\mathcal{X}$ como en la de $G^{(1)}(\mathcal{X})$.

De entre los modelos generalizados, una opción a tener en cuenta podría ser el FGKAM; no obstante, el uso del método no paramétrico funcional lleva asociado el cálculo de distancias, lo que se traduce en mayores necesidades computacionales si lo comparamos con el ajuste de un FGSAM, más adaptado a nuestros datos. De ahí que nos decantemos por este último, en el que también tendremos en cuenta la derivada de los datos, representado ambos (dato funcional y derivada) mediante 4 componentes principales.

A continuación, ajustaremos los cuatro modelos propuestos para analizar el gasto en las cuatro fechas señaladas: San Valentín, Semana Santa, Black Friday y Navidad:

Modelo 1: Modelo lineal funcional de componentes principales (FPC)
Modelo 2: Modelo lineal funcional de mínimos cuadrados parciales (FPLS)
Modelo 3: Modelo lineal funcional con derivada (FLM)
Modelo 4: Modelo aditivo espectral generalizado funcional con derivada (FGSAM)

En la práctica, este análisis ha sido realizado sobre la totalidad de subgrupos; no obstante, dado el gran número de segmentaciones creadas, se mostrará el análisis de regresión al completo para un único subgrupo por acontecimiento, resumiendo los resultados más relevantes de los modelos de los restantes subgrupos al final de cada apartado.

De este modo, analizaremos en detalle el gasto de los clientes jóvenes de A Coruña en San Valentín, el consumo de las clientas de Pontevedra entre 26 y 45 años durante Semana Santa, el importe de las transacciones realizadas en el Black Friday por los clientes lucenses de entre 46 y 59 años y por último el consumo de las clientas sénior durante Navidad.

3.4.1. San Valentín

El primer acontecimiento a estudiar en el análisis de regresión es San Valentín (14 de febrero), correspondiente a la séptima semana del año 2015. El subgrupo analizado en este caso es el de los varones coruñeses de entre 18 y 25 años. De los 442 clientes que conforman la muestra, el 85 % será empleado para realizar el ajuste de regresión y el 15 % restante para validar el modelo mediante predicción.

Empezamos el análisis de regresión por ajustar el modelo funcional de componentes principales (FPC), mostrando en el Cuadro 3.11 los resultados numéricos obtenidos tras realizar dicho ajuste. Observamos que el número de componentes principales elegidas por el Criterio de Información de Schwartz es de 8, las cuales permiten explicar el 71.83 % de la variabilidad, mientras que el coeficiente R^2=0.64 obtenido es relativamente alto.

MODELO 1	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.271	0.029	109.994	$< 10^{-3}$
PC9	0.317	0.026	11.845	$< 10^{-3}$
PC7	0.264	0.024	10.773	$< 10^{-3}$
PC2	-0.189	0.018	-10.079	$< 10^{-3}$
PC8	-0.286	0.026	-10.908	$< 10^{-3}$
PC1	0.087	0.013	6.696	$< 10^{-3}$
PC3	-0.152	0.021	-7.528	$< 10^{-3}$
PC6	0.128	0.024	5.302	$< 10^{-3}$
PC4	0.118	0.022	5.252	$< 10^{-3}$
$R^2 = 0.641$	71.83 % variabilidad explicada con 8 PC's			$\lambda = 0.5$
Varianza residual: 0.332 en 367 grados de libertad				

Cuadro 3.11: Resultados numéricos del modelo 1 en el análisis del gasto en San Valentín.

En cuanto a la estimación del parámetro β en el modelo 1, observamos en la Figura 3.21 que el consumo en San Valentín de los jóvenes coruñeses viene marcado por un elevado gasto en las semanas próximas al evento y la influencia, algo menor, del consumo de las semanas 22 y 39. A su vez, se relaciona con un bajo consumo en la semana 16 y en el período navideño.

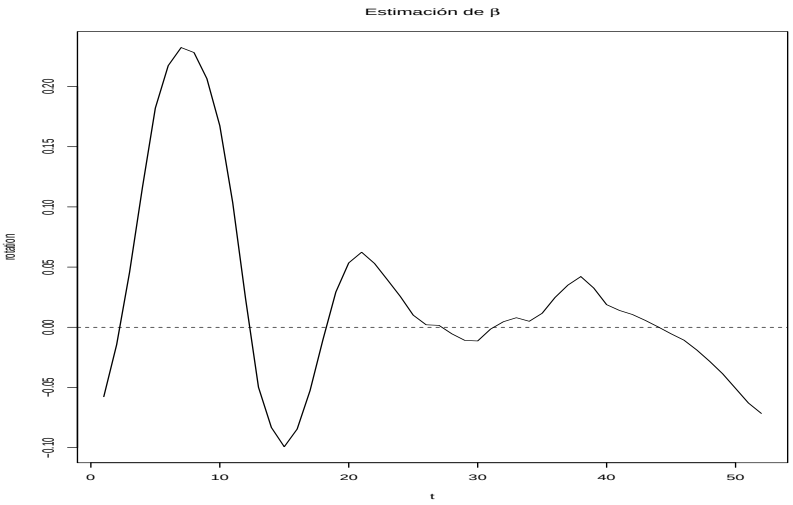

Figura 3.21: Estimación del parámetro de regresión β por el modelo 1 (San Valentín).

En la Figura 3.22, se representan las salidas gráficas habituales en este tipo de métodos. Se muestran, de izquierda a derecha, el ajuste lineal con el coeficiente de determinanción R^2, un gráfico que permite analizar si los residuos poseen media nula y otro para la homocedasticidad de los mismos. En la fila inferior, los *leverage*'s o apalancamientos, el gráfico de normalidad de los residuos y su *box-plot*. A la vista de ellas, en este primer modelo no existen grandes problemas, a excepción de la presencia de una curva influyente y de algún dato funcional atípico, pero sin mayor complicación, por lo que podría ser empleado para la explicación de la respuesta. No obstante, proseguiremos con el ajuste de los restantes modelos, con el objetivo de analizar cual de ellos es el más adecuado para realizar la predicción del gasto en San Valentín.

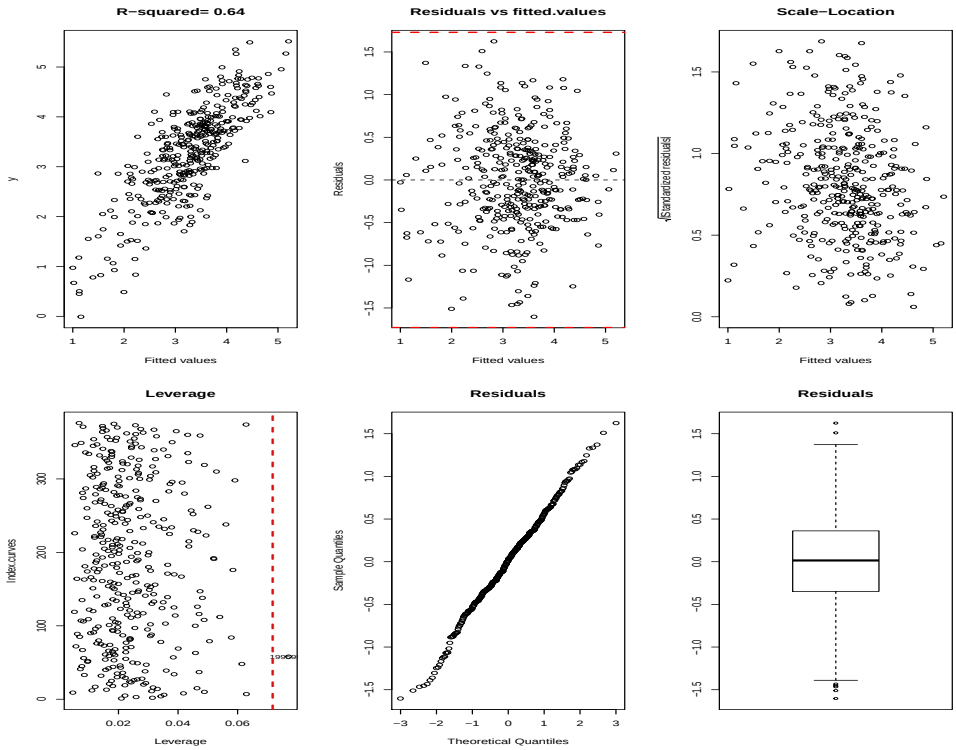

Figura 3.22: Gráficas del modelo 1 para el análisis de gasto en San Valentín. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

En segundo lugar, ajustamos el modelo funcional de mínimos cuadrados parciales (FPLS). A la vista del Cuadro 3.12, vemos que las 4 componentes empleadas resultan significativas, obteniendo un coeficiente R^2=0.93, mayor que en el caso anterior. Observando las gráficas de la Figura 3.23, cabe destacar la presencia de alguna curva atípica e influyente; no obstante, salvo este, no se observan mayores inconvenientes en dicho ajuste.

MODELO 2	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.271	0.013	250.376	$< 10^{-3}$
PLS1	0.314	0.005	57.271	$< 10^{-3}$
PLS2	0.223	0.006	35.585	$< 10^{-3}$
PLS3	0.147	0.007	19.024	$< 10^{-3}$
PLS4	0.146	0.011	13.050	$< 10^{-3}$
$R^2 = 0.932$	Varianza residual: 0.064 en 357.958 grados de libertad			

Cuadro 3.12: Resultados numéricos del modelo 2 en el análisis del gasto en San Valentín.

En lo referente al parámetro β del modelo 2, comprobamos a través de su representación en la Figura 3.24 que el consumo en San Valentín viene determinado por un alto consumo en las semanas más cercanas a esta fecha y asociado a un bajo consumo en las semanas 2 y 12.

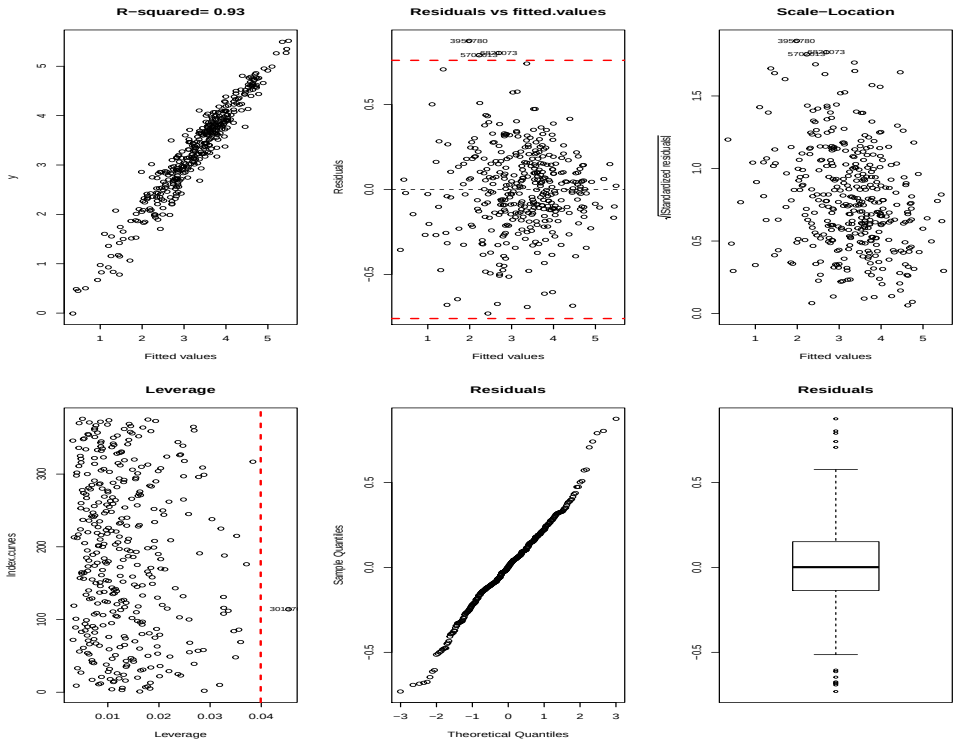

Figura 3.23: Gráficas del modelo 2 para el análisis de gasto en San Valentín. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

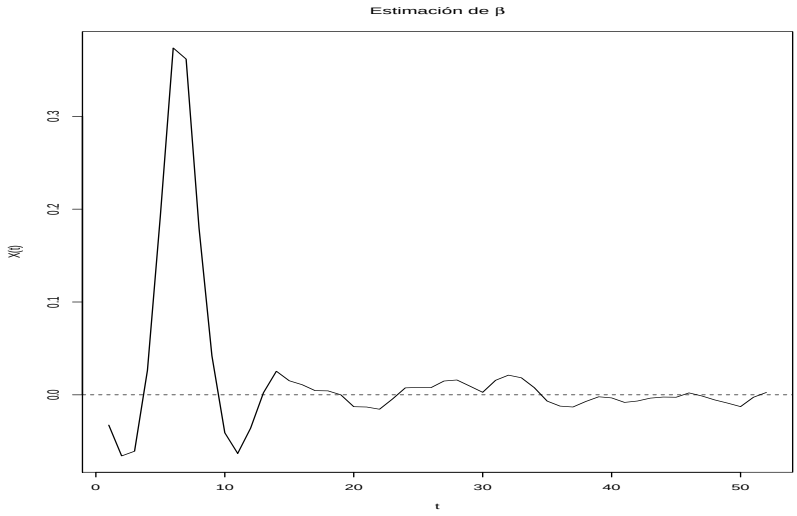

Figura 3.24: Estimación del parámetro de regresión β por el modelo 2 (San Valentín).

El modelo 3 propuesto hemos decidido incorporar la derivada del dato funcional como covariable. Para ello, utilizamos 4 componentes principales tanto en la representación del dato funcional como en la de su derivada. En el Cuadro 3.13 mostramos un resumen numérico de dicho modelo, mediante el que comprobamos que con excepción de la cuarta componente en el caso del dato funcional, las demás

resultan ser significativas. Además, se obtiene un coeficiente R^2 ajustado de 0.53, inferior al de los modelos anteriores.

MODELO 3	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.271	0.03	96.609	$< 10^{-3}$
X.PC1	0.088	0.013	8.788	$< 10^{-3}$
X.PC2	-0.146	0.014	-10.252	$< 10^{-3}$
X.PC3	-0.056	0.016	-3.342	$< 10^{-3}$
X.PC4	0.016	0.019	0.879	0.3791
X1.PC1	0.559	0.055	10.138	$< 10^{-3}$
X1.PC2	-0.193	0.037	-5.155	$< 10^{-3}$
X1.PC3	-0.217	0.043	-5.002	$< 10^{-3}$
X1.PC4	0.383	0.048	7.839	$< 10^{-3}$
$R^2 = 0.535$	R^2 ajustado $= 0.525$		*p-valor*: $< 10^{-3}$	
Error estándar residual: 0.656 en 367 DF			Estad. F: 52.84 en 8 y 367 DF	

Cuadro 3.13: Resultados numéricos del modelo 3 en el análisis del gasto en San Valentín.

En la Figura 3.25, se representan, de izquierda a derecha, una gráfica de media nula para los residuos, una de normalidad y otra de homocedasticidad, mediante las que detectamos un leve incumplimiento de las dos últimas hipótesis. En la fila inferior, gráficas relacionados con la distancia de Cook y los *leverage*'s o apalancamientos. Mediante estas últimas, detectamos la presencia de varias curvas influyentes.

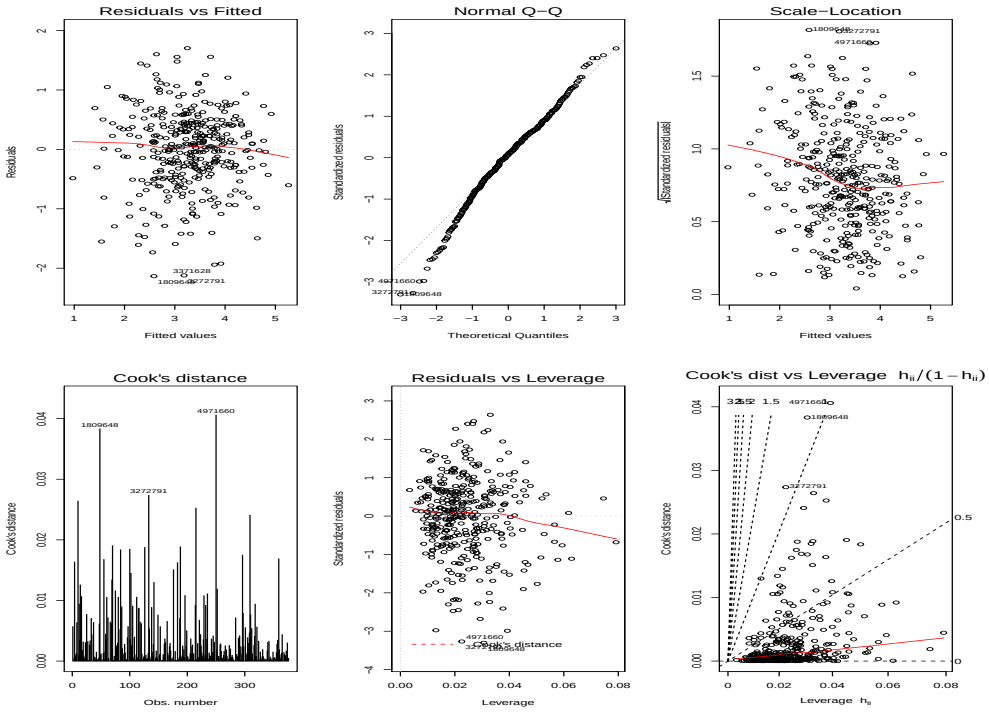

Figura 3.25: Gráficas del modelo 3 para el análisis de gasto en San Valentín. De izquierda a derecha: media cero, gráfico de normalidad Q-Q, homocedasticidad, distancia de Cook y residuos y distancia de Cook frente a apalancamientos.

Las estimaciones asociadas a este ajuste se representan en la Figura 3.26. La curva roja se corresponde a los parámetros $\hat{\beta}_j$ (estimación en los datos originales) y la verde hace referencia a los $\hat{\beta}_k$ (derivada) de la ecuacion (3.1). Estos resultados se corresponden al modelo resultante de eliminar la componente no significativa.

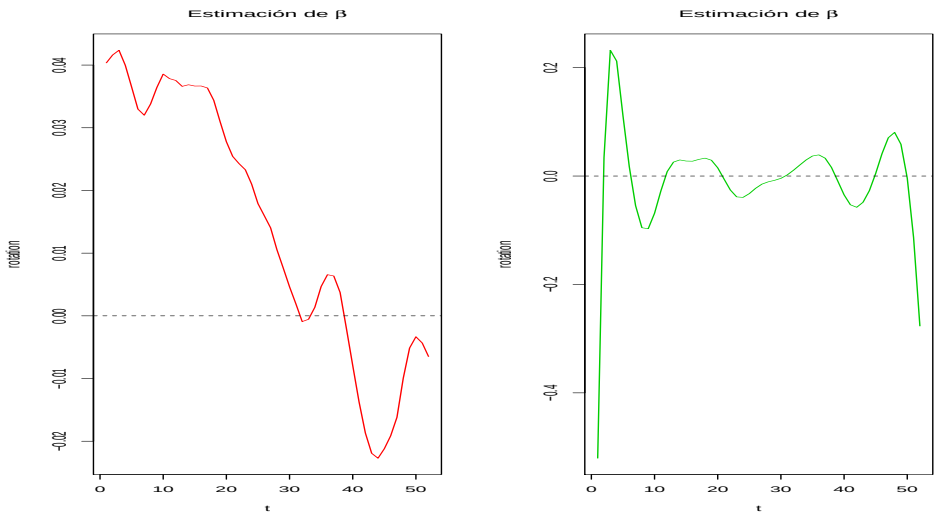

Figura 3.26: Estimación del parámetro de regresión β por el modelo 3 (izquierda) y estimación correspondiente a la derivada (derecha) en el análisis de gasto de San Valentín.

Llegados a este punto, decidimos estudiar la influencia de la variable *actividad*. A pesar de que en su momento se justificó su no utilización en la segmentación, puede ser que dentro de determinados subgrupos de edad, sexo y localidad tenga cierto efecto. Analizando la muestra de entrenamiento, comprobamos que de los 376 clientes que la forman, 152 son inactivos estudiantes, 152 activos ocupados, 48 activos parados y de 24 de ellos se desconoce la actividad que desempeñan. Dado que los inactivos y ocupados se corresponden con los grupos mayoritarios, realizamos un ajuste individual análogo al del modelo 3 para cada subconjunto de clientes frente a la muestra restante en cada caso.

MODELO 3	$\mathcal{X}$	$\mathcal{X}^{(1)}$	R^2	R^2 ajust.	*p-valor*
INACTIVOS	1-2-4	1:4	0.561	0.543	$< 10^{-3}$
- INACTIVOS	1:2	1:3	0.473	0.461	$< 10^{-3}$
OCUPADOS	1:4	1:4	0.563	0.541	$< 10^{-3}$
- OCUPADOS	1-2-4	1:4	0.611	0.599	$< 10^{-3}$

Cuadro 3.14: Resumen del modelo 3 para el análisis de gasto en San Valentín de los clientes inactivos y activos ocupados (de forma separada) frente al resto.

Se muestra en el Cuadro 3.14 un resumen de las principales características de estos modelos: número de componentes empleadas en la representación del dato funcional y su derivada y los coeficientes R^2 y *p-valores* obtenidos. Si comparamos estos resultados con los del modelo 3, en el que tras eliminar la cuarta componente no significativa obtenemos un coeficiente R^2, y su análogo ajustado, igual a 0.53, observamos que los modelos aplicados sobre las muestras para una actividad, son ligeramente mejores. También observamos ciertas diferencias entre estos ajustes y los correspondientes a la muestra restante en cada caso. Por lo tanto, podemos considerar la opción de clasificar a los jóvenes coruñeses por su actividad a la hora de predecir su consumo durante San Valentín.

En el modelo 4, se expresa el predictor lineal en función de términos suaves tal y como se observa en el Cuadro 3.15. Al igual que en el modelo anterior, resultan todos significativos salvo el término relativo a la cuarta componente principal del dato funcional. Cabe salientar que salvo la segunda componente del dato y la primera y la cuarta de su derivada, las demás tienen un efecto lineal. Tras este ajuste se obtiene un coeficiente R^2=0.53, siendo la *deviance* explicada del 54.7 %.

MODELO 4	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.271	0.033	97.42	$< 10^{-3}$
	EDF	Ref. DF	F	p-value
s(X.PC1)	1.000	1.000	78.38	$< 10^{-3}$
s(X.PC2)	1.715	2.179	48.27	$< 10^{-3}$
s(X.PC3)	1.000	1.000	10.37	$< 10^{-3}$
s(X.PC4)	1.000	1.000	21.05	0.3061
s(X1.PC1)	3.199	4.041	26.74	$< 10^{-3}$
s(X1.PC2)	1.000	1.000	24.68	$< 10^{-3}$
s(X1.PC3)	1.000	1.000	23.93	$< 10^{-3}$
s(X1.PC4)	1.662	2.111	29.80	$< 10^{-3}$
R^2 ajustado = 0.533		*Deviance* explicada = 54.7 %		

Cuadro 3.15: Resultados numéricos del modelo 4 en el análisis del gasto en San Valentín.

En la Figura 3.27 se muestran los residuos generados por este modelo, mediante los que se constrasta la hipótesis de homocedasticidad. Dado que presentan cierta heterocedasticidad, vemos que el diagnóstico en base a este modelo, para este subgrupo concreto, no arroja total fiabilidad.

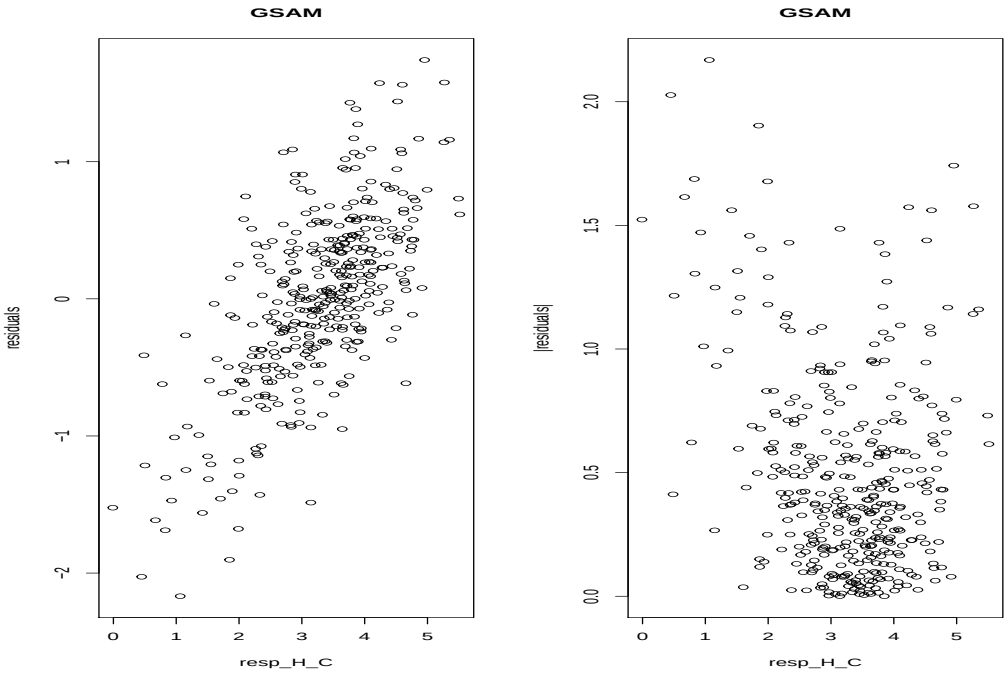

Figura 3.27: Gráfica de los residuos del modelo 4 (izquierda) y valor absoluto de los mismos (derecha) en el análisis de gasto de San Valentín.

Tras el ajuste de los 4 modelos propuestos, emplearemos el 15 % de los clientes de la muestra que no fue utilizado en la regresión para validar cada uno de los modelos mediante predicción.

En la Figura 3.28 representamos la predicciones obtenidas en cada modelo frente a los importes (transformados por logaritmos) de consumo reales durante San Valentín de este conjunto de 66 clientes. A juzgar por los resultados gráficos, el modelo 2 de mínimos cuadrados parciales, que a su vez se corresponde con el de coeficiente R^2 más elevado, es el que realiza una mejor predicción. En el Cuadro 3.16 se muestra el error cuadrático medio cometido en la predicción empleando cada uno de los modelos, mediante el que se verifica el resultado gráfico.

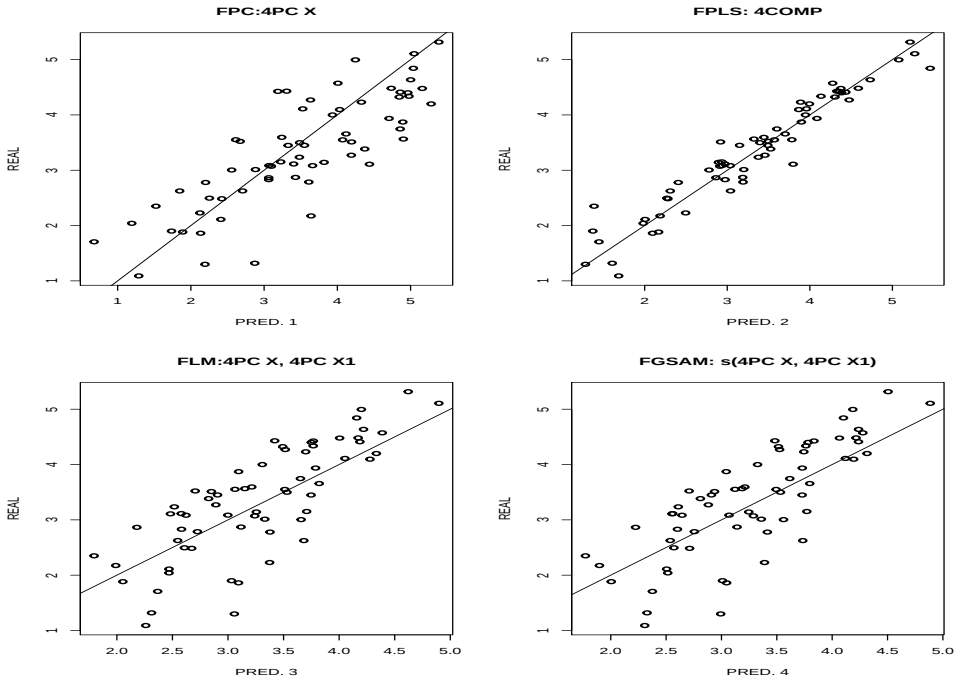

Figura 3.28: Resultados gráficos de predicción frente a consumo real en San Valentín por los modelos funcionales FPC (Modelo 1), FPLS (Modelo 2), FLM con derivada (Modelo 3) y FGSAM (Modelo 4).

MODELO	MSE
1	0.451
2	0.076
3	0.375
4	0.368

Cuadro 3.16: Errores cuadráticos medios cometidos en la predicción mediante los 4 modelos funcionales ajustados: FPC, FPLS, FLM y FGSAM, correspondientes a los datos transformados por logaritmos.

Como en su momento se aplicó la transformación logarítmica a los datos, decidimos deshacerla para conocer la relación entre el valor del consumo real y el obtenido mediante predicción. En el Cuadro 3.17 mostramos la media y mediana de lo que difieren ambas medidas. Así, podemos ver que con el modelo 2 nos desviamos de media 7.88 € en la predicción del gasto de los jóvenes en San Valentín, menos de 4.36 € en la mitad de la muestra.

Llegados a este punto, surge el interés de conocer qué tan necesaria es la segmentación propuesta y hasta que punto es efectiva. Para ello desarrollamos un estudio de predicción cruzada consistente en predecir el consumo durante San Valentín de esta muestra de 66 clientes, pero en lugar de hacerlo con

los modelos correspondientes, la predicción es realizada empleando los modelos de otros subgrupos del sector joven. Los ajustes empleados en este punto serán los de los jóvenes que realizan sus compras en Pontevedra y los de las jóvenes que lo hacen en A Coruña y Pontevedra.

MODELO	Media	Mediana
1	25.71 €	13.56 €
2	7.88 €	4.36 €
3	17.39 €	11.93 €
4	17.31 €	11.34 €

Cuadro 3.17: Media y mediana de las diferencias entre valor real y predicción (en euros) del gasto en San Valentín mediante los 4 modelos ajustados.

En el Cuadro 3.18 se muestran los errores cuadráticos medios (de los datos transformados por logaritmos) cometidos en la predicción cruzada en comparación con los obtenidos en la predicción empleando los modelos correspondientes al subgrupo estudiado. Observamos que dentro de cada modelo, los mejores resultados se obtienen en el caso de emplear los modelos específicos para el subgrupo en cuestión; no obstante, para ambos grupos de mujeres, los resultados obtenidos por el modelo 2 no difieren en demasía. A pesar de que existen otras combinaciones que podríamos validar, en base a estos resultados podemos justificar en este caso la segmentación del subgrupo por sexo y localidad, puesto que los mejores resultados se obtienen con los modelos específicos de este.

MODELO	Hombre Coruña	Hombre Pontevedra	Mujer Coruña	Mujer Pontevedra
1	0.451	0.519	0.509	0.483
2	0.076	0.084	0.078	0.079
3	0.365	0.387	0.471	0.386
4	0.368	0.397	0.482	0.390

Cuadro 3.18: Errores cuadráticos medios cometidos tras el estudio de predicción cruzada en base a modelos de diferentes subgrupos del sector joven.

En los Cuadros 3.19 y 3.20 se muestran los resultados de regresión (R^2) y predicción (MSE) del gasto en San Valentín para los diferentes modelos y la totalidad de subgrupos. Para cada uno de ellos, se resalta en rojo el coeficiente R^2 más elevado y en azul, el modelo para el que se comete menos error en la predicción. A pesar de que no necesariamente tienen porqué coincidir, en este caso sí se asocia mayor R^2 con menor error cuadrático medio, siendo el modelo funcional PLS el que realiza el mejor ajuste de los datos.

	SÉNIOR			
M	***Hombre***		***Mujer***	
	R^2	MSE	R^2	MSE
1	0.82	0.483	0.77	0.458
2	0.95	0.059	0.95	0.056
3	0.58	0.538	0.59	0.428
4	0.61	0.565	0.60	0.418

Cuadro 3.19: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo (transformado por logaritmos) durante la semana de San Valentín de los subgrupos de clientes sénior.

		JÓVENES				ADULTOS 1				ADULTOS 2			
L	M	*Hombre*		*Mujer*		*Hombre*		*Mujer*		*Hombre*		*Mujer*	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
A Coruña	1	0.64	0.451	0.54	0.704	0.79	0.546	0.76	0.518	0.74	0.539	0.81	0.455
	2	0.93	0.076	0.94	0.072	0.95	0.063	0.95	0.066	0.95	0.056	0.95	0.059
	3	0.53	0.375	0.55	0.567	0.59	0.537	0.55	0.576	0.58	0.478	0.56	0.521
	4	0.53	0.368	0.58	0.551	0.60	0.518	0.57	0.552	0.60	0.477	0.58	0.508
Lugo	1	0.83	0.237	0.81	0.354	0.81	0.505	0.70	0.477	0.77	0.431	0.85	0.558
	2	0.96	0.100	0.94	0.060	0.95	0.059	0.95	0.062	0.95	0.079	0.95	0.069
	3	0.79	0.631	0.68	0.522	0.61	0.577	0.44	0.568	0.59	0.652	0.56	0.647
	4	-	-	0.72	0.546	0.62	0.571	0.46	0.551	0.62	0.669	0.59	0.653
Ourense	1	0.91	0.708	0.93	0.518	0.63	0.630	0.79	0.472	0.80	0.444	0.78	0.420
	2	0.96	0.118	0.96	0.096	0.95	0.073	0.95	0.078	0.95	0.062	0.95	0.054
	3	0.73	0.505	0.90	0.203	0.50	0.586	0.50	0.588	0.62	0.489	0.64	0.499
	4	-	-	0.91	0.224	0.52	0.597	0.53	0.560	0.65	0.470	0.66	0.490
Pontevedra	1	0.80	0.291	0.78	0.552	0.74	0.553	0.76	0.500	0.79	0.561	0.78	0.480
	2	0.94	0.092	0.94	0.085	0.95	0.058	0.95	0.063	0.95	0.063	0.95	0.057
	3	0.70	0.414	0.65	0.578	0.54	0.519	0.55	0.530	0.66	0.544	0.57	0.493
	4	0.73	0.467	0.67	0.526	0.56	0.506	0.57	0.515	0.67	0.532	0.58	0.480
Resto	1	0.55	0.840	0.77	0.510	0.76	0.662	0.72	0.699	0.76	0.938	0.74	0.981
	2	0.94	0.057	0.95	0.061	0.95	0.065	0.95	0.085	0.94	0.056	0.96	0.085
	3	0.49	0.624	0.67	0.361	0.61	0.560	0.56	0.752	0.65	0.587	0.73	0.333
	4	0.52	0.592	0.69	0.369	0.62	0.552	0.57	0.699	0.71	0.542	0.78	0.353

Cuadro 3.20: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo (transformado por logaritmos) durante la semana de San Valentín de los subgrupos de clientes jóvenes, adultos 1 y adultos 2.

3.4.2. Semana Santa

El segundo acontecimiento que nos interesa analizar es el consumo durante Semana Santa, que en el 2015 tuvo lugar del 30 de marzo al 5 de abril, semana correspondiente a la décimo cuarta del año. En este caso, para ejemplificar este análisis, se empleará la muestra de clientas de entre 26 y 45 años que realizan sus compras en la provincia de Pontevedra. Realizaremos el ajuste de regresión con 10 413 curvas, correspondientes al 85 % de la muestra, y se emplearán las 1 838 restantes para la validación de la predicción.

Los primeros resultados del análisis de regresión se muestran en el Cuadro 3.21, correspondientes al modelo 1 de componentes principales funcional. En este ajuste se emplean 9 componentes principales que explican el 76.55 % de la variabilidad, obteniendo un coeficiente R^2=0.71, relativamente alto. En la Figura 3.29 se observan una gran cantidad de curvas atípicas e influyentes, en cuyo análisis no nos vamos a demorar porque no describen ningún patrón reseñable. Por lo demás, este modelo cumple las hipótesis de normalidad, media nula y homocedasticidad de los residuos.

MODELO 1	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.901	0.005	711.702	$< 10^{-3}$
PC1	0.215	0.002	101.619	$< 10^{-3}$
PC9	-0.473	0.004	-98.655	$< 10^{-3}$
PC8	0.241	0.004	51.601	$< 10^{-3}$
PC10	0.221	0.005	42.741	$< 10^{-3}$
PC5	0.091	0.004	20.517	$< 10^{-3}$
PC3	-0.086	0.004	-19.588	$< 10^{-3}$
PC6	0.085	0.004	19.227	$< 10^{-3}$
PC2	-0.062	0.004	-15.706	$< 10^{-3}$
PC4	0.033	0.004	8.063	$< 10^{-3}$
$R^2 = 0.712$	76.55 % variabilidad explicada con 9 PC's			$\lambda = 0.5$
Varianza residual: 0.312 en 10403 grados de libertad				

Cuadro 3.21: Resultados numéricos del modelo 1 en el análisis del gasto en Semana Santa.

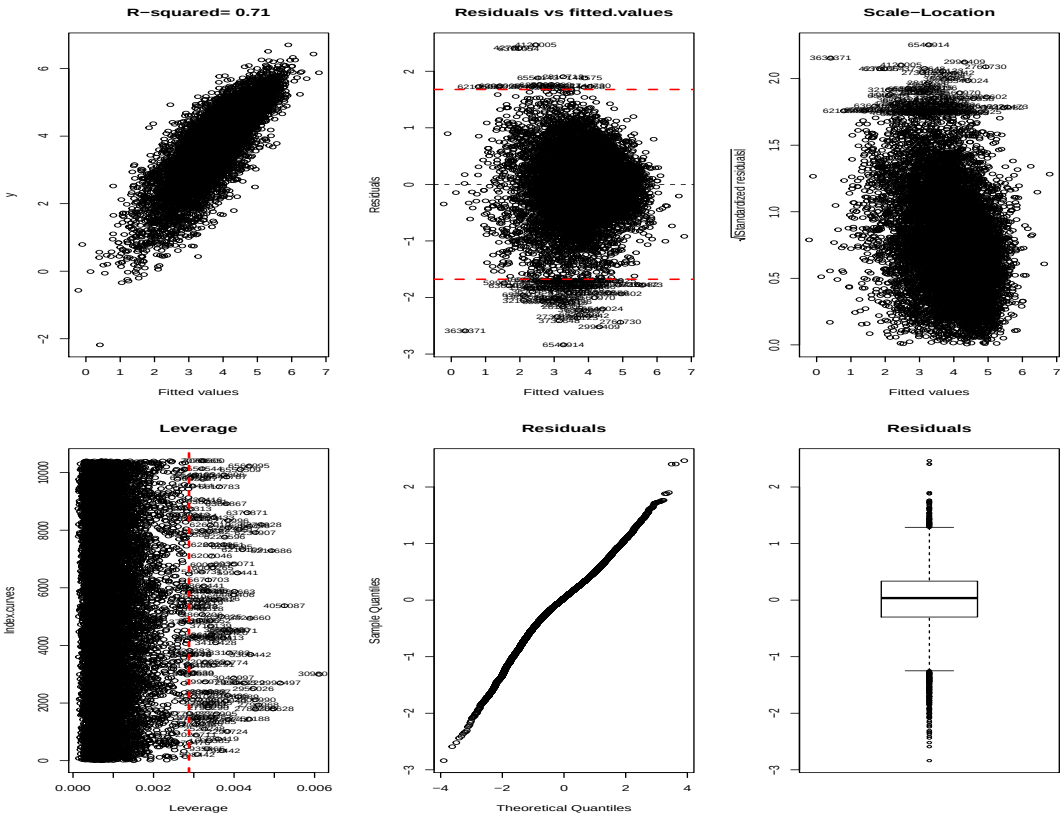

Figura 3.29: Gráficas del modelo 1 para el análisis de gasto durante Semana Santa. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

En la Figura 3.30 se representa la estimación del parámetro β. Se observa que el consumo de este subgrupo de clientes durante Semana Santa está determinado por el gasto en las semanas más cercanas y asociado al bajo consumo en las semanas 6 y 22.

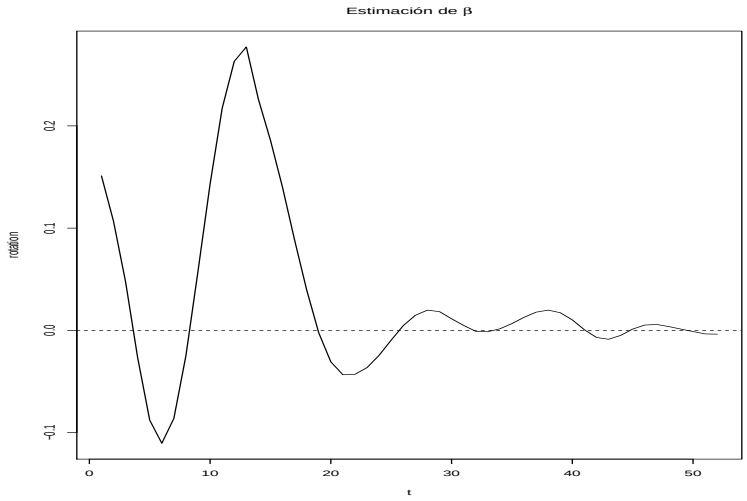

Figura 3.30: Estimación del parámetro de regresión β por el modelo 1 (Semana Santa).

En el modelo FPLS las 4 componentes empleadas resultan todas ellas significativas, como se puede observar en el Cuadro 3.22, obteniendo un coeficiente R^2=0.95, mayor que en el caso anterior.

MODELO 2	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.901	0.002	1707.617	$< 10^{-3}$
PLS1	0.221	0.002	319.283	$< 10^{-3}$
PLS2	0.391	0.001	281.068	$< 10^{-3}$
PLS3	0.184	0.001	102.528	$< 10^{-3}$
PLS4	0.135	0.002	66.621	$< 10^{-3}$
$R^2 = 0.949$	Varianza residual: 0.054 en 10394.57 grados de libertad			

Cuadro 3.22: Resultados numéricos del modelo 2 en el análisis del gasto en Semana Santa.

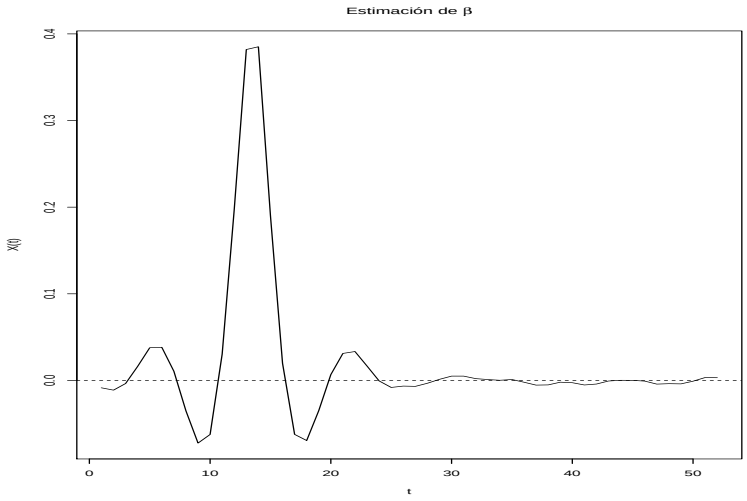

Figura 3.31: Estimación del parámetro de regresión β por el modelo 2 (Semana Santa).

En esta ocasión, la Figura 3.31 representa una estimación del parámetro β bastante plana, salvo por el elevado gasto ocasionado en las semanas más próximas a Semana Santa. En menor medida, cabe mencionar la influencia del gasto realizado en las semanas 5 y 12 y del bajo consumo de las semanas 8 y 19 en la explicación del gasto durante este acontecimiento para este subgrupo de clientes.

Cabe destacar, de nuevo, la presencia de un gran número de curvas atípicas e influyentes, que quedan reflejadas en la Figura 3.32, en la que salvo este, el modelo parece no presentar otros inconvenientes, aunque es difícil realizar esta afirmación dada la cantidad de curvas analizadas.

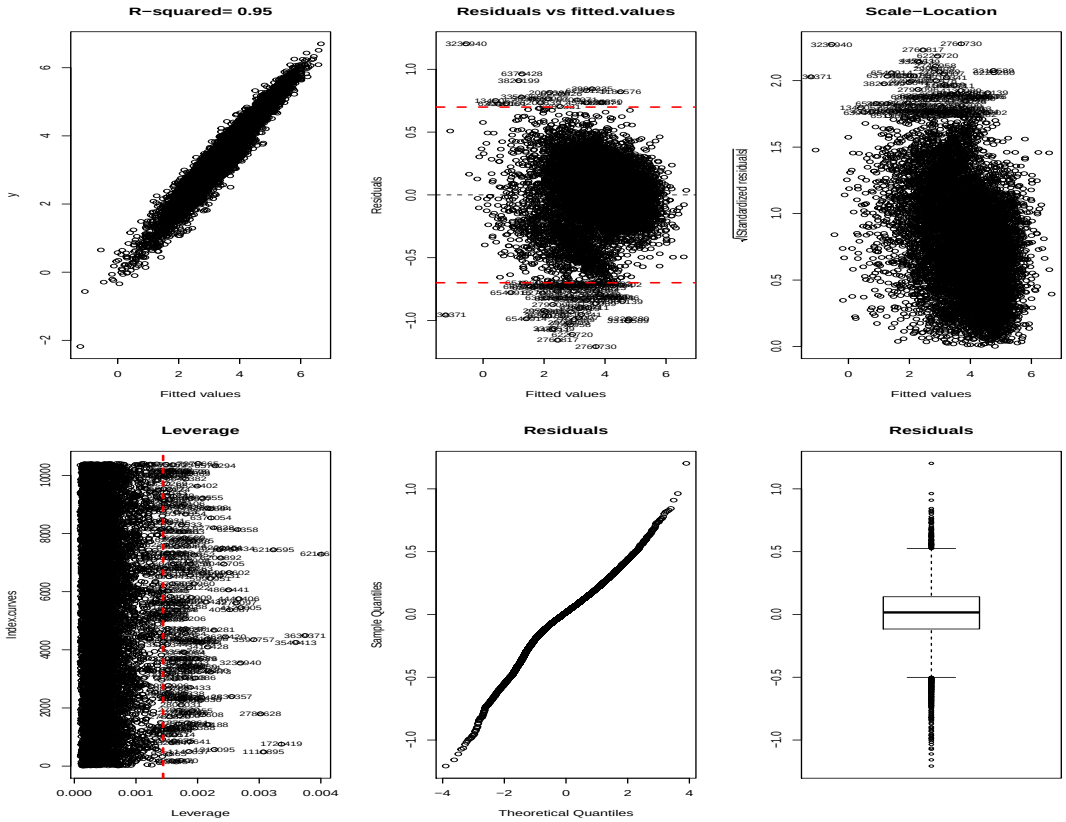

Figura 3.32: Gráficas del modelo 2 para el análisis de gasto durante Semana Santa. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

Mediante el modelo 3, el modelo lineal funcional que incluye la derivada del dato funcional como covariable, se obtienen los resultados del Cuadro 3.23. Se emplean 4 componentes principales en la representación del dato funcional y la derivada de éste, siendo todas ellas significativas, y se obtiene un coeficiente R^2 ajustado de 0.30, inferior al obtenido en los ajustes anteriores.

Mediante las gráficas inferiores de la Figura 3.33 se detectan algunas curvas influyentes, como ya veníamos observando en otros modelos y la segunda y tercera gráfica de la fila superior ponen en entredicho la normalidad y homocedasticidad de los residuos.

En la Figura 3.34 se representan las estimaciones del parámetro β asociadas al modelo 3, correspondiéndose la curva roja con los parámetros $\hat{\beta}_j$ (estimación en los datos originales) y la verde a los $\hat{\beta}_k$ (derivada) de la ecuación (3.1).

MODELO 3	Estimación	Error estándar	t *valor*	$Pr(>\|t\|)$
Intercepto	3.902	0.008	456.207	$< 10^{-3}$
X.PC1	0.135	0.002	61.268	$< 10^{-3}$
X.PC2	-0.021	0.004	-4.982	$< 10^{-3}$
X.PC3	-0.082	0.004	-16.923	$< 10^{-3}$
X.PC4	0.023	0.004	5.438	$< 10^{-3}$
X1.PC1	-0.032	0.013	-2.299	$< 10^{-3}$
X1.PC2	-0.033	0.012	-2.633	$< 10^{-3}$
X1.PC3	0.029	0.009	3.101	$< 10^{-3}$
X1.PC4	0.028	0.013	2.659	$< 10^{-3}$
$R^2 = 0.298$	R^2 ajustado $= 0.298$		p-*valor*: $< 10^{-3}$	
Error estándar residual: 0.873 en 10404 DF			Estad. F: 553.9 en 8 y 10404 DF	

Cuadro 3.23: Resultados numéricos del modelo 3 en el análisis del gasto en Semana Santa.

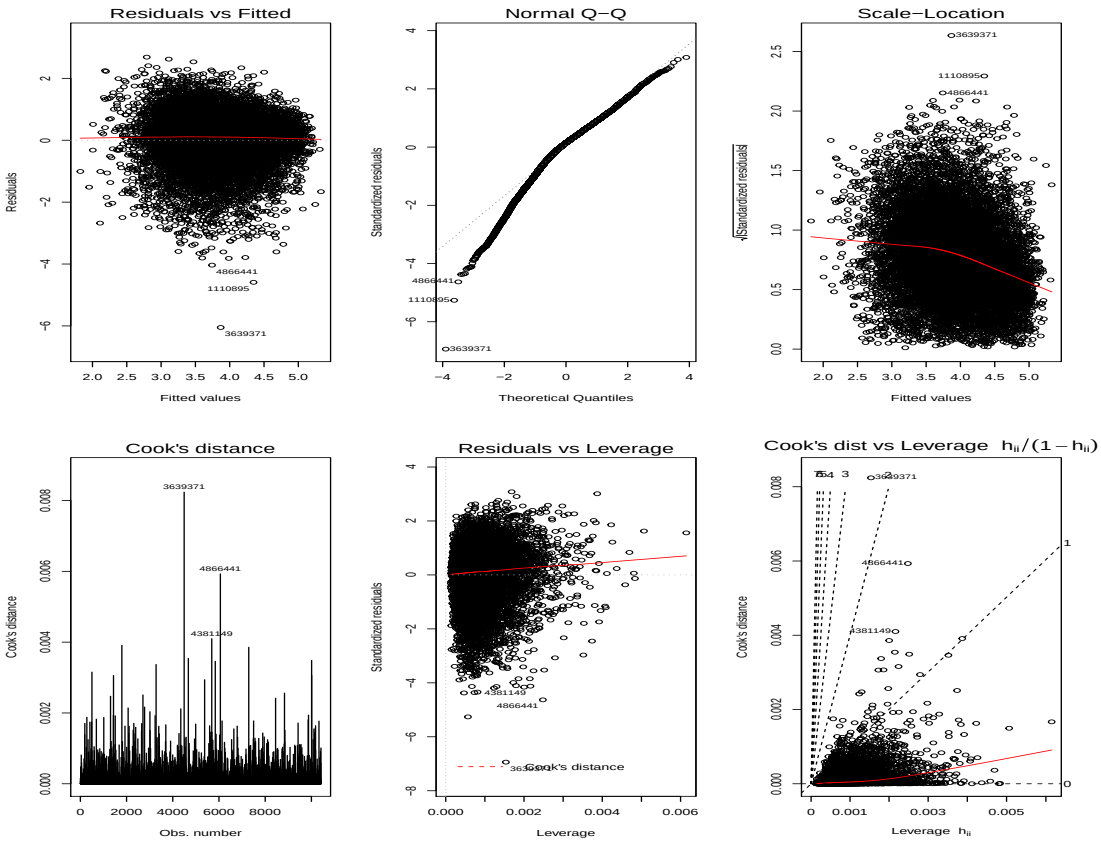

Figura 3.33: Gráficas del modelo 3 para el análisis de gasto durante Semana Santa. De izquierda a derecha: media cero, gráfico de normalidad Q-Q, homocedasticidad, distancia de Cook y residuos y distancia de Cook frente a apalancamientos.

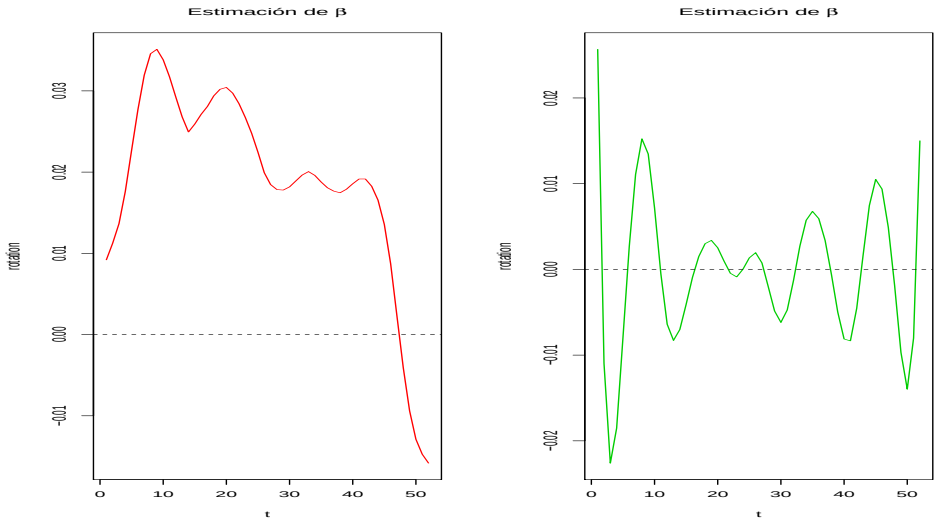

Figura 3.34: Estimación del parámetro de regresión β por el modelo 3 (izquierda) y estimación correspondiente a la derivada (derecha) en el análisis de gasto de Semana Santa.

A continuación, analizaremos la influencia que tiene la variable *actividad* dentro del subgrupo de mujeres de entre 26 y 45 años que realizan sus compras mayoritariamente en Pontevedra. La muestra de entrenamiento está compuesta de 4 702 clientas activas ocupadas, 1 032 activas paradas, 1 954 inactivas y de 2 725 de ellas se desconoce la actividad que desempeñan, por lo que decidimos crear un modelo para cada segmentación de esta variable.

MODELO 3	$\mathcal{X}$	$\mathcal{X}^{(1)}$	R^2	R^2 ajust.	*p-valor*
OCUPADAS	1:4	1-3-4	0.378	0.372	$< 10^{-3}$
PARADAS	1:4	2-3	0.313	0.312	$< 10^{-3}$
INACTIVAS	1:4	1:4	0.397	0.391	$< 10^{-3}$
SE DESCONOCE	1:4	2:4	0.337	0.332	$< 10^{-3}$

Cuadro 3.24: Resumen del modelo 3 para el análisis de gasto durante Semana Santa de las clientas adultas 1 clasificadas según su actividad contable.

Analizando los resultados del Cuadro 3.24, comprobamos que en el caso de las clientas de entre 26 y 45 años que realizan sus compras en la provincia de Pontevedra, podría ser interesante considerar la posibilidad de clasificarlas por la variable actividad en el análisis del consumo que realizan durante Semana Santa, puesto que el ajuste mejora ligeramente en comparación con el que no tenía en cuenta esta segmentación (R^2=0.30). La mejora más significativa se produce en el grupo de las clientas inactivas, lo cuál tiene sentido puesto que al no tener ingresos fijos mensuales, se espera que tengan un patrón de gasto distinto al de las clientas activas.

En el Cuadro 3.25 se muestran los resultados numéricos tras el ajuste del modelo FGSAM para la explicación del consumo durante Semana Santa de las clientas pontevedresas pertenecientes al subgrupo de adultos 1. Todas las componentes resultan significativas teniendo en cuenta un nivel de significación del 10 %. Se obtiene un coeficiente R^2=0.30, siendo la *deviance* explicada del 30.4 %.

En la Figura 3.35 se muestran los residuos generados por este modelo, mediante la que se observa que presentan una clara heterocedasticidad.

MODELO 4	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	3.901	0.008	457.6	$< 10^{-3}$
	EDF	Ref. DF	F	p-value
s(X.PC1)	1.000	1.000	3462.872	$< 10^{-3}$
s(X.PC2)	4.269	5.378	10.032	$< 10^{-3}$
s(X.PC3)	4.061	5.125	46.092	$< 10^{-3}$
s(X.PC4)	2.559	3.315	10.125	$< 10^{-3}$
s(X1.PC1)	3.276	4.215	6.839	$< 10^{-3}$
s(X1.PC2)	1.000	1.000	8.699	$< 10^{-3}$
s(X1.PC3)	1.000	1.000	9.389	$< 10^{-3}$
s(X1.PC4)	1.992	2.585	2.265	0.0864
R^2 ajustado = 0.303		*Deviance* explicada =30.4 %		

Cuadro 3.25: Resultados numéricos del modelo 4 en el análisis del gasto en Semana Santa.

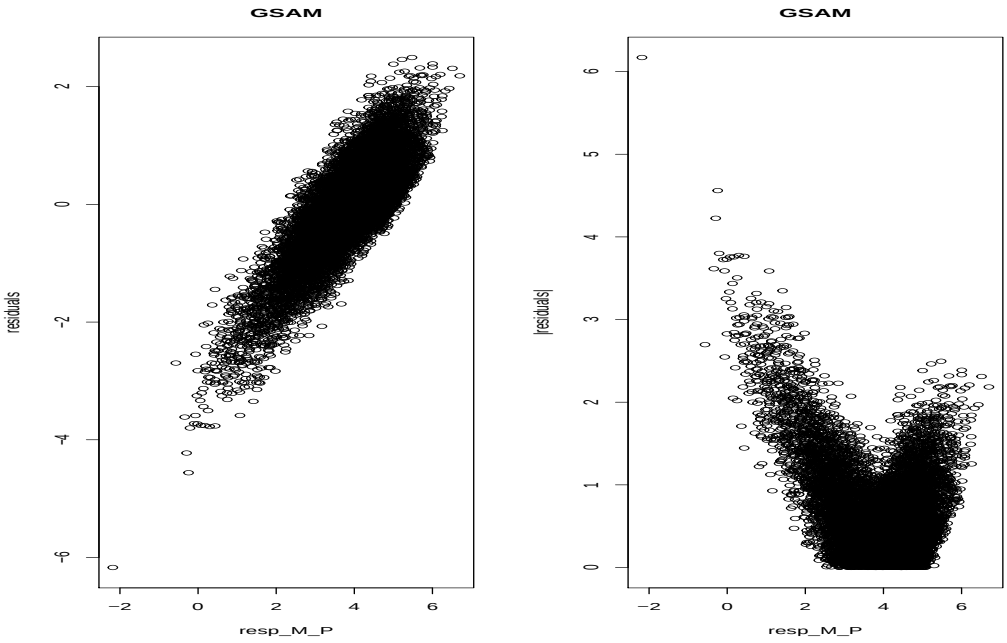

Figura 3.35: Gráfica de los residuos del modelo 4 (izquierda) y valor absoluto de los mismos (derecha) en el análisis de gasto de Semana Santa.

Ajustados los modelos de regresión, procedemos a presentar los resultados de predicción en base a los 4 ajustes. Para ello se emplea el 15 % de la muestra que no fue utilizado en la regresión, correspondiente a 1 838 curvas de consumo de las clientas del subgrupo analizado, con el objetivo de validar cada uno de los modelos mediante predicción.

Se muestran la predicciones obtenidas mediante cada uno de los modelos frente a los importes (transformados por logaritmos) de consumo reales durante Semana Santa en la Figura 3.36. El modelo FPLS, correspondiente al mayor coeficiente R^2, es el que realiza una predicción más próxima al consumo real en comparación al resto de modelos. En el Cuadro 3.26 puede consultarse el error cuadrático medio cometido en la predicción para cada uno de los modelos empleados. Los resultados numéricos verifican que, efectivamente, es el modelo 2 el asociado a un menor error cuadrático medio de predicción, significativamente inferior al de los otros modelos.

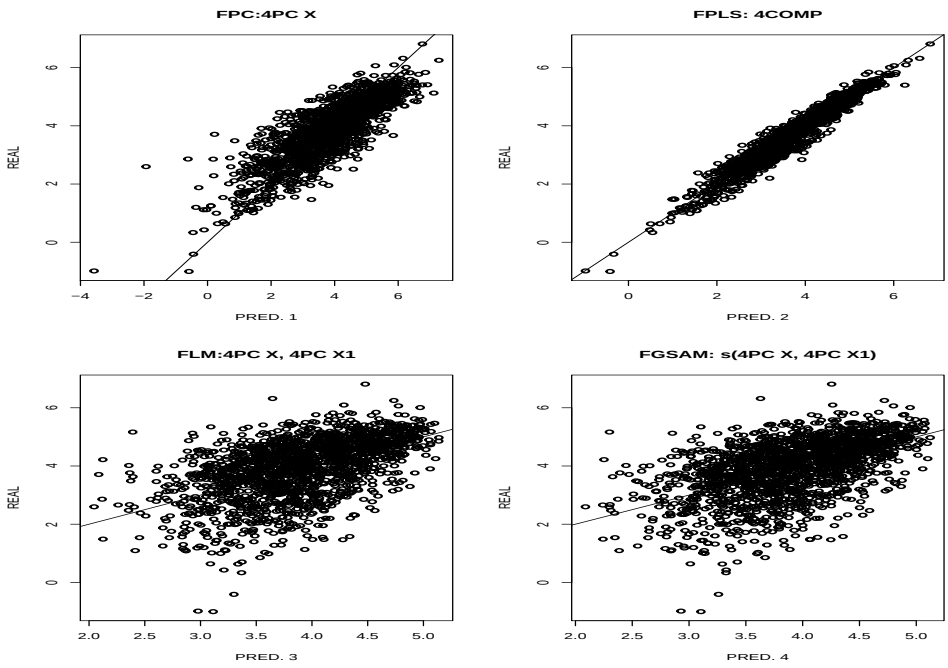

Figura 3.36: Resultados gráficos de predicción frente a consumo real en Semana Santa por los modelos funcionales FPC (Modelo 1), FPLS (Modelo 2), FLM con derivada (Modelo 3) y FGSAM (Modelo 4).

MODELO	MSE
1	0.494
2	0.053
3	0.717
4	0.713

Cuadro 3.26: Errores cuadráticos medios cometidos en la predicción mediante los 4 modelos funcionales ajustados: FPC, FPLS, FLM y FGSAM, correspondientes a los datos transformados por logaritmos.

En el Cuadro 3.27 mostramos los resultados acerca de la media y mediana de la diferencia entre el consumo real y el obtenido mediante predicción al deshacer la transformación logarítmica. Así, podemos ver que con el modelo 2 nos desviamos de media 10.30 € en la predicción del gasto en Semana Santa y menos de 6.36 € en la mitad de la muestra analizada.

MODELO	Media	Mediana
1	45.92 €	21.30 €
2	10.30 €	6.36 €
3	38.09 €	25.86 €
4	37.91 €	25.59 €

Cuadro 3.27: Media y mediana de las diferencias entre valor real y predicción (en euros) del gasto en Semana Santa mediante los 4 modelos ajustados.

Desarrollaremos a continuación un estudio de predicción cruzada, con el objetivo de analizar qué tan efectiva y necesaria es la segmentación propuesta para este subgrupo. Para ello, realizaremos la predicción del consumo de las clientas que compran en Pontevedra por medio de los modelos de otros subgrupos de clientes de entre 26 y 45 años. Los subgrupos empleados en el cruce de predicción serán el de las clientas que realizan sus compras en Ourense y el de lo clientes que compran en Pontevedra y Lugo.

Los errores cuadráticos medios cometidos en la predicción cruzada en comparación con los obtenidos en la predicción empleando los modelos correspondientes al subgrupo de mujeres adultas 1 de Pontevedra se muestran en el Cuadro 3.28. Para cada uno de los modelos, observamos que los mejores resultados se obtienen en el caso de emplear el modelo específico para el subgrupo en cuestión. Además, independientemente del subgrupo empleado, observamos que con el modelo FPLS se obtienen mejores resultados que con los restantes modelos.

MODELO	Mujer Pontevedra	Mujer Ourense	Hombre Pontevedra	Hombre Lugo
1	0.494	0.519	0.512	0.553
2	0.053	0.064	0.062	0.063
3	0.717	0.723	0.723	0.729
4	0.713	0.733	0.733	0.731

Cuadro 3.28: Errores cuadráticos medios cometidos tras el estudio de predicción cruzada en base a modelos de diferentes subgrupos del sector adultos 1.

		JÓVENES				**ADULTOS 1**				**ADULTOS 2**			
L	M	***Hombre***		***Mujer***		***Hombre***		***Mujer***		***Hombre***		***Mujer***	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
A Coruña	1	0.67	0.462	0.70	0.626	0.68	0.540	0.68	0.541	0.70	0.526	0.69	0.458
	2	0.95	0.054	0.94	0.076	0.95	0.052	0.95	0.052	0.96	0.051	0.95	0.050
	3	0.32	0.935	0.30	0.886	0.31	0.798	0.35	0.763	0.40	0.690	0.35	0.660
	4	0.34	0.947	0.32	0.918	0.31	0.801	0.35	0.759	0.41	0.694	0.36	0.656
Lugo	1	0.90	0.300	0.82	0.253	0.64	0.450	0.62	0.554	0.74	0.462	0.61	0.627
	2	0.95	0.093	0.94	0.093	0.95	0.059	0.95	0.053	0.95	0.051	0.95	0.053
	3	0.65	0.348	0.46	0.878	0.37	0.612	0.34	0.786	0.45	0.644	0.37	0.864
	4	-	-	0.55	0.900	0.38	0.629	0.36	0.785	0.46	0.667	0.39	0.880
Ourense	1	0.85	0.246	0.58	0.626	0.69	0.457	0.71	0.466	0.73	0.416	0.72	0.471
	2	0.96	0.059	0.94	0.049	0.95	0.061	0.95	0.066	0.96	0.054	0.95	0.050
	3	0.68	0.411	0.54	0.432	0.39	0.620	0.30	0.748	0.45	0.663	0.50	0.447
	4	-	-	0.58	0.423	0.41	0.656	0.32	0.754	0.49	0.625	0.52	0.452
Pontevedra	1	0.77	0.613	0.75	0.452	0.68	0.537	0.71	0.494	0.72	0.486	0.70	0.463
	2	0.94	0.067	0.94	0.068	0.95	0.054	0.95	0.053	0.95	0.059	0.95	0.053
	3	0.58	0.500	0.44	0.602	0.33	0.668	0.30	0.717	0.43	0.637	0.36	0.622
	4	0.58	0.495	0.47	0.624	0.33	0.671	0.30	0.713	0.43	0.639	0.37	0.616
Resto	1	0.60	0.770	0.73	0.490	0.68	0.618	0.65	0.675	0.76	0.892	0.62	0.700
	2	0.94	0.095	0.95	0.066	0.95	0.065	0.94	0.086	0.95	0.086	0.95	0.091
	3	0.51	0.688	0.39	0.728	0.34	0.905	0.36	0.742	0.54	0.741	0.57	0.816
	4	0.52	0.668	0.39	0.718	0.36	0.896	0.36	0.734	0.60	0.766	0.62	0.941

Cuadro 3.29: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo durante Semana Santa de los subgrupos de clientes jóvenes, adultos 1 y adultos 2.

	SÉNIOR			
M	***Hombre***		***Mujer***	
	R^2	MSE	R^2	MSE
1	0.73	0.537	0.68	0.475
2	0.96	0.052	0.95	0.043
3	0.38	0.677	0.39	0.524
4	0.39	0.698	0.39	0.522

Cuadro 3.30: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo durante Semana Santa de los subgrupos de clientes sénior.

En los Cuadros 3.29 y 3.30 se muestran los resultados de regresión (R^2) y predicción (MSE) del consumo durante Semana Santa para los diferentes modelos y la totalidad de subgrupos. Para cada uno de ellos, se resalta en rojo el coeficiente R^2 más elevado y en azul, el modelo para el que se comete menos error en la predicción. Vemos que en todos los casos, los mejores resultados se obtienen con el ajuste PLS.

3.4.3. Black Friday

El tercer evento a analizar es el consumo durante el Black Friday, que tuvo lugar el 27 de noviembre de 2015. Para la entidad es importante analizar el gasto producido en esta fecha, no solo por el interés en sí mismo si no también, porque de una forma u otra, para determinados establecimientos marca el inicio de la campaña navideña. Por lo tanto, en esta ocasión se analizarán las transacciones acaecidas en la semana 48 del año 2015. El subgrupo empleado para visualizar el análisis es el de los varones entre 46 y 59 años de edad que realizan sus compras de forma mayoritaria en la provincia de Lugo. De nuevo, de los 609 clientes que conforman la muestra, el 85 % será empleado para realizar el ajuste de regresión y el 15 % restante para validar el modelo mediante predicción.

Comenzamos el análisis de regresión ajustando el modelo 1, correspondiente al modelo funcional de componentes principales. En el Cuadro 3.31 mostramos los resultados obtenidos tras realizar dicho ajuste. En este modelo se emplean 7 componentes principales que explican un 64.32 % de la variabilidad, siendo el coeficiente R^2=0.60.

MODELO 1	Estimación	Error estándar	$t\ valor$	$Pr(>\|t\|)$
Intercepto	4.088	0.028	144.154	$< 10^{-3}$
PC1	-0.187	0.009	-18.970	$< 10^{-3}$
PC8	0.314	0.024	13.049	$< 10^{-3}$
PC10	0.292	0.027	10.659	$< 10^{-3}$
PC5	-0.207	0.022	-9.256	$< 10^{-3}$
PC7	0.095	0.025	3.806	$< 10^{-3}$
PC9	0.077	0.024	3.143	$< 10^{-3}$
PC4	-0.064	0.021	-3.044	$< 10^{-3}$
$R^2 = 0.599$	64.32 % variabilidad explicada con 7 PC's			$\lambda = 0.5$
Varianza residual: 0.417 en 5107 grados de libertad				

Cuadro 3.31: Resultados numéricos del modelo 1 en el análisis del gasto durante el Black Friday.

Este ajuste detecta ciertas curvas atípicas e influyentes, como se puede observar en la Figura 3.37. No obstante, salvo la presencia de curvas que difieren del comportamiento general del grupo, no presenta otros inconvenientes y satisface las hipótesis de media nula, homocedasticidad y normalidad de los residuos.

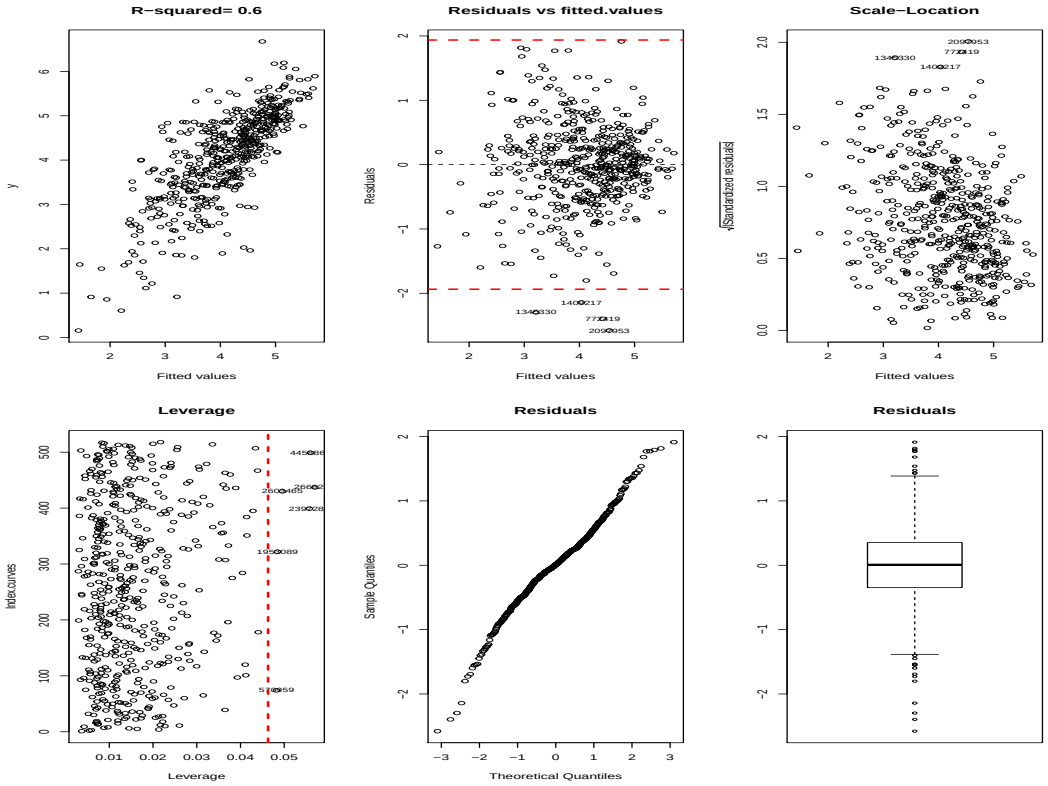

Figura 3.37: Gráficas del modelo 1 para el análisis de gasto en Black Friday. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

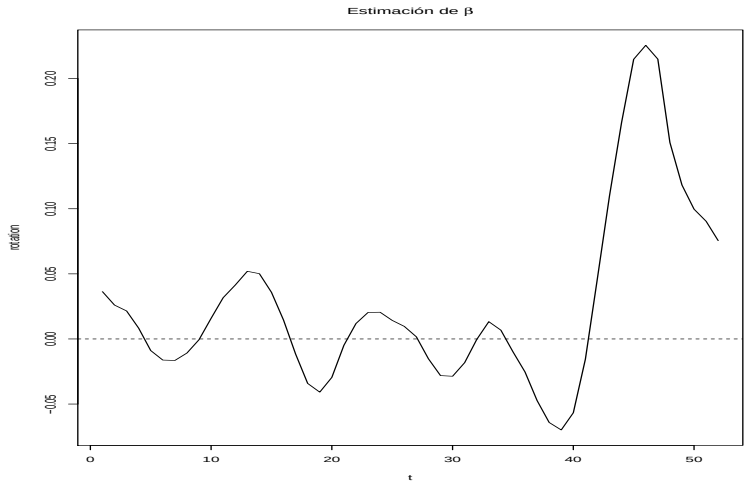

Figura 3.38: Estimación del parámetro de regresión β por el modelo 1 (Black Friday).

En la Figura 3.38 se representa la estimación del parámetro β, en la que se observa una clara influencia del consumo de las semanas en torno al Black Friday en la explicación del gasto producido durante el evento, que a su vez está asociado con el gasto en las semanas 11, 23 y 33 y el bajo consumo durante las semanas 6, 18, 30 y 39.

MODELO 2	Estimación	Error estándar	t *valor*	$Pr(>\|t\|)$
Intercepto	4.088	0.009	409.349	$< 10^{-3}$
PLS1	0.187	0.003	69.164	$< 10^{-3}$
PLS2	0.417	0.006	66.034	$< 10^{-3}$
PLS3	0.255	0.008	29.991	$< 10^{-3}$
PLS4	0.099	0.009	12.312	$< 10^{-3}$
$R^2 = 0.951$	Varianza residual: 0.052 en 499.9258 grados de libertad			

Cuadro 3.32: Resultados numéricos del modelo 2 en el análisis del gasto durante el Black Friday.

En el modelo 2 de mínimos cuadrados parciales las 4 componentes empleadas resultan significativas, tal y como podemos observar en el Cuadro 3.32. Se obtiene un coeficiente R^2=0.95, mayor que en el caso anterior. Cabe destacar la presencia de ciertas curvas atípicas e influyentes, tal como se observa en la Figura 3.39, en la que salvo este, el modelo parece no presentar inconvenientes en el cumplimiento de hipótesis.

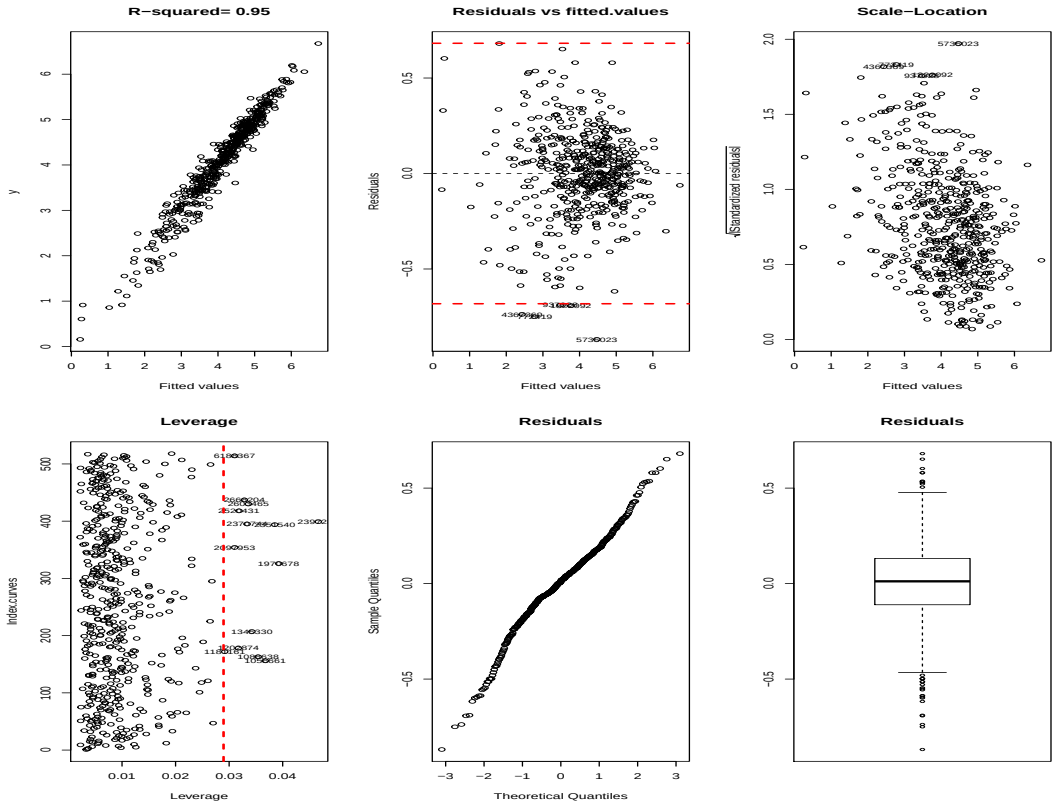

Figura 3.39: Gráficas del modelo 2 para el análisis de gasto en Black Friday. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

A la vista de la Figura 3.40, la estimación del parámetro β es bastante plana en esta ocasión, con excepción de la semana 43, cuyo consumo se relaciona inversamente al ocasionado en Black Friday. El gasto durante este evento viene asociado al consumo en las semanas más próximas a la fecha en cuestión.

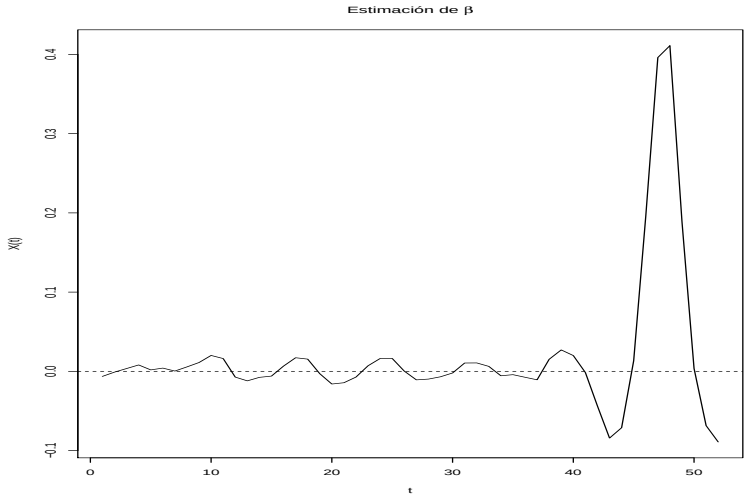

Figura 3.40: Estimación del parámetro de regresión β por el modelo 2 (Black Friday).

En el modelo lineal funcional incorporando la derivada del dato funcional como covariable, todas las componentes resultan significativas considerando un nivel de significación del 10 %, tal y como se deriva del Cuadro 3.33. El coeficiente R^2 ajustado obtenido en esta ocasión es de 0.48, inferior al obtenido tras los ajustes anteriores.

MODELO 3	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	4.088	0.032	127.503	$< 10^{-3}$
X.PC1	-0.123	0.007	-16.662	$< 10^{-3}$
X.PC2	-0.027	0.016	-1.724	0.0853
X.PC3	0.066	0.016	3.991	$< 10^{-3}$
X.PC4	-0.119	0.018	-6.796	$< 10^{-3}$
X1.PC1	-0.361	0.049	-7.256	$< 10^{-3}$
X1.PC2	-0.536	0.052	-10.233	$< 10^{-3}$
X1.PC3	-0.207	0.035	-5.745	$< 10^{-3}$
X1.PC4	-0.168	0.044	-3.819	$< 10^{-3}$
$R^2 = 0.489$	R^2 ajustado = 0.481		*p-valor*: $< 10^{-3}$	
Error estándar residual: 0.729 en 509 DF			Estad. F: 60.82 en 8 y 509 DF	

Cuadro 3.33: Resultados numéricos del modelo 3 en el análisis del gasto durante el Black Friday.

Mediante las gráficas de la fila superior de la Figura 3.41 se puede presenciar cierto incumplimiento de las hipótesis del modelo, mientras que en la fila inferior se observa la existencia de alguna curva influyente.

En la Figura 3.42 se representan las estimaciones del parámetro β asociadas a este ajuste, correspondiéndose la curva roja con los parámetros $\hat{\beta}_j$ (estimación en los datos originales) y la verde a los $\hat{\beta}_k$ (derivada) de la ecuación (3.1).

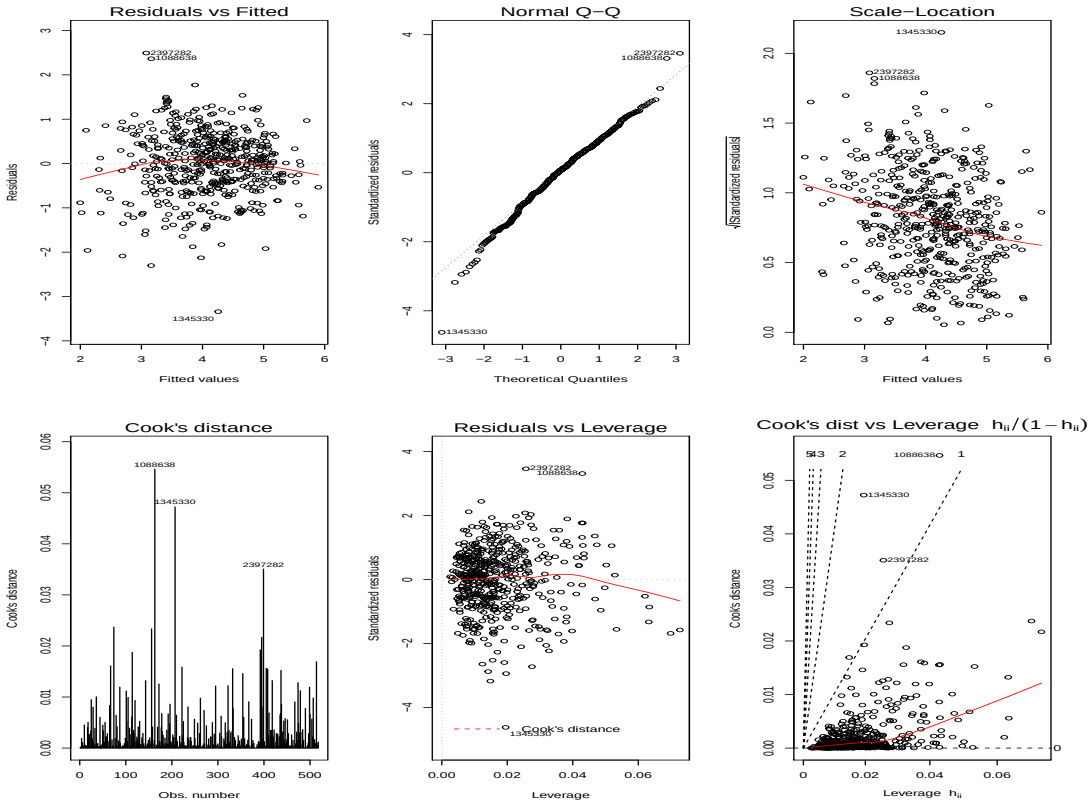

Figura 3.41: Gráficas del modelo 3 para el análisis de gasto en Black Friday. De izquierda a derecha: media cero, gráfico de normalidad Q-Q, homocedasticidad, distancia de Cook y residuos y distancia de Cook frente a apalancamientos.

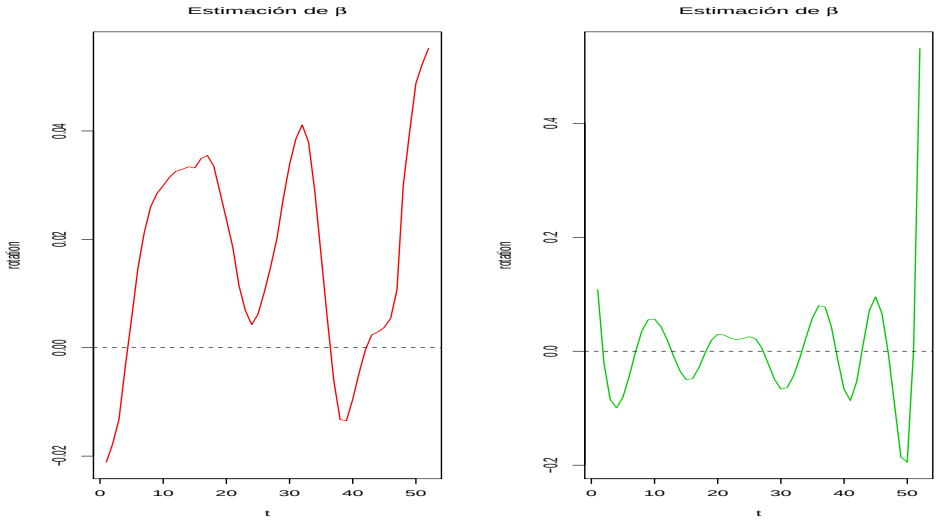

Figura 3.42: Estimación del parámetro de regresión β por el modelo 3 (izquierda) y estimación correspondiente a la derivada (derecha) en el análisis de gasto del Black Friday.

Analizaremos ahora el efecto que tiene la variable *actividad* dentro del subgrupo de varones de entre 46 y 59 años que realizan sus compras en Lugo. La muestra de entrenamiento está formada por 518 clientes, de los cuáles 436 son activos ocupados, 31 activos parados, 10 inactivos y de 41 clientes se desconoce la actividad que desempeñan. El 84 % de la muestra de entrenamiento corresponde a clientes activos ocupados, un porcentaje superior al 67 % de la muestra total de clientes en este rango de edad, cuyo comportamiento podría estar enmascarando el del 16 % restante de clientes al tener unos ingresos fijos mensuales, lo que se puede traducir en un patrón de gasto bastante diferente. Realizaremos un ajuste diferenciado para explicar el consumo de este subgrupo de actividad y de los restantes subgrupos contables.

MODELO 3	$\mathcal{X}$	$\mathcal{X}^{(1)}$	R^2	R^2 ajust.	*p-valor*
OCUPADOS	1:4	1:4	0.582	0.578	$< 10^{-3}$
- OCUPADOS	1:2	1:2	0.513	0.506	$< 10^{-3}$

Cuadro 3.34: Resumen del modelo 3 para el análisis de gasto durante el Black Friday de los clientes adultos 2 activos ocupados (de forma separada) frente al resto.

Analizando los resultados del Cuadro 3.34, comprobamos que en el caso de los clientes pertenecientes al subgrupo de edad de los adultos 2 que realizan sus compras en la provincia de Lugo, podría ser interesante tener en cuenta la posibilidad de clasificar a los clientes por la variable actividad en el análisis del consumo que realizan durante el Black Friday; ya que, pese que el ajuste no mejora en demasía de forma general, si lo hace, ligeramente, para el subgrupo de activos ocupados y, sobre todo, para el subgrupo restante.

En el Cuadro 3.35 se muestran los resultados numéricos tras el ajuste del modelo 4, el FGSAM, para la explicación del consumo de los clientes adultos 2 que compran en Lugo durante el Black Friday. Todas las componentes resultan significativas, obteniendo un coeficiente R^2=0.53 y siendo la *deviance* explicada del 56.1 %.

MODELO 4	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	4.088	0.034	134.5	$< 10^{-3}$
	EDF	Ref. DF	F	p-value
s(X.PC1)	7.166	8.205	31.956	$< 10^{-3}$
s(X.PC2)	7.064	8.096	2.201	$< 10^{-3}$
s(X.PC3)	2.978	3.775	7.700	$< 10^{-3}$
s(X.PC4)	2.117	2.720	17.692	$< 10^{-3}$
s(X1.PC1)	2.360	3.030	21.022	$< 10^{-3}$
s(X1.PC2)	7.362	8.349	15.113	$< 10^{-3}$
s(X1.PC3)	1.000	1.000	31.772	$< 10^{-3}$
s(X1.PC4)	1.029	1.058	11.913	$< 10^{-3}$
R^2 ajustado = 0.533		*Deviance* explicada = 56.1 %		

Cuadro 3.35: Resultados numéricos del modelo 4 en el análisis del gasto durante el Black Friday.

En la Figura 3.43 se muestran los residuos generados por este modelo, mediante los que se rechaza la hipótesis de homocedasticidad, ya que se evidencia la presencia de heterocedasticidad en los residuos y su valor absoluto.

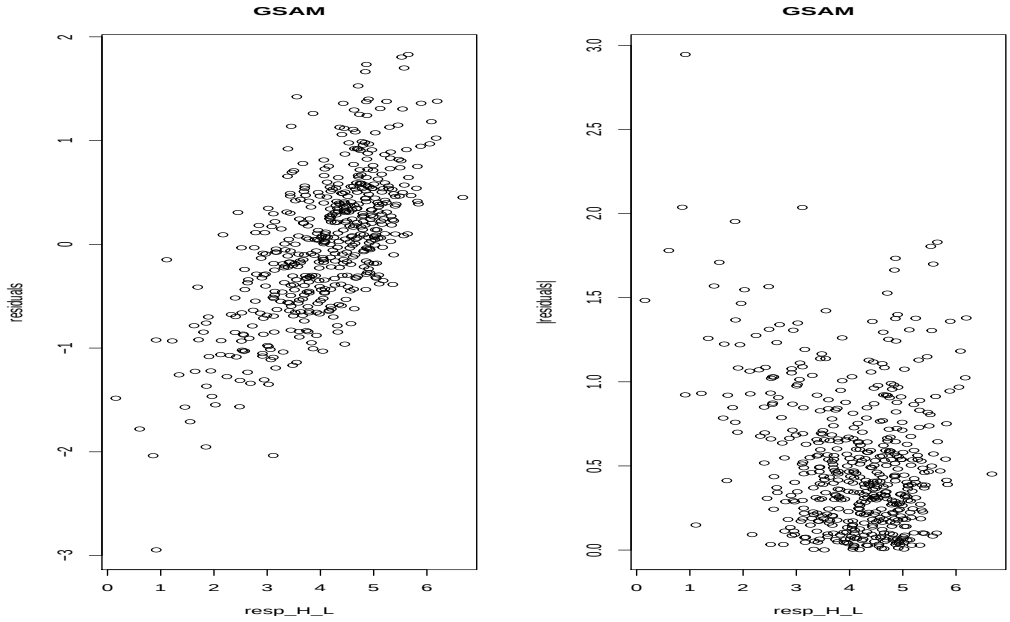

Figura 3.43: Gráfica de los residuos del modelo 4 (izquierda) y valor absoluto de los mismos (derecha) en el análisis de gasto del Black Friday.

Una vez ajustados los modelos de regresión, se desarrolla la predicción en base a los 4 ajustes. Para ello se emplea el 15 % de la muestra restante, correspondiente al consumo de 91 clientes del subgrupo analizado, con el objetivo de validar cada uno de los modelos mediante predicción.

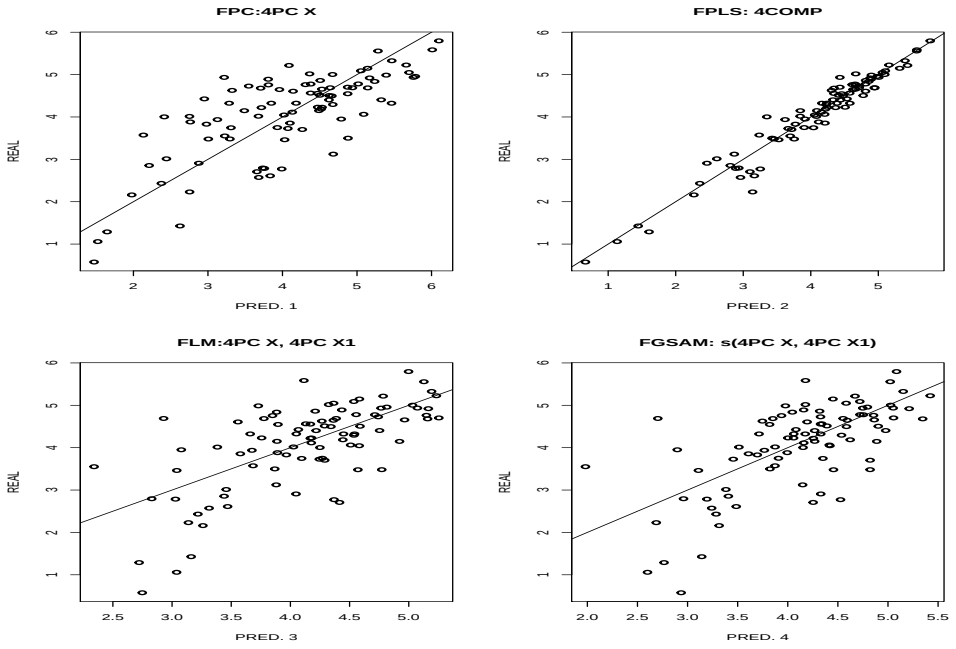

Figura 3.44: Resultados gráficos de predicción frente a consumo real en Black Friday por los modelos funcionales FPC (Modelo 1), FPLS (Modelo 2), FLM con derivada (Modelo 3) y FGSAM (Modelo 4).

En la Figura 3.44 se muestran la predicciones obtenidas mediante cada uno de los modelos frente a los importes (transformados por logaritmos) de consumo reales durante San Valentín del conjunto de 91 clientes adultos 2. De nuevo, el modelo 2 de mínimos cuadrados parciales, que a su vez se corresponde con el de coeficiente R^2 más elevado, es el que realiza una predicción notablemente más acertada en comparación al resto de modelos. El error cuadrático medio cometido en la predicción empleando cada uno de los modelos, que puede consultarse en el Cuadro 3.36, corrobora el resultado gráfico.

MODELO	MSE
1	0.541
2	0.053
3	0.559
4	0.569

Cuadro 3.36: Errores cuadráticos medios cometidos en la predicción mediante los 4 modelos funcionales ajustados: FPC, FPLS, FLM y FGSAM, correspondientes a los datos transformados por logaritmos.

En el Cuadro 3.37 mostramos la media y mediana de la diferencia, en valor absoluto, entre el consumo real y el obtenido mediante predicción al deshacer la transformación logarítmica. Observamos que con el modelo 2 nos desviamos de media 10.15 € en la predicción del gasto en Black Friday y en la mitad de la muestra analizada menos de 6.26 €.

MODELO	Media	Mediana
1	48.82 €	28.51 €
2	10.15 €	6.26 €
3	35.52 €	25.23 €
4	35.54 €	21.35 €

Cuadro 3.37: Media y mediana de las diferencias entre valor real y predicción (en euros) del gasto en Black Friday mediante los 4 modelos ajustados.

En este punto, resulta interesante conocer qué tan efectiva es la segmentación propuesta para este subgrupo en particular y hasta que punto es necesaria. Para ello desarrollamos un estudio de predicción cruzada, mediante el que predeciremos el consumo realizado en el Black Friday por la muestra de 91 clientes lucenses en base a modelos de otros subgrupos del sector adultos 2. Los ajustes empleados en este punto serán los de los clientes de entre 46 y 59 años que realizan sus compras en Ourense, las clientas de ese rango de edad que compran en Lugo y las que lo hacen en Pontevedra.

MODELO	Hombre Lugo	Hombre Ourense	Mujer Lugo	Mujer Pontevedra
1	0.524	0.609	0.732	0.545
2	0.053	0.075	0.057	0.056
3	0.359	0.478	0.466	0.413
4	0.364	0.474	0.470	0.408

Cuadro 3.38: Errores cuadráticos medios cometidos tras el estudio de predicción cruzada en base a modelos de diferentes subgrupos del sector adultos 2.

En el Cuadro 3.38 se muestran los errores cuadráticos medios cometidos en la predicción cruzada en comparación con los obtenidos en la predicción empleando los modelos correspondientes al subgrupo de los adultos 2 varones de Lugo. Fijando el modelo, observamos que los mejores resultados se obtienen en el caso de emplear el modelo específico para el subgrupo en cuestión; no obstante, para ambos grupos de mujeres, los resultados obtenidos por el modelo 2 se asemejan bastante. En este caso, estos resultados apoyan la segmentación de este subgrupo por las variables sexo y localidad.

En los Cuadros 3.39 y 3.40 se muestran los resultados de regresión (R^2) y predicción (MSE) del consumo durante el Black Friday para la totalidad de subgrupos. Para cada uno de ellos, se resalta en rojo el coeficiente R^2 más elevado y en azul, el modelo para el que se comete menos error en la predicción. Vemos que en todos los casos, los mejores resultados se obtienen con el ajuste PLS.

		JÓVENES				**ADULTOS 1**				**ADULTOS 2**			
L	M	***Hombre***		***Mujer***		***Hombre***		***Mujer***		***Hombre***		***Mujer***	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
A Coruña	1	0.53	0.607	0.69	0.599	0.53	0.668	0.53	0.669	0.61	0.601	0.63	0.617
	2	0.95	0.091	0.94	0.064	0.95	0.058	0.95	0.064	0.95	0.046	0.95	0.052
	3	0.63	0.452	0.62	0.450	0.65	0.388	0.55	0.502	0.62	0.454	0.65	0.361
	4	0.64	0.445	0.63	0.439	0.66	0.383	0.57	0.482	0.63	0.440	0.66	0.351
Lugo	1	0.87	0.822	0.60	0.849	0.69	0.782	0.67	0.618	0.60	0.541	0.60	0.649
	2	0.96	0.101	0.95	0.120	0.95	0.079	0.95	0.055	0.95	0.053	0.95	0.049
	3	0.82	0.091	0.41	0.614	0.57	0.592	0.67	0.418	0.48	0.559	0.56	0.488
	4	-	-	0.49	0.648	0.58	0.591	0.68	0.405	0.53	0.569	0.58	0.480
Ourense	1	0.80	0.711	0.57	0.445	0.57	0.785	0.54	0.609	0.69	0.607	0.61	0.508
	2	0.94	0.045	0.95	0.052	0.95	0.085	0.94	0.060	0.96	0.079	0.95	0.050
	3	0.77	0.345	0.49	0.489	0.65	0.391	0.57	0.450	0.67	0.513	0.51	0.555
	4	-	-	0.55	0.432	0.67	0.373	0.58	0.435	0.69	0.476	0.54	0.545
Pontevedra	1	0.67	0.823	0.64	0.559	0.54	0.684	0.58	0.631	0.62	0.638	0.62	0.548
	2	0.95	0.089	0.94	0.071	0.95	0.062	0.95	0.055	0.95	0.062	0.95	0.054
	3	0.71	0.339	0.50	0.515	0.53	0.489	0.57	0.470	0.71	0.301	0.62	0.412
	4	0.72	0.321	0.51	0.513	0.54	0.479	0.59	0.451	0.71	0.299	0.63	0.396
Resto	1	0.51	0.994	0.64	0.667	0.62	0.790	0.68	0.805	0.71	1.048	0.55	1.105
	2	0.94	0.077	0.94	0.085	0.95	0.070	0.94	0.086	0.94	0.084	0.95	0.082
	3	0.41	0.833	0.60	0.568	0.59	0.634	0.71	0.439	0.43	0.724	0.52	0.390
	4	0.46	0.811	0.61	0.565	0.61	0.661	0.72	0.465	0.50	0.752	0.56	0.376

Cuadro 3.39: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo durante la semana de Black Friday de los subgrupos de clientes jóvenes, adultos 1 y adultos 2.

	SÉNIOR			
M	***Hombre***		***Mujer***	
	R^2	MSE	R^2	MSE
1	0.56	0.670	0.58	0.524
2	0.95	0.050	0.95	0.058
3	0.63	0.367	0.59	0.460
4	0.63	0.359	0.61	0.423

Cuadro 3.40: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo durante la semana de Black Friday de los subgrupos de clientes sénior.

3.4.4. Navidad

Por último, trataremos de analizar el consumo en Navidad. A pesar de que tiene lugar en las últimas semanas del año, es sabido que el gasto en muchas ocasiones empieza con antelación, ya no tanto el ocasionado por la compra de productos alimenticios pero sí en regalos. De hecho, los descuentos del Black Friday potencian el inicio de las compras navideñas. Es por eso que hemos decidido agrupar el gasto del mes de diciembre, realizando el promedio de las semanas de la 49 a la 53 del año 2015. El subgrupo empleado en esta ocasión es el de las clientas de entre 60 y 65 años de la entidad, pues recordemos que en el subgrupo sénior no se hizo distinción por sexo y localidad. De la muestra completa, formada por 4 565 mujeres, se empleará el 85 % para la regresión y el 15 % restante en la predicción.

En el Cuadro 3.41 se muestran los resultados obtenidos tras el ajuste del modelo funcional de componentes principales, en el que se emplean 5 componentes que explican el 53.65 % de la variabilidad. El coeficiente R^2=0.20 es sumamente bajo en este caso.

MODELO 1	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	4.364	0.012988056	335.986	$< 10^{-3}$
PC1	-0.161	0.005	-30.205	$< 10^{-3}$
PC10	-0.082	0.014	-5.698	$< 10^{-3}$
PC2	-0.047	0.011	-4.633	$< 10^{-3}$
PC3	-0.042	0.012	-4.071	$< 10^{-3}$
PC9	-0.047	0.013	-3.493	$< 10^{-3}$
$R^2 = 0.204$	53.65 % variabilidad explicada con 5 PC's			$\lambda = 0.5$
Varianza residual: 0.654 en 3874 grados de libertad				

Cuadro 3.41: Resultados numéricos del modelo 1 en el análisis del gasto en Navidad.

En cuanto a la estimación del parámetro β del modelo 1, observamos en la Figura 3.45 que el gasto en Navidad viene determinado por una tendencia ascendente del consumo en las semanas previas al evento, salvo ciertas bajadas puntuales en las semanas 9, 20, 31 y 41.

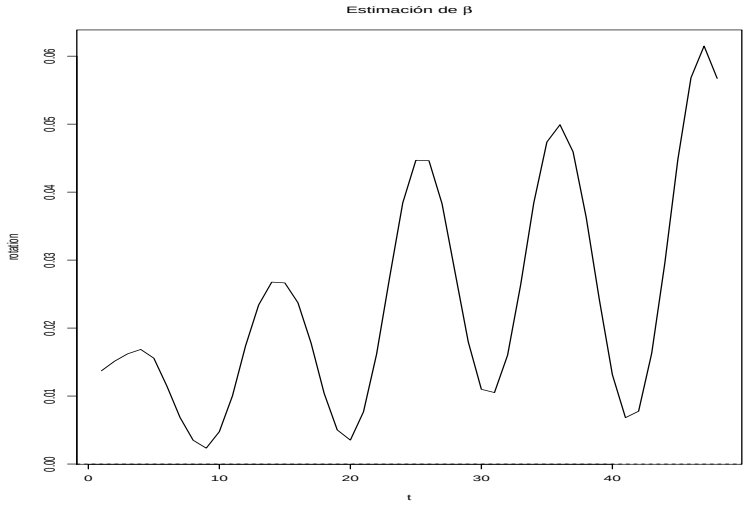

Figura 3.45: Estimación del parámetro de regresión β por el modelo 1 (Navidad).

Mediante la Figura 3.46 presenciamos un gran número de curvas atípicas e influyentes, lo que nos hace pensar que quizá sea debido a que se trata de un subgrupo muy heterogéneo en cuanto a localidad y actividad. Además, se observa problemas evidentes en el ajuste lineal derivados del coeficiente R^2 tan bajo obtenido. Dado que este modelo no realiza un ajuste precisamente bueno, habrá que tener precaución con el diagnóstico en base a él.

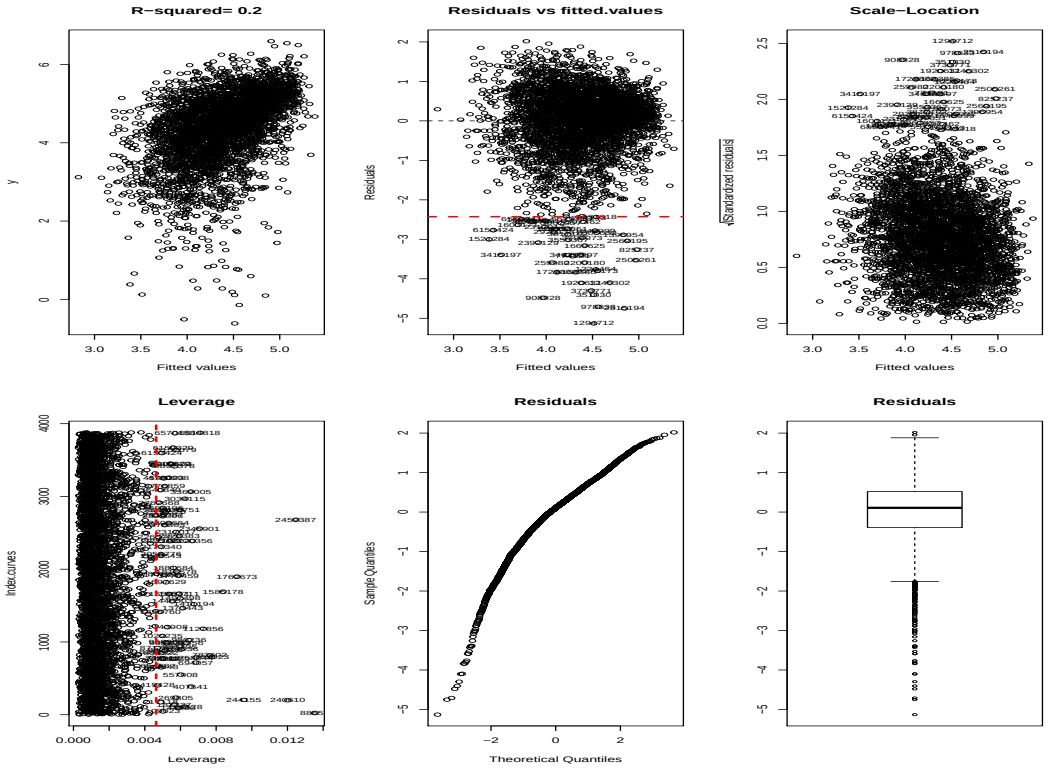

Figura 3.46: Gráficas del modelo 1 para el análisis de gasto durante Navidad. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

Tras el ajuste del modelo 2, el modelo funcional de mínimos cuadrados parciales, obtenemos los resultados que se muestran en el Cuadro 3.42. Las 4 componentes empleadas en el ajuste son significativas; no obstante, se obtiene un coeficiente R^2=0.33, que a pesar de ser mayor que el obtenido en el modelo 1, es muy bajo. De nuevo, mediante este modelo se detecta la presencia de numerosas curvas atípicas e influyentes, tal como se aprecia en la Figura 3.47, en la que también se observa el claro incumplimiento de la hipótesis de normalidad de los residuos.

MODELO 2	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	4.363	0.012	365.881	$< 10^{-3}$
PLS1	0.157	0.004	42.509	$< 10^{-3}$
PLS2	0.292	0.011	25.694	$< 10^{-3}$
PLS3	0.194	0.011	17.092	$< 10^{-3}$
PLS4	0.141	0.014	10.391	$< 10^{-3}$
$R^2 = 0.331$	Varianza residual: 0.552 en 3859.273 grados de libertad			

Cuadro 3.42: Resultados numéricos del modelo 2 en el análisis del gasto en Navidad.

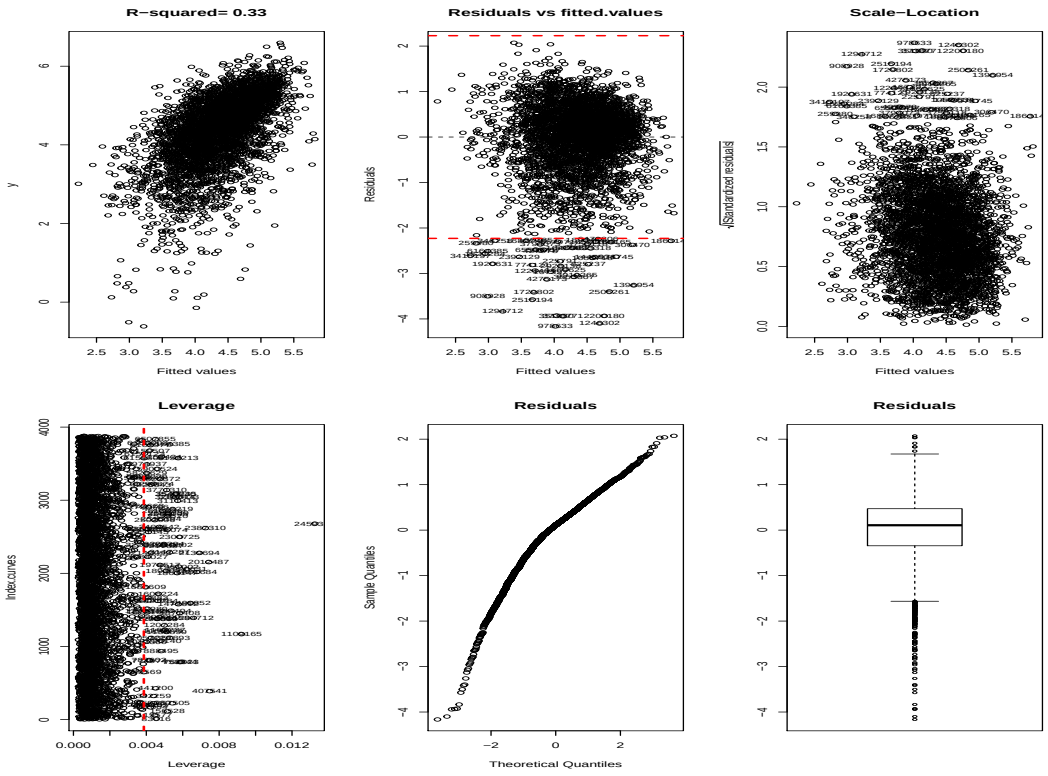

Figura 3.47: Gráficas del modelo 2 para el análisis de gasto durante Navidad. De izquierda a derecha: ajuste lineal (R^2), media cero, homocedasticidad, apalancamientos, normalidad de residuos y *box-plot* de residuos.

La representación de la estimación del parámetro β en la Figura 3.48 es bastante plana, con excepción de la influencia del gasto de la semana 40 y la del bajo consumo en la 45 en la explicación el gasto navideño.

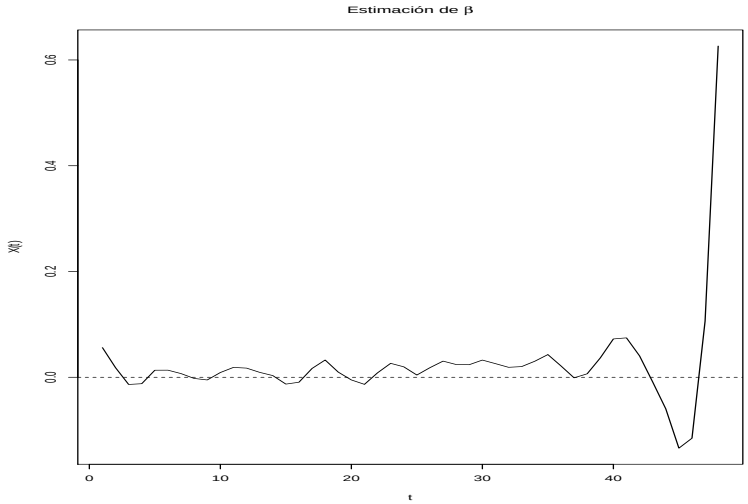

Figura 3.48: Estimación del parámetro de regresión β por el modelo 2 (Navidad).

En el Cuadro 3.43 se muestran los resultados obtenidos tras realizar el ajuste por el modelo 3. Tanto las 4 componentes principales empleadas en la representanción del dato funcional como en su derivada son significativas. Se obtiene un coeficiente R^2 ajustado de 0.23, inferior al del modelo FPLS.

MODELO 3	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	4.364	0.012	9340.451	$< 10^{-3}$
X.PC1	-0.112	0.003	-31.326	$< 10^{-3}$
X.PC2	-0.039	0.006	-5.797	$< 10^{-3}$
X.PC3	-0.025	0.006	-3.713	$< 10^{-3}$
X.PC4	-0.021	0.007	-3.057	$< 10^{-3}$
X1.PC1	0.104	0.021	4.830	$< 10^{-3}$
X1.PC2	0.112	0.017	6.262	$< 10^{-3}$
X1.PC3	-0.148	0.019	-7.821	$< 10^{-3}$
X1.PC4	0.124	0.016	7.557	$< 10^{-3}$
$R^2 = 0.225$	R^2 ajustado = 0.224		*p-valor*: $< 10^{-3}$	
Error estándar residual: 0.798 en 3871 DF			Estad. F: 140.7 en 8 y 3871 DF	

Cuadro 3.43: Resultados numéricos del modelo 3 en el análisis del gasto en Navidad.

En la Figura 3.49 se observa, de nuevo, la presencia de alguna curva influyente y la clara falta de normalidad en los residuos. Las estimaciones asociadas a este ajuste se muestan en la Figura 3.50, correspondiéndose la curva roja a los parámetros $\hat{\beta}_j$ y la verde a los $\hat{\beta}_k$ de la ecuación (3.1).

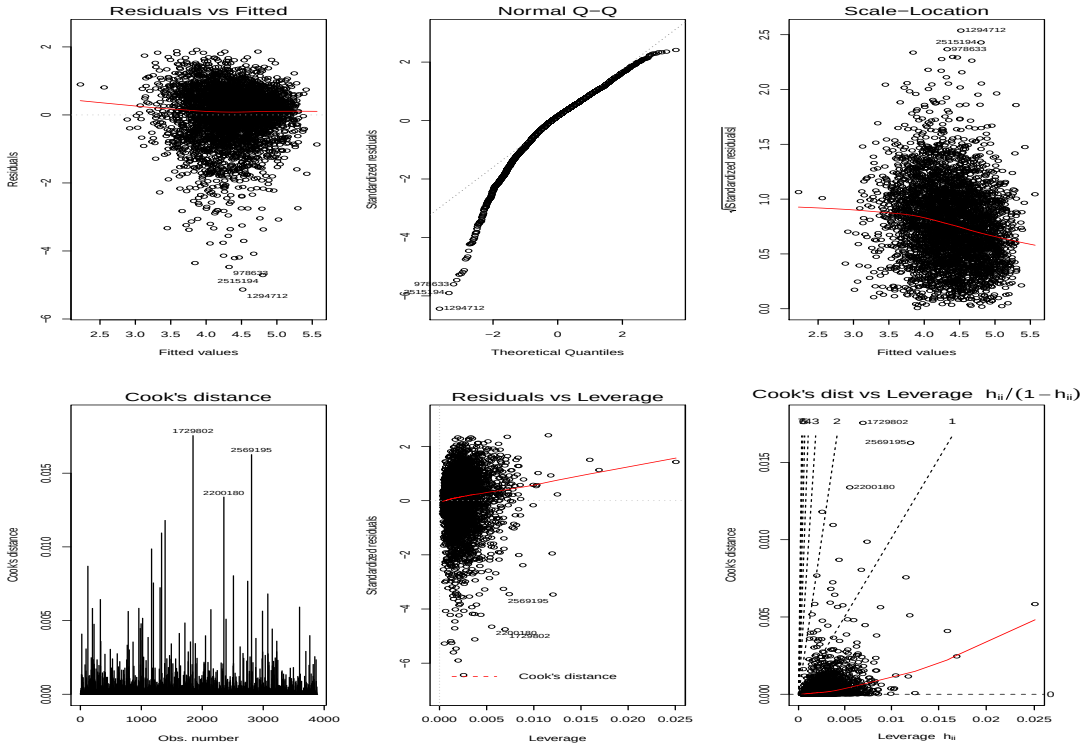

Figura 3.49: Gráficas del modelo 3 para el análisis de gasto durante Navidad. De izquierda a derecha: media cero, gráfico de normalidad Q-Q, homocedasticidad, distancia de Cook y residuos y distancia de Cook frente a apalancamientos.

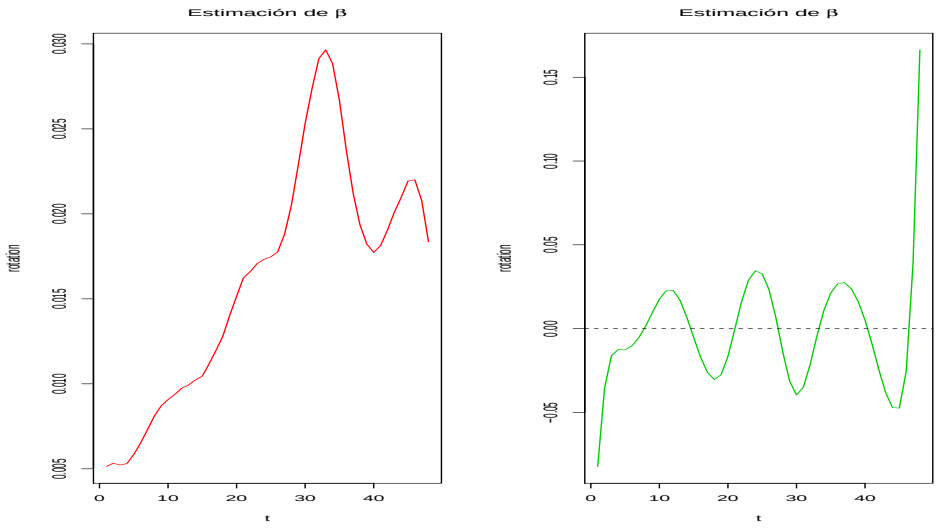

Figura 3.50: Estimación del parámetro de regresión β por el modelo 3 (izquierda) y estimación correspondiente a la derivada (derecha) en el análisis de gasto de Navidad.

Llegados a este punto, decidimos estudiar la influencia de las variables *actividad* y *localidad*, con el objetivo de determinar si la no segmentación del subgrupo de las mujeres sénior en base a estas variables fue una decisión acertada. La muestra de entrenamiento, se compone por 636 clientas inactivas, 2 412 activas ocupadas, 52 activas paradas y de 780 de ellas se desconoce la actividad que desempeñan. Dado que las mujeres activas ocupadas representan el 62 % de la muestra se realiza el ajuste diferenciado para este subgrupo y los restantes.

Se muestran en el Cuadro 3.44 los resultados obtenidos tras el ajuste de la variante del modelo 3 considerando la actividad contable de las clientas objeto de estudio. En este caso, para el subgrupo de clientas sénior, podemos prescindir de la segmentación en base a esta variable, puesto que no implicaría mejoras significativas en el ajuste, por lo menos en el análisis del gasto navideño, en comparación al modelo que no considera dicha variable.

MODELO 3	$\mathcal{X}$	$\mathcal{X}^{(1)}$	R^2	R^2 ajust.	*p-valor*
OCUPADAS	1:4	1:4	0.237	0.231	$< 10^{-3}$
- OCUPADAS	1-2-4	1:4	0.228	0.224	$< 10^{-3}$

Cuadro 3.44: Resumen del modelo 3 para el análisis de gasto en Navidad de las clientas sénior activas ocupadas (de forma separada) frente al resto.

En cuanto a la variable localidad, la muestra de entrenamiento, se compone por 1 931 clientas que realizan sus compras en A Coruña, 346 que lo hacen en Lugo, 291 en Ourense, 1 283 en Pontevedra y 29 en el resto de España. Dado que A Coruña y Pontevedra representan la mayor parte de la muestra, decidimos ajustar el modelo 3 de forma diferenciada para cada uno de los 2 subgrupos y para el que engloba las 3 localidades restantes.

En el Cuadro 3.45 se muestran los resultados obtenidos tras este ajuste. Vemos que, pese a que se observan diferencias relativas al número de componentes empleadas, en ninguno de los 3 casos se observan diferencias del ajuste lineal del modelo. Por lo tanto, en el subgrupo de clientas sénior, prescindiremos de la segmentación en base a la variable localidad, en lo que a la explicación del consumo navideño se refiere.

MODELO 3	$\mathcal{X}$	$\mathcal{X}^{(1)}$	R^2	R^2 ajust.	*p-valor*
A CORUÑA	1:4	1:4	0.233	0.232	$< 10^{-3}$
PONTEVEDRA	1:2	1-3	0.233	0.231	$< 10^{-3}$
- COR:PONTE	1:2	1:4	0.234	0.231	$< 10^{-3}$

Cuadro 3.45: Resumen del modelo 3 para el análisis de gasto en Navidad de las clientas sénior que compran en A Coruña y Pontevedra (de forma separada) frente al resto.

Para finalizar con el ajuste de modelos, mostramos en el Cuadro 3.46 los resultados del FGSAM. A pesar de resultar significativas todas las componentes del modelo, el R^2=0.23 obtenido es de nuevo inferior al del modelo 2, el cuál ya era de por sí bastante bajo. La *deviance* explicada es del 23.3 %.

MODELO 4	Estimación	Error estándar	*t valor*	$Pr(>\|t\|)$
Intercepto	4.363	0.012	341.8	$< 10^{-3}$
	EDF	Ref. DF	F	p-value
s(X.PC1)	2.530	3.222	299.433	$< 10^{-3}$
s(X.PC2)	2.156	2.784	13.248	$< 10^{-3}$
s(X.PC3)	1.541	1.936	8.949	$< 10^{-3}$
s(X.PC4)	1.000	1.000	9.338	$< 10^{-3}$
s(X1.PC1)	1.000	1.000	20.660	$< 10^{-3}$
s(X1.PC2)	1.000	1.000	39.962	$< 10^{-3}$
s(X1.PC3)	4.472	5.605	12.416	$< 10^{-3}$
s(X1.PC4)	2.130	2.766	21.196	$< 10^{-3}$
R^2 ajustado = 0.23		*Deviance* explicada = 23.3 %		

Cuadro 3.46: Resultados numéricos del modelo 4 en el análisis del gasto en Navidad.

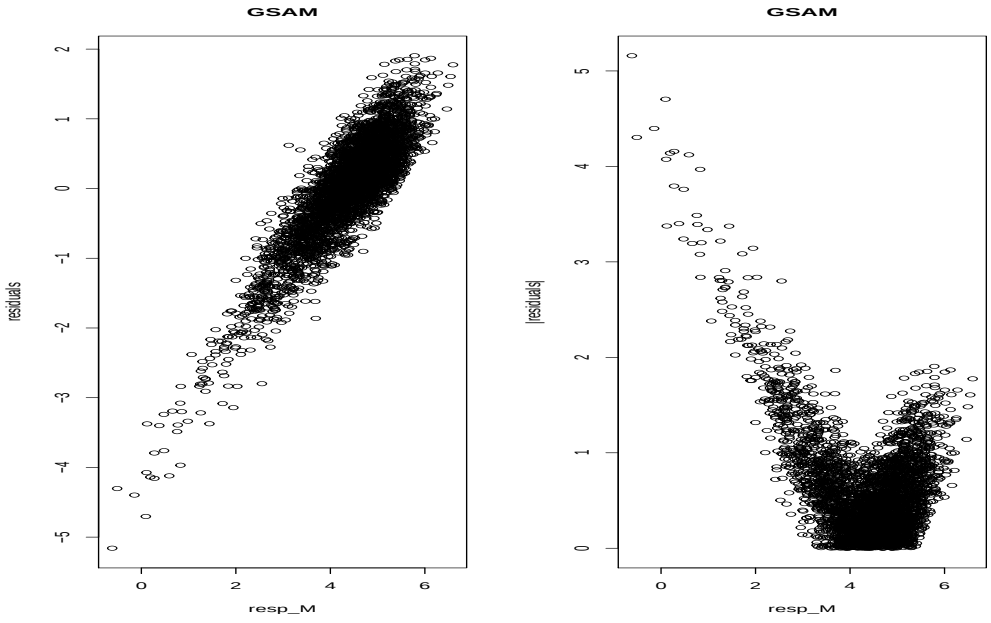

Figura 3.51: Gráfica de los residuos del modelo 4 (izquierda) y valor absoluto de los mismos (derecha) en el análisis de gasto de Navidad.

Se presenta en la Figura 3.51 el gráfico de homocedasticidad de los residuos, mediante el que observamos el claro incumplimiento de la hipótesis por medio de una clara heterocedasticidad.

Estos resultados nos hacen pensar en que el problema de estos ajustes pueda residir en la consideración de todo el mes de diciembre como la respuesta en el modelo que intenta explicar el consumo navideño. Alternativamente, cabría centrar el análisis considerando que el consumo se da, por ejemplo, en dos semanas concretas del mes de diciembre; no obstante, al estar tan próximos los descuentos del Black Friday, es complicado discernir el período exacto de consumo.

A continuación, se muestran los resultados de predicción en base a los cuatro modelos ajustados. Para ello se emplea el 15 % de la muestra restante, correspondiente al consumo de 685 clientas sénior. En la Figura 3.52 se representan la predicciones obtenidas mediante cada uno de los modelos frente a los importes (transformados por logaritmos) de consumo reales durante el mes de diciembre de la muestra de validación. En este caso, ninguna de las predicciones realizadas parece en conjunto realmente buena; no obstante, semeja que la que menos se desvía del consumo real de los clientes es la correspondiente al modelo 2 de mínimos cuadrados parciales.

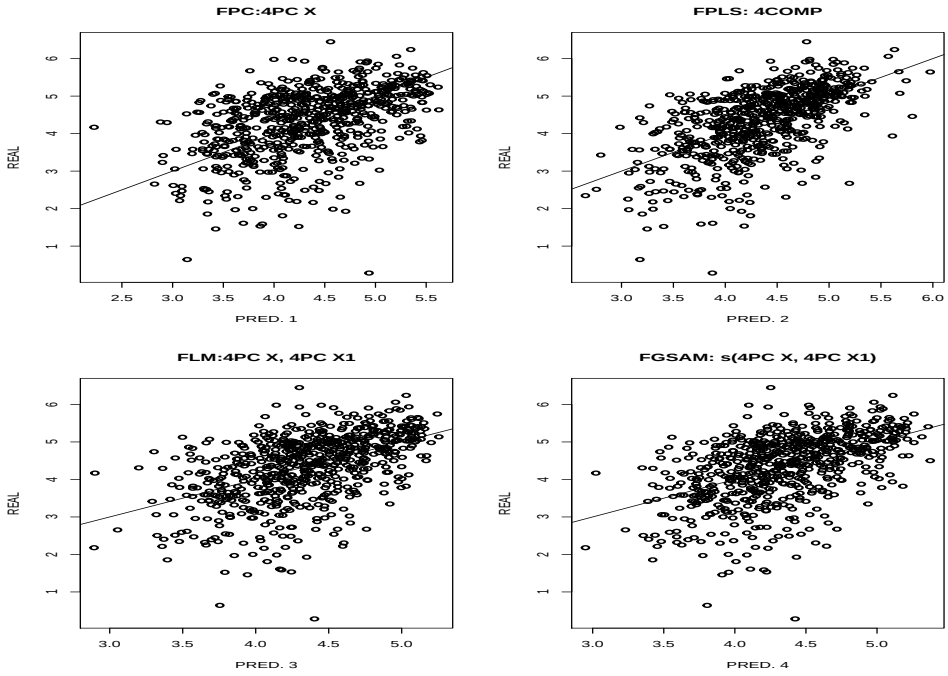

Figura 3.52: Resultados gráficos de predicción frente a consumo real en Navidad por los modelos funcionales FPC (Modelo 1), FPLS (Modelo 2), FLM con derivada (Modelo 3) y FGSAM (Modelo 4).

MODELO	MSE
1	0.674
2	0.526
3	0.613
4	0.613

Cuadro 3.47: Errores cuadráticos medios cometidos en la predicción mediante los 4 modelos funcionales ajustados FPC, FPLS, FLM y FGSAM, correspondientes a los datos transformados por logaritmos.

El error cuadrático medio cometido en la predicción empleando cada uno de los modelos puede consultarse en el Cuadro 3.47. Se comprueba la similitud entre ellos, siendo ligeramente inferior en el caso del modelo 2. En el Cuadro 3.48 mostramos la media y mediana de la diferencia, en valor absoluto, entre el consumo real y el obtenido mediante predicción al deshacer la transformación logarítmica. Con el modelo 2 estamos cometiendo un error medio de 45.61 €. Se obviará la predicción cruzada en este caso, puesto que cabe ser cautelosos respecto a estos resultados debido al incumplimiento de las hipótesis subyacentes a los modelos que anticipamos previamente.

MODELO	Media	Mediana
1	52.17 €	37.60 €
2	45.61 €	32.03 €
3	49.26 €	35.05 €
4	49.21 €	35.69 €

Cuadro 3.48: Media y mediana de las diferencias entre valor real y predicción (en euros) del gasto en Navidad mediante los 4 modelos ajustados.

		JÓVENES				**ADULTOS 1**				**ADULTOS 2**			
L	M	***Hombre***		***Mujer***		***Hombre***		***Mujer***		***Hombre***		***Mujer***	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
A Coruña	1	0.15	0.715	0.12	0.878	0.18	0.858	0.20	0.855	0.22	0.807	0.21	0.730
	2	0.31	0.655	0.28	0.735	0.32	0.712	0.31	0.675	0.36	0.662	0.35	0.571
	3	0.15	0.708	0.11	0.838	0.19	0.806	0.21	0.781	0.21	0.755	0.22	0.649
	4	0.20	0.786	0.11	0.848	0.20	0.793	0.22	0.773	0.23	0.750	0.24	0.652
Lugo	1	0.36	2.752	0.08	1.512	0.18	0.952	0.18	0.800	0.25	1.052	0.22	0.728
	2	0.56	1.545	0.35	1.223	0.36	0.782	0.33	0.643	0.38	0.931	0.36	0.564
	3	0.36	0.969	0.14	1.136	0.26	0.868	0.18	0.730	0.26	0.967	0.24	0.628
	4	-	-	0.23	1.192	0.28	0.897	0.20	0.749	0.30	0.976	0.26	0.619
Ourense	1	0.23	2.427	0.17	0.748	0.16	0.769	0.18	0.885	0.25	0.883	0.26	0.864
	2	0.61	1.827	0.38	0.634	0.31	0.653	0.30	0.670	0.35	0.715	0.37	0.623
	3	0.23	1.426	0.19	0.735	0.18	0.667	0.21	0.814	0.28	0.747	0.27	0.776
	4	-	-	0.34	0.883	0.21	0.678	0.23	0.800	0.30	0.773	0.29	0.773
Pontevedra	1	0.26	0.985	0.17	1.021	0.17	0.927	0.17	0.898	0.20	0.757	0.21	0.739
	2	0.35	0.852	0.31	0.913	0.30	0.743	0.31	0.695	0.33	0.630	0.34	0.585
	3	0.24	0.817	0.17	0.958	0.18	0.880	0.19	0.834	0.22	0.733	0.22	0.685
	4	0.36	0.855	0.20	0.998	0.19	0.880	0.20	0.830	0.23	0.739	0.24	0.680
Resto	1	0.11	1.294	0.10	1.385	0.16	1.199	0.20	1.252	0.19	0.907	0.29	0.806
	2	0.27	1.143	0.26	1.054	0.31	0.939	0.32	0.991	0.34	0.885	0.43	0.636
	3	0.11	1.135	0.19	1.347	0.22	1.202	0.21	1.247	0.18	0.806	0.38	0.603
	4	0.14	1.144	0.23	1.363	0.24	1.220	0.23	1.261	0.23	0.854	0.41	0.749

Cuadro 3.49: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo durante el período de Navidad de los subgrupos de clientes jóvenes, adultos 1 y adultos 2.

	SÉNIOR			
M	*Hombre*		*Mujer*	
	R^2	MSE	R^2	MSE
1	0.20	0.807	0.20	0.674
2	0.35	0.680	0.33	0.526
3	0.21	0.752	0.22	0.613
4	0.22	0.750	0.23	0.613

Cuadro 3.50: Resultados del estudio de regresión (R^2) y predicción (MSE) sobre el consumo durante el período de Navidad de los subgrupos de clientes sénior.

En los Cuadros 3.49 y 3.50 se muestran los resultados de regresión (R^2) y predicción (MSE) del consumo durante Navidad para los diferentes modelos y la totalidad de subgrupos. Para cada uno de ellos, se resalta en rojo el coeficiente R^2 más elevado y en azul, el modelo para el que se comete menos error en la predicción. Vemos que independientemente del subgrupo considerado, en todos los casos se obtienen coeficientes R^2 bajos y elevados errores de predicción, pues recordemos que están expresados en escala logarítmica.

Estos resultados afianzan la idea de que el problema de estos ajustes reside en la respuesta considerada, y nada tiene que ver con la mayor o menor segmentación de la muestra de clientes en subgrupos más o menos pequeños.

3.5. Exigencias computacionales

A lo largo de este trabajo se han sucedido diferentes etapas que han facilitado el manejo del conjunto de datos, desde la reducción de dimensión por medio de diversos métodos hasta la segmentación de la muestra en base a diferentes variables, creando así submuestras más sencillas de analizar, definidas en los Cuadros 3.7, 3.8, 3.9 y 3.10, que a su vez facilitan la clasificación del cliente.

En la exploración del conjunto de datos funcionales, se empleó un ordenador con capacidades básicas; no obstante, llegado el análisis de regresión, fue necesario recurrir a una máquina de mayor capacidad para realizar el estudio de ciertos subgrupos concretos. Las características de procesador y memoria de dichos ordenadores son las siguientes:

1. Intel (R) Core (TM) i5-3337U CPU @ 1.80GHz 1.80GHz || RAM: 4G
2. Intel (R) Core (TM) i7-3770 CPU @ 3.40GHz 3.40GHz || RAM: 32GB

SUBGRUPO	REGRESIÓN	PREDICCIÓN
Jóvenes	2 h	30 min
Adultos 1	50 h	30 h
Adultos 2	20 h	10 h
Sénior	4 h	1 h
TOTAL	76 h	41 h 30 min

Cuadro 3.51: Tiempos de ejecución aproximados del análisis de regresión y el de predicción para cada subgrupo de edad.

Para los subgrupos de adultos 1 y 2 de clientes que realizan sus compras en A Coruña y Pontevedra, tanto hombres como mujeres, y para el subgrupo sénior femenino, fue necesaria la utilización de la máquina con procesador i7 y mayor memoria RAM; mientras que para los restantes subgrupos fue suficiente con el ordenador de menor capacidad.

En el Cuadro 3.51 se muestran los tiempos aproximados de ejecución en el análisis de regresión y predicción de cada subgrupo de edad, donde cabe destacar los elevados consumos de tiempo para el subgrupo de adultos 1.

Capítulo 4

Conclusiones y perspectivas de futuro

Del estudio de caso realizado a petición de una entidad bancaria concluimos que se han cumplido los objetivos planteados. A partir de un problema inicial de carácter complejo, se ha demostrado que la aplicación del análisis de datos funcionales es válida como aproximación a la solución; al menos, para los principales casos de uso planteados en el negocio de medios de pago. Al mismo tiempo, el estudio resulta prometedor para otras situaciones de interés, sin detrimento de que la solución tentativa requiera un mayor nivel de desarrollo para su puesta en producción en los sistemas de la empresa.

De una manera eminentemente práctica, el trabajo proporciona una herramienta para clasificar a los usuarios de tarjetas en base a su perfil sociodemográfico más básico y una serie de modelos para predecir el gasto esperado en al menos 3 de las 4 campañas anuales de consumo que han sido analizadas. Con ello se han podido confirmar las sospechas de que existen grupos de usuarios de tarjeta que aumentan su nivel de consumo en épocas concretas, asociadas a campañas promocionales, vacaciones y épocas de descuentos.

1. De un lado, ha quedado mostrada la utilidad del análisis de datos funcionales para el tratamiento de las bases de datos correspondientes a los pagos de tarjeta. Pese a las reservas previas sobre la adecuación del FDA a datos que carecen *per se* de forma funcional, las estrategias adoptadas en la preparación de los datos se han mostrado existosas para aportar soluciones al problema. Con ello se ha contribuido a introducir en el ámbito de la empresa una serie de prácticas analíticas con las que no se había trabajado hasta la fecha, dejando entrever las posibilidades que ofrecen para aumentar la comprensión de los patrones de uso de sus tarjetas y anticipar los niveles de gasto en secuencias semanales.

2. De otro, se han obtenido resultados prácticos en cuanto a la segmentación de la base de clientes de la entidad bancaria, ampliando la segmentación de edad prestablecida por la empresa, cuya validez se ha confirmado y pormenorizado, pasando de 4 grandes grupos de edad a 32 subgrupos que resultan de la combinación de su edad, su sexo y la localización de sus compras. Las evidencias aportadas en cuanto a la existencia de diferencias significativas entre distintos subgrupos de usuarios ha suscitado gran interés entre los responsables del negocio, puesto que abre la posibilidad de dar un tratamiento diferenciado a su base de clientes, a la hora de diseñar productos específicos para cada público, elaborar protocolos de comercialización y venta de tarjetas, así como elaborar campañas de marketing adaptadas a cada perfil.

3. Finalmente, los modelos ajustados permiten realizar predicciones útiles para los escenarios de consumo de diferentes épocas del año, arrojando estimaciones que se comportan razonablemente

bien en la evaluación de su error y conforme a lo esperado en la bondad del ajuste. En este sentido, se aportan soluciones para más del 99 % de los clientes en su consumo durante San Valentín, Semana Santa y Black Friday. Por el contrario, la mezcla de patrones que subyacen a los datos registrados durante la época de Navidad impide que se puedan hacer predicciones robustas y lleva a pensar que la selección del período estudiado, en éste caso concreto, no ha resultado satisfactoria para el objetivo de confirmar la hipótesis de partida. Cabe atribuir este hecho bien a la falta de variables que informen los modelos, bien a que se trate de un período merecedor de un tratamiento más pormenorizado (posiblemente, fragmentando el punto de discretización del período navideño en un grid semanal más fino).

Los niveles de gasto que reflejan los datos permiten hacer observaciones sobre los grupos que gastan permanentemente más que otros. Como es conocido y se observa en las representaciones gráficas, el gasto medio aumenta con la edad y esta pauta se mantiene regular a lo largo de todo el año. Así mismo, métricas descriptivas ya manejadas en la entidad dan muestra de que existen picos de consumo en eventos comerciales como las semana de Black Friday; que el consumo juvenil disminuye en períodos escolares vacacionales, como la Semana Santa y la Navidad; o que los niveles de gasto están muy relacionados con el asentamiento territorial de la Galicia rural y urbana. Aunque es obvio que la base de datos permitiría hacer inferencias de este tipo, sin embargo, no ha existido la necesidad de profundizar en la descripción de dichos subgrupos, debido a que se trata de métricas conocidas para la entidad y sujetas al cambio de coyuntura que están experimentando los niveles de consumo en la realidad actual. De mayor interés han sido las diferencias entre e intragrupos y la posibilidad de hacer predicciones bien ajustadas sobre los mismos.

La pauta de uso de los modelos se proporciona en el Cuadro 4.1, donde se muestra para cada subgrupo de clientes cuál es el escenario donde se obtienen ajustes adecuados. Según su necesidad, el usuario final podrá decidir la época del año en la que contactar a su cliente apoyándose en la desviación esperada para la predicción puntual del gasto. Por ejemplo, con los hombres jóvenes de Coruña (HJC), parece que el error cometido en los ajustes para la semana santa es menor que en el resto de períodos analizados, en concreto, inferior a 3 € en el 50 % de los clientes (error de predicción expresado en Mediana). Con estos resultados se propondrá una actuación o no, en función de criterios de negocio que determinarán comunicaciones de marketing, descuentos u ofertas para aquellos clientes cuya predicción de gasto esperado permita amortizar la inversión en incentivos por la vía de comisiones de servicio.

En cuanto a las recomendaciones para el mayor aprovechamiento del trabajo, cabe mencionar el valor de disponer de datos para un período más amplio en el tiempo, de tal forma que contásemos con las transacciones de los 12 meses anteriores a cada evento de interés. La actualización de los modelos ajustados con dicha información es una vía a explorar para aumentar su precisión y, sobre todo, para proporcionar predicciones dinámicas que vayan reajustando los coeficientes de los modelos validados con cada nueva entrada de datos. En este sentido, desde una aproximación *machine learning*, se propone replicar el modelado con datos entrantes de nuevos períodos de consumo, conjeturando que los modelos que se han mostrado robustos serán aplicables a las nuevas situaciones.

Mención especial merece la exigencia computacional de la solución aportada y la conveniencia de abrir una línea de futuros desarrollos para mejorar la programación empleada de cara a reducir los tiempos de cálculo. Al respecto, queda abierta la posibilidad de paralelizar la programación empleada, así como de reducir la dimensión de los datos para poder abordar modelos más complejos pero también con mayores necesidades computacionales. En este sentido, presumimos que el ajuste FGKAM puede ser un candidato a competir con el FPLS. Así mismo, los modelos FPLS se han mostrado los más robustos en la mayoría de los escenarios, existiendo coherencia entre las medidas de evaluación aportadas y las desviaciones entre las muestras de observación y predicción. Por el contrario, los ajustes donde existen dudas sobre el cumplimiento de las hipótesis del modelo y los errores son más abultados, también presentan desviaciones entre observaciones y predicciones más abultadas.

Grupo	N	San Valentín			Semana Santa			Black Friday			Navidad		
		R^2	Me		R^2	Me		R^2	Me		R^2	Me	
J H C	442	↑	4.36	▼ [b]	↑	2.93	▼	↑	4.50	▼	↓	9.65	▼
J H L	38	↑	10.57	◆	↑	4.85	▼	↑	8.67	▼ [a]	↘	10.43	◆ [a]
J H O	39	↑	10.87	◆	↑	4.64	▼	↑	3.26	▼	↘	17.46	◆ [a]
J H P	247	↑	5.46	▼	↑	5.99	▼	↑	3.98	▼	↓	17.18	◆ [a]
J H R	452	↑	5.03	▼	↑	4.76	▼	↑	6.18	▼	↓	16.02	◆ [a]
J M C	1 085	↑	4.78	▼	↑	3.85	▼	↑	5.14	▼	↓	16.84	◆
J M L	130	↑	5.46	▼	↑	5.32	▼	↑	6.32	▼	↓	24.43	▲ [a]
J M O	145	↑	10.85	◆	↑	4.97	▼	↑	4.79	▼	↓	23.26	▲
J M P	655	↑	4.70	▼	↑	3.50	▼	↑	4.74	▼	↓	19.08	◆
J M R	554	↑	4.28	▼	↑	4.51	▼	↑	6.20	▼	↓	16.67	◆
A1 H C	6 664	↑	5.44	▼	↑	5.24	▼	↑	7.32	▼	↓	24.54	▲
A1 H L	883	↑	5.19	▼	↑	5.45	▼	↑	6.55	▼	↓	23.63	▲
A1 H O	804	↑	6.59	▼	↑	4.68	▼	↑	6.64	▼	↓	23.73	▲
A1 H P	4 950	↑	5.87	▼	↑	5.40	▼	↑	6.62	▼	↓	25.63	▲
A1 H R	1 563	↑	5.89	▼	↑	5.96	▼	↑	6.82	▼	↓	20.93	▲
A1 M C	15 868	↑	5.87	▼	↑	6.30	▼	↑	7.99	▼	↓	29.31	▲
A1 M L	2 726	↑	6.44	▼	↑	6.41	▼	↑	7.64	▼	↓	28.19	▲
A1 M O	2 488	↑	6.86	▼	↑	7.19	▼	↑	8.22	▼	↓	27.87	▲
A1 M P	12 251	↑	6.01	▼	↑	6.36	▼ [b]	↑	7.86	▼	↓	30.15	▲
A1 M R	1 266	↑	7.20	▼	↑	5.68	▼	↑	9.31	▼	↓	24.15	▲
A2 H C	4 132	↑	6.24	▼	↑	6.27	▼	↑	6.71	▼	↓	26.75	▲
A2 H L	609	↑	5.82	▼	↑	6.58	▼	↑	6.26	▼ [b]	↓	33.88	▲
A2 H O	572	↑	7.23	▼	↑	4.72	▼	↑	7.46	▼	↓	28.44	▲
A2 H P	2 800	↑	6.84	▼	↑	5.90	▼	↑	7.20	▼	↓	28.76	▲
A2 H R	299	↑	5.49	▼	↑	6.06	▼	↑	9.70	▼	↓	30.23	▲ [a]
A2 M C	10 000	↑	6.27	▼	↑	7.05	▼	↑	8.07	▼	↓	33.82	▲
A2 M L	1 824	↑	5.88	▼	↑	6.42	▼	↑	8.95	▼	↓	35.83	▲
A2 M O	1 599	↑	4.41	▼	↑	6.41	▼	↑	7.04	▼	↓	29.71	▲
A2 M P	7 224	↑	6.60	▼	↑	6.93	▼	↑	8.03	▼	↓	33.38	▲
A2 M R	167	↑	11.38	◆	↑	6.35	▼	↑	14.88	◆	↓	20.68	▲ [a]
S H	2 203	↑	6.20	▼	↑	5.49	▼	↑	9.96	▼	↓	29.57	▲
S M	4 565	↑	5.64	▼	↑	5.83	▼	↑	7.66	▼	↓	32.03	▲ [c]

Cuadro 4.1: Resumen del estudio de caso con los resultados de los modelos FPLS ajustados para cada período analizado con los subgrupos de clientes que el test ANOVA demostró significativamente diferentes en su nivel de gasto, donde J son jóvenes de 18 a 25 años, $A1$ son adultos de 26 a 45 años, $A2$ son adultos de 46 a 59 años y S son sénior mayores de 60 años; H son Hombres y M son mujeres, mientras que C, L, O, P y R son las siglas de las provincias de Coruña, Lugo, Ourense, Pontevedra y Resto de España. En cada subgrupo y período, R^2 está representado con ↑ si es superior a 0,9, con ↓ si es inferior a 0,5 y con ↘ si está comprendido entre 0,5 y 0,9. En cada subgrupo y período, si las diferencias entre las predicciones y las observaciones son inferiores a 10 € se representan con ▼. Por su parte, si las diferencias entre las predicciones y las observaciones son superiores a 20 € se representan con ▲. Con ◆ se representan las diferencias entre las predicciones y las observaciones comprendidas entre 10 € y 20 €. El indicativo [a] significa que el MSE es inferior en el modelo 3 frente al PLS. El [b] implica que podemos considerar la opción de clasificar a los jóvenes coruñeses por su actividad a la hora de predecir su consumo durante San Valentín. Finalmente, el [c] significa que podemos prescindir de la segmentación por actividad, dado que no implicaría mejoras significativas en el ajuste.

En esta competición de modelos se ha puesto especial atención a evitar la presencia de posibles sobreajustes por redudancia de información. De una manera parsimoniosa, con factores sociodemográficos estables en el tiempo, fáciles de obtener y respetuosos con la protección de datos de carácter personal, se han planteado modelos de carácter aplicado. No obstante, existe recorrido para evaluar los niveles de consumo de los usuarios de tarjeta a la luz de otras variables. En este terreno conviene mencionar la limitación que impone la calidad de los datos registrados por la entidad, existiendo una relación entre la bondad en el registro de las variables que han resultado significativas para los análisis realizados (como la edad, el sexo o la localización) y la ausencia de registro o desactualización de la información de otras variables, como es el caso de la actividad ocupacional, que no han contribuido de igual manera a acallar el ruido de los datos.

Bibliografía

[1] Aneiros-Pérez G, Vieu P (2006) Semi-functional partial linear regression. Statistics & Probability Letters, 76(11):1102-1110.

[2] Antoniadis A, Sapatinas T (2003) Wavelet methods for continuous-time prediction using hilbert-valued autoregressive processes. Journal of Multivariate Analysis, 87(1):133-158.

[3] Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. John Wiley & Sons.

[4] Benko M (2007) Functional data analysis with applications in finance. Doctoral dissertation, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

[5] Cao R (2015) Inferencia estadística con datos de gran volumen. La Gaceta de la RSME 18(2):393-417.

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

[7] Cardot H, Ferraty F, Sarda P (1999) Functional linear model. Statistics & Probability Letters, 45(1):11-22.

[8] Cardot H, Ferraty F, Sarda P (2003) Spline estimators for the functional linear model. Statistica Sinica, 13(3):571-592.

[9] Cardot H, Mas A, Sarda P (2007) Clt in functional linear regression models. Probability Theory and Related Fields, 138(3):325-361.

[10] Cardot H, Sarda P (2005) Estimation in generalized linear models for functional data via penalized likelihood. Journal of Multivariate Analysis, 92(1):24-41.

[11] Chen K, Zhang X, Petersen A, Müller HG (2015) Quantifying Infinite-Dimensional Data: Functional Data Analysis in Action. Statistics in Biosciences (Journal of the International Chinese Statistical Association). Springer, pp 1-23. `http://rdcu.be/mReW` Accedido 1 de septiembre de 2016.

[12] Cuesta-Albertos J, Febrero-Bande M (2010) A simple multiway ANOVA for functional data. TEST 19, 537-557. Springer.

[13] Cuesta-Albertos J, Nieto-Reyes A (2008) The random tukey depth. Computational Statistics & Data Analysis, 52(11):4979-4988.

[14] Cuevas A, Febrero M, Fraiman R (2004) An anova test for functional data. Computational statistics & data analysis, 47(1):111-122.

[15] 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.

[16] Dablemont S, Bellegem SV, Verleysen M (2007) Modelling and Forecasting financial time series of "tick data" by functional analysis and neural networks. Conference of Forecasting Financial Markets `http://perso.uclouvain.be/michel.verleysen/papers/ffm07sd.pdf` Accedido 1 de septiembre de 2016.

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

[18] Escabias M, Aguilera A, Valderrama M (2005) Modeling environmental data by functional principal component logistic regression. Environmetrics, 16(1):95-107.

[19] Escabias M, Aguilera A, Valderrama M (2007) Functional pls logit regression model. Computational Statistics & Data Analysis, 51(10):4891-4902.

[20] Febrero M, Galeano P, González-Manteiga W (2007) Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. Environmetrics, 19(4):331-345.

[21] Febrero-Bande M, González-Manteiga W (2013) Generalized additive models for functional data. TEST, pages 1-15.

[22] Febrero-Bande M, Oviedo M (2012) Statistical Computing in Functional Data Analysis: The R Package fda.usc. Journal of Statistical Software, 51:1-28. `https://cran.r-project.org/package=fda.usc` Accedido 1 de febrero de 2016.

[23] Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer Science & Business Media, USA.

[24] Fraiman R, Muniz G (2001) Trimmed means for functional data. Test, 10(2):419-440.

[25] Hall P, Müller H, Wang J (2006) Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics, 34(3):1493-1517.

[26] Hand DJ, Henley WE (1997) Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3):523-541.

[27] Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer Science & Business Media, USA.

[28] Tecnocom (2016) Tendencias en Medios de Pago 2016. `http://www.afi.es/afi/libre/pdfs/grupo/documentos/Informe_Tecnocom16_WEB.pdf` Accedido 2 de enero de 2017.

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

[30] Kavousian A, Rajagopal R, Fischer M (2012) A method to analyze large data sets of residential electricity consumption to inform data-driven energy efficiency. Centre for Integrated Facility Engineering (CIFE) Working Paper #WP130, Standford University. `http://cife.stanford.edu/publications` Accedido 1 de septiembre de 2016.

[31] Krämer N, Boulesteix A, Tutz G (2008) Penalized partial least squares with applications to b-spline transformations and functional data. Chemometrics and Intelligent Laboratory Systems, 94(1):60-69.

[32] Lardin-Puech P, Cardot H, Goga C (2014) Analysing large datasets of functional data: a survey sampling point of view. Journal de la Société Française de Statistique, 155(4):70-94.

[33] Laukaitis A, Rackauskas A (2002) Functional data analysis of payment systems. Nonlinear Analysis: Modeling and Control 7(2):53-68.

[34] Laukaitis A, Rackauskas A (2005) Functional data analysis for clients segmentation tasks. European journal of operational research 163:210-216.

[35] Laukaitis A (2008) Functional data analysis for cash flow and transactions intensity continuous-time prediction using Hilbert-valued autoregressive processes. European Journal of Operational Research 185:1607-1614.

[36] Linden G, Smith B and York J (2003) "Amazon.com recommendations: item-to-item collaborative filtering" in IEEE Internet Computing, 7(1):76-80.

[37] Lohr S (2012) The Age of Big Data. The New York Times. `http://nyti.ms/18LduZA` Accedido 1 de septiembre de 2016.

[38] Miao H (2013) Potential Applications of Function Data Analysis in High-frequency Financial Research. Business & Financial Affairs 2:e125.

[39] Müller H, Stadtmüller U (2005) Generalized functional linear models. The Annals of Statistics, 33(2):774-805.

[40] Müller H, Yao F (2008) Functional additive models. Journal of the American Statistical Association, 103(484):1534-1544.

[41] Peña D, Tiao GC, Tsay RS (2011) A Course in Time Series Analysis. John Wiley & Sons, New York, USA.

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

[43] Ramsay J, Dalzell C (1991) Some tools for functional data analysis. Journal of the Royal Statistical Society. Series B (Methodological) 53:539-572.

[44] Ramsay J, Silverman B (2002) Applied functional data analysis: methods and case studies. Springer, New York.

[45] Ramsay J, Silverman B (2005) Functional data analysis. Springer Science & Business Media, USA.