Trabajo Fin de Máster

Selección de modelos para estimación de áreas pequeñas. Aplicación a datos socioeconómicos de la Comunidad de Galicia

Manuel Antonio Novo Pérez

Máster en Técnicas Estadísticas

Curso 2017-2018

Propuesta de Trabajo Fin de Máster

Título en galego: Selección de modelos para estimación de áreas pequenas. Aplicación a datos socioeconómicos da Comunidade de Galicia
Título en español: Selección de modelos para estimación de áreas pequeñas. Aplicación a datos socioeconómicos de la Comunidad de Galicia
English title: Selection of models for estimation of small areas. Application to socio-economic data of the Community of Galicia
Modalidad: Modalidad B
Autor/a: Manuel Antonio Novo Pérez, USC
Director/a: María José Lombardía Cortiña, UDC; ,
Tutor/a: Esther López Vizcaíno, IGE; ,
Breve resumen del trabajo: Los modelos mixtos son muy flexibles y ampliamente utilizados en aplicaciones. Pero una parte clave del análisis es la selección de los modelos. El problema se centra en seleccionar las variables auxiliares y los efectos aleatorios, y varios procedimientos de selección han sido propuestos en la literatura. Este proyecto fin de máster se enfoca al estudio de la selección de modelos lineales mixtos con aplicaciones particulares en áreas pequeñas

Doña María José Lombardía Cortiña, de la UDCy doña Esther López Vizcaíno, responsable del Servicio de Difusión e Información del IGE, informan que el Trabajo Fin de Máster titulado

Selección de modelos para estimación de áreas pequeñas. Aplicación a datos socioeconómicos de la Comunidad de Galicia

fue realizado bajo su dirección por don Manuel Antonio Novo Pérez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 19 de Septiembre de 2018.

La directora:

Doña María José Lombardía Cortiña

La tutora:

Doña Esther López Vizcaíno

El autor/a:

Don Manuel Antonio Novo Pérez

Agradecimientos

A mi directora del TFM, María José Lombardía Cortiña por su paciencia, apoyo y su disponibilidad.

Al IGE por permitirme realizar las prácticas con ellos, especialmente a mi tutora, Esther López Vizcaíno por su paciencia, apoyo y disponibilidad.

A mi familia.

Índice general

Resumen

Resumen en español

El objetivo de este trabajo es la selección de modelos mixtos en el contexto de áreas pequeñas, para lo que se utilizarán varias versiones del criterio AIC adaptado aplicándolo a un ejemplo con datos reales proporcionados por el Instituto Galego de Estadística (IGE). En concreto, en la selección de modelos mixtos para explicar los distintos tipos de ingresos de los hogares gallegos.

De esta forma, en primer lugar se introducirán los conceptos y la metodología, mostrando en detalle los modelos mixtos y los criterios de selección que utilizaremos posteriormente. Tras esto, utilizando un conjunto de datos reales proporcionados por el IGE, aplicaremos dicha metodología para estimar los distintos tipos de ingresos medios en el hogar en el período 2007-2016, obteniendo estimadores más precisos que los obtenidos por los estimadores directos.

Finalmente aplicaremos los criterios AIC adaptados al contexto de áreas pequeñas para seleccionar cuál es el mejor modelo y cuales son sus variables explicativas, mejorando los modelos ajustados anteriormente. Por esto, este trabajo no solo nos permitirá obtener estimadores más precisos si no que nos permitirá comprobar el comportamiento de los distintos métodos de selección de modelos en áreas pequeñas, permitiendo que nos decantemos por alguno de ellos.

English abstract

The objective of this work is the selection of mixed models in the context of small areas, for which several versions of the AIC criterion will be adapted, applying it to an example with real data provided by the Instituto Galego de Estadística (IGE). Specifically, in the selection of mixed models to explain the different types of income of Galician households.

n this way, the concepts and methodology will be introduced first, showing in detail the mixed models and the selection criteria that we will use later. After this, using a set of real data provided by the IGE, we will apply this methodology to estimate the different types of average income in the household in the period 2007-2016, obtaining more precise estimators than those obtained by direct estimators.

Finally, we will apply the AIC criteria adapted to the context of small areas to select which is the best model and which are its explanatory variables, improving the previously adjusted models. Therefore, this work will not only allow us to obtain more precise estimators, but it will allow us to check the behavior of the different models selection methods in small areas, which will allow us to chose the best of them.

Capítulo 1

Introducción al problema y los datos.

En ocasiones cuando se realiza un estudio contamos con pocos datos de una población, como por ejemplo ocurre si queremos investigar alguna enfermedad rara a partir de las personas que la padecen o si estudiamos características socioeconómicas de una zona de Galicia a partir de una encuesta que tenga como objetivo toda la población española. En estos casos, debido a los pocos datos de los que disponemos, los estimadores directos ofrecen unos resultados pobres resultando en predicciones o estimaciones con una alta varianza, lo que provoca a su vez que, por ejemplo, las estimaciones de características de una población en años consecutivos muestren cambios bruscos, por lo que es necesario el uso de otra metodología. En este contexto destaca la metodología de estimación en áreas pequeñas (SAE), que logra reducir los problemas de los estimadores directos mediante el uso de herramientas como los modelos mixtos. Esta metodología ha sido estudiada en múltiples ocasiones por diversos autores, entre los que podemos destacar los trabajos de *Rao(2003)*, *Jiang y Lahiri (2006)* y *Rao (2003)*, en los que se realiza una revisión general de la metodología SAE y los avances realizados en la misma.

En el presente trabajo nos centraremos en estos últimos, concretamente en el problema la selección de los mejores modelos y sus variables, en el que debido a la naturaleza de dichos modelos, no podemos aplicar criterios ya conocidos como el AIC o el BIC. Este problema ha sido estudiado por diversos autores que utilizan diversas metodologías, destacando entre ellas los criterios AIC adaptados al contexto SAE que pueden verse en algunas referencias como *Vaida y Blanchard (2005)*, *Han (2013)* o más recientemente en *Lombardía et al. (2017)*, aunque no existe todavía una solución definitiva a este problema. Por ello, estudiaremos el comportamiento de algunos de los criterios más utilizados aplicándolos en un caso real, concretamente a los datos de la Encuesta estructural a hogares (EEH) proporcionados por el Instituto Galego de Estadística (IGE).

La EEH se realiza anualmente por el IGE desde 1999 y está dirigida a los hogares gallegos con el objetivo de obtener información sobre las características socieconómicas de los mismos, siendo sus principales objetivos:

- Analizar y describir las principales características de los hogares gallegos y la diferencia entre las distintas áreas territoriales gallegas.
- Proporcionar información sobre las características socioeconómicas de la población y los hogares gallegos.
- Conocer la cuantía de los ingresos de los hogares y su tipología, proporcionando información significativa de los ingresos monetarios menores.

- Conocer la cuantía de los gastos comunes de los hogares y su relación con otras variables socioeconómicas.

En este trabajo no trataremos todos estos aspectos, si no que nos centraremos en conocer y explicar los ingresos de los hogares según su tipología en las distintas áreas territoriales gallegas. Para ello utilizaremos la metodología SAE, diversas variables socioeconómicas y diversos criterios para seleccionar tanto los modelos como las variables, puesto que los estimadores directos no funcionan correctamente a nivel de área, pues muestran una gran variabilidad entre los distintos años.

De esta forma este trabajo está estructurado de manera que, en lo que resta de Capítulo 1, comentaremos que datos concretos estudiaremos y su procedencia. En el Capítulo 2 introduciremos la metodología que aplicaremos. En el Capítulo 3 realizaremos un análisis descriptivo de las variables descritas en este capítulo y seleccionaremos aquellas más adecuadas para los modelos explicados en el Capítulo 2, que ajustaremos en el Capítulo 4. En el Capítulo 5 seleccionaremos las variables y los modelos utilizando diversos criterios de selección y finalmente en el Capítulo 6 sacaremos conclusiones de los resultados obtenidos.

1.1. Encuesta estructural a hogares.

Como ya hemos mencionado, en la EEH se obtienen datos para obtener información sobre las características socioeconómicas. En concreto trabajaremos con los datos de dicha encuesta entre los años 2007 y 2016 (siendo, respectivamente, las ediciones de 2008 y 2017). En dichos años se fragmenta el territorio gallego en diversas áreas (estando estas preestablecidas por el IGE), las cuales mostramos a continuación según la provincia en la que se sitúan, las comarcas que las forman y su código numérico dentro de su misma provincia:

Provincia de A Coruña.

1. **A Coruña suroriental**: comprende las comarcas de Arzúa, Ordes y Terra de Melide.
2. **Ferrol-Eume-Ortegal**: comprende las comarcas de Ferrol, Eume y Ortegal.
3. **Área da Costa da morte**: comprende las comarcas de Bergantiños, Fisterra, Muros, Terra de Soneira y Xallas.
4. **A Barbanza-Noia**: comprende las comarcas de Barbanza y Noia
5. **Área da A Coruña**: Comprende las comarcas de A Coruña y Betanzos.
6. **Área de Santiago**: comprende las comarcas da Barcala, O Sar y Santiago.

Provincia de Lugo.

1. **Lugo sur**: comprende las comarcas de Chantana, Quiroga y Terra de Lemos.
2. **Lugo oriental**: comprende las comarcas da Fonsagrada, Os Ancares y Sarria.
3. **Lugo central**: comprende las comarcas da Ulloa, Lugo, Meira y A Terra Chá.
4. **A Mariña**: comprende las comarcas da Mariña Central, A Mariña Oriental y A Mariña Occidental.

Provincia de Ourense.

1. **O Carballiño-O Ribeiro**: comprende las comarcas de O Carballiño y O Ribeiro.
2. **Ourense central**: comprende las comarcas de Allariz e Maceda, Terra de Caldelas, Terra de Trives y Valdeorras.
3. **Ourense sur**: comprende las comarcas da Limia, Baixa Limia, Terra de Celanova, Verín y Viana.
4. **Área de Ourense**: comprende la comarca de Ourense.

Provincia de Pontevedra.

1. **Pontevedra nororiental**: comprende las comarcas de Deza y Tabeiros-Terra de Montes.
2. **Pontevedra sur**: comprende las comarcas da Paradanta, O Baixo Miño y O Condado.
3. **Caldas-O Salnés**: comprende las comarcas de Caldas y O Salnés.
4. **O Morrazo**: comprende la comarca de O Morrazo.
5. **Área de Pontevedra**: comprende la comarca de Pontevedra.
6. **Área de Vigo**: comprende la comarca de Vigo.

En la Figura 1.1 podemos ver también como se distribuyen las áreas en la geografía gallega. A su vez cada una de estas áreas se dividen en distintos estratos según el número de habitantes, considerando los estrados:

- **Estrato 0**: concellos autorepresentados.
- **Estrato 1**: concellos de más de 20000 habitantes.
- **Estrato 2**: concellos de 15000 a 20000 habitantes.
- **Estrato 3**: concellos de 10000 a 15000 habitantes.
- **Estrato 4**: concellos de 5000 a 10000 habitantes.
- **Estrato 5**: concellos de menos de 5000 habitantes.

Se unen estratos contiguos en algunas áreas para evitar estratos de escasa representatividad. Además de esta división, también fragmentamos el territorio en las distintas secciones censales. En las ediciones de 2008 y 2009 se investigaron 642 secciones, repartiéndose 231 para A Coruña, 112 para Lugo, 116 para Ourense y 183 para Pontevedra, entrevistando en cada una a 16 viviendas, dando un tamaño muestral total de 10272, mientras que de 2010 en adelante fueron 512 secciones, repartiéndose 180 para A Coruña, 90 para Lugo, 91 para Ourense y 151 para Pontevedra, entrevistando en cada una a 18 viviendas, dando un tamaño muestral total de 9216 viviendas.

Considerando las divisiones territoriales ya mencionadas se procede a seleccionar la muestra, cuyo marco es el Padrón de habitantes. En los concellos autorepresentados, el muestreo es unietápico, es decir, se ordenan los hogares según características sociodemográficas y a continuación se escoge la muestra mediante muestreo sistemático con arranque aleatorio. En los demás estratos de las distintas áreas el muestreo es bietápico, es decir, primero se ordenan las secciones según características sociodemográficas escogiéndose la muestra de secciones mediante muestreo sistemático con arranque aleatorio y luego en una segunda etapa, para la muestra de secciones escogida en la primera etapa, se ordenan los hogares dentro de esas secciones siguiendo los mismos criterios, escogiendo entonces la muestra de

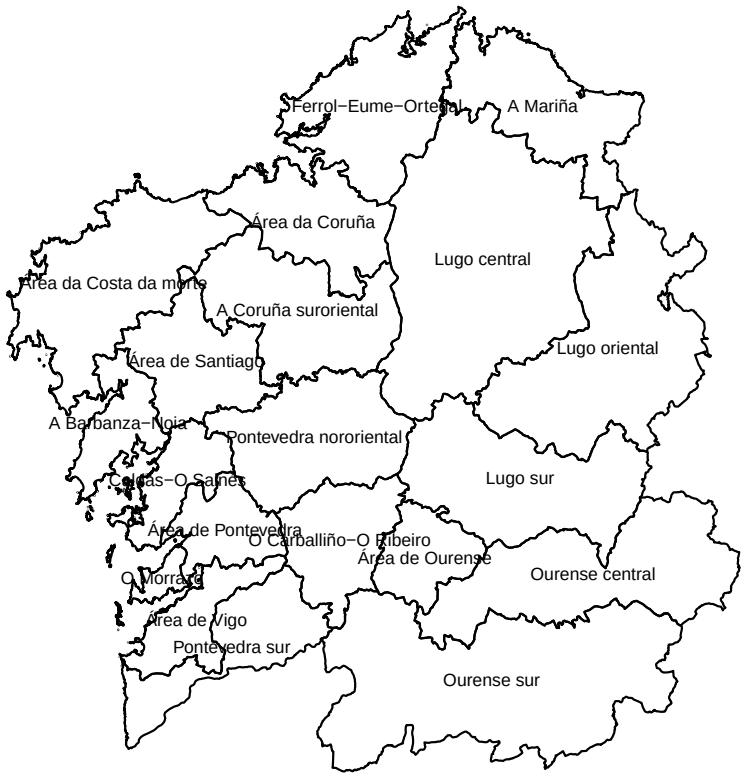

Figura 1.1: Mapa de Galicia separado en las distintas áreas de estudio

hogares mediante muestreo sistemático con arranque aleatorio.

Como ya mencionamos anteriormente, nuestro interés se centra en estudiar los ingresos medios de los hogares en la áreas comentadas anteriormente (calculados mediante los estimadores que veremos en el Capítulo 2), según su tipo, los cuales son estimados en cada área mediante estimadores directos (que veremos en el Capítulo 2). De esta forma consideramos los siguientes tipos de ingreso:

- **Ingresos totales medios**. Se trata de la media de los ingresos totales de cada uno de los hogares en cada área.
- **Ingresos por cuenta ajena medios**. Se trata de la media de los ingresos por cuenta ajena de cada uno de los hogares que poseen ingresos de este tipo en cada área.
- **Ingresos por cuenta propia medios**. Se trata de la media de los ingresos por cuenta propia de cada uno de los hogares que poseen ingresos de este tipo en cada área.
- **Ingresos por prestaciones contributivas medios**. Se trata de la media de los ingresos por cuenta ajena de cada uno de los hogares que poseen ingresos de este tipo en cada área.

Capítulo 2

Introducción a la metodología.

Como ya mencionamos en el capítulo anterior, la estimación de los distintos tipos de ingresos medios en cada área se calcula utilizando estimadores directos. Esto puede ocasionar que las estimaciones de un año para otro experimenten cambios bruscos que no deberían suceder, esto es debido a que en cada una de las áreas el tamaño muestral es reducido, provocando inestabilidad en los estimadores y altos errores de estimación. Por ello una forma razonable de lidiar con este problema es el uso de la metodología propia de áreas pequeñas, en la que se utilizan los datos de todas las áreas para conseguir estimadores en cada una de ellas.

En este capítulo definiremos qué entendemos por áreas pequeñas, comentando cuales son los principales modelos que se utilizan en este contexto, centrándonos en los llamados modelos mixtos (ya sea con o sin efecto temporal) y como podemos seleccionar cuál de ellos es el mejor, centrándonos en este caso en distintas versiones del AIC adaptados a este contexto.

2.1. Modelos lineales mixtos en SAE.

El término *área pequeña* se refiere a una población en la que no tenemos estadísticos fiables debido a las limitaciones de los datos disponibles. Ejemplos de esto pueden ser regiones geográficas, como un municipio o incluso grupos demográficos, como puede ser un grupo de individuos de una determinada edad o raza, de los que disponemos de pocos datos, lo que provoca que estimadores como la media muestral se comporten de forma inestable o, cómo ocurre con los modelos de regresión habituales, las estimaciones sean deficientes. En la literatura pueden verse distintos métodos de estimación que tratan de solventar esto, como es el caso de, entre los que destacan los basados en el diseño, en modelos asistidos y basados en el modelo, que serán en las que nos centraremos.

Los **modelos lineales mixto** son modelos que además de incorporar efectos fijos (tal y como hacen los modelos de regresión lineal habituales), incorporan efectos aleatorios. Son por ello adecuados en el contexto de áreas pequeñas por su flexibilidad a la hora de combinar información de diferentes fuentes y explicar distintos tipos de errores. En el caso de áreas pequeñas es habitual que estos incorporen efectos de área específicos para explicar las variaciones no explicadas por los efectos fijos del modelo, diferenciándose así de los modelos de regresión más habituales, que no contemplan diferencias entre las distintas áreas o las incluyen como efectos fijos, añadiendo un gran número de parámetros.

En el contexto de áreas pequeñas, los modelos mixtos suelen clasificarse según el tipo de información que dispongamos para la variable respuesta en dos tipos: **modelos mixtos a nivel de área** y los **modelos mixtos a nivel de individuo**, de los que hablaremos a continuación.

2.1.1. Modelos mixtos a nivel de área.

Esta clase de modelos son aquellos en los que las variables explicativas están disponibles solamente a nivel de área. Dentro de estos modelos destaca el **modelo de Fay-Herriot** (Fay y Herriot (1979)), que fue utilizado inicialmente para estimar la renta per cápita en pequeñas localidades que veremos en detalle a continuación.

Modelo de Fay-Herriot.

El modelo de Fay-Herriot para D áreas es un modelo Bayesiano de dos niveles:

- **Nivel 1(modelo muestral)**: $y_d|\mu_d \in N(\mu_d, \sigma^2_{\epsilon_d})$, $d = 1, ..., D$
  Donde y_d es el estimador de la media de la variable de interés en el área d y μ_d es la media de dicha variable en el área. Este nivel se utiliza para explicar la variabilidad debida al muestreo, $\sigma^2_{\epsilon_d}$ de y_d en cada área, la cual se considera conocida.
- **Nivel 2(modelo de conexión)**: $\mu_d \in N(\mathbf{x}'_d\boldsymbol{\beta}, \sigma^2_u)$, $d = 1, ..., D$
  donde $\mathbf{x}'_d = (x_{d1}, ..., x_{dp})$ es el vector constituido por las p covariables, $\boldsymbol{\beta}$ el vector de parámetros y σ^2_u la varianza de μ_d, siendo $\boldsymbol{\beta}$ y σ^2_u habitualmente desconocidos.

Si consideramos los dos niveles anteriores podemos escribir el modelo como un modelo lineal mixto de la siguiente forma:

$$y_d = \mu_d + \epsilon_d = \mathbf{x}'_d\boldsymbol{\beta} + u_d + \epsilon_d \tag{2.1}$$

donde para $d = 1, ..., D$, $\epsilon_d \in N(0, \sigma^2_{\epsilon_d})$ y $u_d \in N(0, \sigma^2_u)$ son independientes, suponiendo $\sigma^2_{\epsilon_d}$ conocida y siendo σ^2_u desconocida. Con este modelo *Fay y Herriot (1979)* capturaron los efectos específicos de cada área utilizando efectos aleatorios, algo que no es posible utilizando modelos de regresión clásicos.

2.1.2. Modelo a nivel de unidad.

En ocasiones para cada área d disponemos de una variable respuesta $y_{d,i}$ y de p variables auxiliares $x_{k,d,i}$, $k = 1, ..., p$, para cada individuo i. Es decir, disponemos de información de individuos concretos en cada área. Un ejemplo de esta clase de modelos es el siguiente:

$$y_{d,i} = u_d + \beta_1 x_{1,d,i} + \beta_2 x_{2,d,i} + ... + \beta_{p-1} x_{(p-1),d,i} + \beta_p x_{p,d,i} + e_{d,i} \tag{2.2}$$

donde u_d es el efecto del área d y $e_{d,i}$ son los errores, ambos gaussianos i.i.d con media 0 y con varianzas σ^2_u y σ^2_e respectivamente, siendo ambas desconocidas.

Un ejemplo de aplicación del modelo (2.2) es el realizado por *Battese et al. (1988)*, buscando relacionar el número de hectáreas de maíz en los segmentos de los condados, a partir del número de *pixels* clasificados como maíz y soja en los mismos.

En el ejemplo anterior sucede que los tamaños muestrales para cada condado son muy pequeños, lo que provoca medias muestrales con grandes errores. También hay dificultades en la estimación de la varianza por métodos muestrales estándar, pues en algunos condados no es posible hacerlo, debido a que solo tenemos un dato. A diferencia de los modelos de regresión habituales, esos modelos si reflejan un efecto aleatorio debido al área u_d, lo que les permite capturar variabilidad adicional específica del área.

Como se comenta en *Herrador et al. (2009)*, los modelos de unidad suelen tener una alta precisión de estar bien especificados, aunque los estimadores derivados de estos modelos necesitan combinar datos auxiliares de las unidades con los datos agregados de los dominios, algo no siempre posible. Otro problema de estos modelos es la dificultad de obtener las variables a nivel de unidad. En nuestro caso nos centraremos en los modelos de área, puesto que nuestros datos se encuentran a dicho nivel.

2.1.3. Modelos lineales mixtos.

Los modelos lineales (2.1) y (2.2) son casos particulares del llamado **modelo lineal mixto**. Dicho modelo engloba los modelos que se pueden expresar de la forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{2.3}$$

donde $\mathbf{y}$ es el vector de las observaciones, $\mathbf{X}$ la matriz de covariables conocida, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión desconocidos, a menudo denominados efectos fijos, $\mathbf{Z}$ es una matriz conocida, $\mathbf{u}$ es un vector de efectos aleatorios y $\boldsymbol{\varepsilon}$ es el vector de errores. Nótese que tanto $\mathbf{u}$ como $\boldsymbol{\varepsilon}$ son no observables. Es interesante comparar el modelo lineal mixto con el modelo lineal general, expresado como $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, pues la diferencia entre ellos es el término $\mathbf{Zu}$, que refleja los efectos aleatorios del modelo, que puede tomar una gran variedad de formas, permitiéndonos obtener una amplia cantidad de modelos y establecer estructuras de dependecia entre las distintas áreas.

Como hemos mencionado anteriormente $\mathbf{u}$ y $\boldsymbol{\varepsilon}$ son no observables, por lo que tendremos que hacer ciertas suposiciones, que serán las hipótesis de nuestro modelo. Habitualmente se considera que ambos son incorrelados con media cero y varianza finita. Adicionalmente, tal y como se comenta en *Jiang y Lahiri (2006)*, puede asumirse normalidad para facilitar los cálculos, denominándose el modelo resultante como **modelo gaussiano**.

2.1.4. Estimación de los parámetros.

Uno de los principales objetivos es la realización de predicciones de la media real μ_d en cada área, es decir $\boldsymbol{\mu_d} = \mathbf{X_d}\boldsymbol{\beta} + \mathbf{Z_d}\mathbf{u}$. Observamos que para obtener dichas predicciones requerimos estimar $\boldsymbol{\beta}$ y predecir $\mathbf{u}$. El problema mencionado depende de si las varianzas de $\boldsymbol{\varepsilon}$ y $\mathbf{u}$ son conocidas o no, puesto que, como veremos, los estimadores de $\mathbf{u}$ y de $\boldsymbol{\beta}$ dependen de dichas varianzas, por lo que tenemos dos escenarios diferenciados:

- **Escenario 1: componentes de la varianza conocidas**. En este caso tendremos que estimar $\boldsymbol{\beta}$ y $\mathbf{u}$.
- **Escenario 2: componentes de la varianza desconocidas**. En este escenario, además de $\boldsymbol{\beta}$ y $\mathbf{u}$, deberemos estimar las varianzas de $\mathbf{u}$ y de los errores $\boldsymbol{\varepsilon}$.

Escenario 1: Componentes de la varianza conocidas.

En este escenario tanto $\mathbf{V}_u = Var(\mathbf{u})$ como $\mathbf{V}_\varepsilon = Var(\boldsymbol{\varepsilon})$ son conocidas. En *Jiang y Lahiri (2006)* se sugiere predecir $\mathbf{u}$ de la forma:

$$\tilde{\mathbf{u}} = \mathbf{V_u}\mathbf{Z}^{-1}\mathbf{V}_u^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.4}$$

donde $\mathbf{V}_y = Var(\mathbf{y}) = \mathbf{V}_\varepsilon + \mathbf{V}_u$. Observamos que $\tilde{\mathbf{u}}$ depende de $\boldsymbol{\beta}$, el cuál es desconocido. En *Jiang y Lahiri (2006)* se comenta que suele reemplazarse por su estimador de máxima verosimilitud (MLE), suponiendo normalidad en $\boldsymbol{\beta}$ y que V_u y V_ε son conocidos, siendo dicho estimador

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V_y^{-1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V_y^{-1}}\mathbf{y}.$$

El predictor resultante de sustituir el estimador $\tilde{\boldsymbol{\beta}}$ en $\tilde{\boldsymbol{u}}$ es el llamado predictor BLUP (mejor predictor linear insesgado).

Escenario 2: Componentes de la varianza desconocidas.

En el Escenario 1 hemos visto estimadores tanto de $\boldsymbol{\beta}$ como de $\mathbf{u}$. Observamos que dichos estimadores dependen ya sea de $\mathbf{V}_u$ o de $\mathbf{V}_y$, lo que significa que dependen de las componentes de la varianza. Esto supone un problema, pues significa que en el caso de ser desconocidas, no podemos obtener los estimadores. En *Jiang y Lahiri (2006)* se sugiere reemplazar dichas varianzas por estimadores consistentes de las mismas, $\hat{\mathbf{V}}_u$ y $\hat{\mathbf{V}}_y$. Al realizar dicha sustitución obtenemos el llamado BLUP empírico (EBLUP).

Obtener los estimadores $\hat{\mathbf{V}}_u$ y $\hat{\mathbf{V}}_y$ se trata de un problema en sí, pues depende de la distribución de las componentes de la varianza. En *Jiang y Lahiri (2006)* se menciona que en el caso general se pueden utilizar como métodos el ANOVA (análisis de la varianza) o el MINQUE (*Rao 1972*), aunque ambos métodos tienen dificultades, pues el ANOVA genera estimadores ineficientes para *unbalanced data* y el MINQUE depende de otros factores como los valores iniciales.

El problema se simplifica si suponemos normalidad, puesto que, como se señala en *Jiang y Lahiri (2006)*, los estimadores eficientes son los de máxima verosimilitud (MLE), aunque muestra problemas referentes a sus propiedades asintóticas, pues a diferencia del caso *i.i.d* , en este las observaciones están correlacionadas bajo el modelo lineal mixto. Además los estimadores MLE son, en general, sesgados, ocurriendo que dicho sesgo no se anula al aumentar el tamaño muestral si el número de efectos fijos es proporcional al tamaño muestral. Para lidiar con el problema anterior en *Jiang y Lahiri (2006)* se propone el uso del llamado método de máxima verosimilitud residual o restringido (REML), propuesto por *Thompson (1962)* y extendido por *Patterson y Thompson (1971).*

Si bien el MLE y el REML están desarrollados bajo normalidad, en *Jiang y Lahiri (2006)* se menciona su uso para situaciones no normales utilizando una aproximación *quasi-likehood.* Los autores hablan de la consistencia asintótica y la normalidad del REML, así como también condiciones necesarias y suficientes para propiedades similares para el MLE.

2.2. Modelos lineales mixtos con efectos temporales en SAE.

En este trabajo disponemos de datos recogidos a lo largo del tiempo, por lo que puede resultar de interés considerar el efecto temporal de dichos datos. Para lo anterior *Rao y Yu (1994)* sugirieron un modelo basado en el modelo de Fay-Herriot, añadiendo un efecto temporal en el mismo. El modelo para un área $d \in (1, ..., D)$ y el instante $t \in (1, ..., T_d)$ es el siguiente:

$$y_{dt} = \boldsymbol{x}_{dt}^{'}\boldsymbol{\beta} + u_{1,d} + u_{2,dt} + \epsilon_{dt} \tag{2.5}$$

donde $y_{d,t}$ la variable respuesta en el área d para el instante t, $\boldsymbol{x}_{d,t}$ el vector de variables explicativas, $u_{1,d} \in N(0, \sigma_1^2)$ es el efecto aleatorio del área d, $(u_{2,d1}, ..., u_{2,dT_d})$ es i.i.d. AR(1) con misma varianza σ_2^2 y autocorrelación ρ para cada d y ϵ_{dt} son los errores, que son independientes, normales de media 0 y varianza $\sigma_{d,t}^2$ (siendo esta conocida, a diferencia de σ_1^2, σ_2^2 y ρ, que son desconocidas), siendo además ϵ_{dt}, $(u_{2,d1}, ..., u_{2,dT_d})$ y $u_{1,d}$ independientes. Observamos que el efecto aleatorio $u_{2,dt}$ es el encargado de capturar el efecto temporal en los datos. En *Rao y Yu (1994)* se comenta que se podría considerar que los efectos siguen un proceso ARMA, pero que dichos modelos no muestran ser significativamente mejores.

Observamos que el modelo (2.5) puede expresarse, al igual que los modelos anteriores, de forma matricial. En este caso:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{2.6}$$

donde $\mathbf{y} = col_{1\leq d\leq D}(\mathbf{y}_d)$, $\mathbf{y}_d = col_{1\leq t\leq T_d}(y_{dt})$, $\mathbf{u}_1 = col_{1\leq d\leq D}(u_{1,d})$, $\mathbf{u}_2 = col_{1\leq d\leq D}(\mathbf{u}_{2,d})$, $\mathbf{u}_{2,d} = col_{1\leq t\leq T_d}(u_{2,dt})$, $\mathbf{u} = (\mathbf{u}_1^{'}, \mathbf{u}_2^{'})^{'}$ $\boldsymbol{\varepsilon} = col_{1\leq d\leq D}(\boldsymbol{\epsilon}_d)$, $\boldsymbol{\epsilon}_d = col_{1\leq t\leq T_d}(\epsilon_{dt})$, $\mathbf{X} = col_{1\leq d\leq D}(\mathbf{X}_d)$, $\mathbf{X}_d = col_{1\leq t\leq T_d}(x_{dt})$, siendo x_{dt} los vectores con las p covariables, $\boldsymbol{\beta}$ el vector de efectos fijos, $\mathbf{Z}_1 = diag_{1\leq d\leq D}(1_{T_d})$, $\mathbf{Z}_2 = I_M$, denotando $\mathbf{I}_M$ la matriz identidad de dimensión M, con $M = \sum_{d=1}^{D} T_d$. Asumimos también que $\mathbf{u}_1 \sim N(0, \mathbf{V}_{\mathbf{u}_1})$, $\mathbf{u}_2 \sim N(0, \mathbf{V}_{\mathbf{u}_2})$ y $\boldsymbol{\varepsilon} \sim N(0, \mathbf{V}_{\boldsymbol{\varepsilon}})$ independientes con varianzas σ_{dt}^2 conocidas y matriz de covarianzas $\mathbf{V}_{\mathbf{u}} = diag(V_{\mathbf{u}_1}, V_{\mathbf{u}_2})$, $\mathbf{V}_{\mathbf{u}_1} = \sigma_1^2$, $\mathbf{V}_{\mathbf{u}_2} = \sigma_2^2\Omega(\rho)$, $\Omega(\rho) = diag_{1\leq d\leq D}(\Omega_d(\rho))$, $\mathbf{V}_{\boldsymbol{\varepsilon}} = diag_{1\leq d\leq D}(V_{\epsilon,d})$, $V_{\epsilon,d} = diag_{1\leq t\leq T_d}$ y

$$\Omega_d(\rho) = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \dots & \rho^{T_d-2} & \rho^{T_d-1} \\ \rho & 1 & \ddots & & \rho^{T_d-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{T_d-2} & & \ddots & 1 & \rho \\ \rho^{T_d-1} & \rho^{T_d-2} & \dots & \rho & 1 \end{pmatrix}$$

En la ecuación (2.6), observamos que el modelo planteado es un caso particular del modelo (2.3). Por lo tanto la estimación de los parámetros del modelo se calcula de igual forma, es decir $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ y $\tilde{\mathbf{u}} = \mathbf{V}_u Z^{'}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ son los estimadores BLUP cuando $\mathbf{V} = Var(\mathbf{y})$ es conocido , es decir, tal y como se comenta en *Rao y Yu (1994)*, cuando σ_2^2, σ_1^2 y ρ son conocidos.

En la práctica σ_2^2, σ_1^2 y ρ son desconocidos, por lo que deberemos estimarlos. Para ello en *Esteban et al. (2012)* se sugiere el uso del REML, utilizando el algoritmo Fisher scoring para calcularlo (pueden verse los detalles en la referencia). De esta forma, al tratarse de un caso particular del modelo (2.3), podemos estimar $\boldsymbol{\beta}$ y predecir $\mathbf{u}$ de igual manera que en dicho modelo.

2.3. Estimadores mixtos.

Uno de nuestros objetivos es, a partir de la muestra, conocer la media de la característica de interés en toda la población o las distintas áreas. Para ello se suelen estimar dichos valores utilizando los llamados estimadores directos, que utilizan solamente los valores de interés en cada dominio para la estimación. Tal y cómo se comenta en *Rao (2003)*, estos estimadores funcionan de forma razonable cuando la muestra en cada dominio es grande, algo que no se cumple en el contexto de áreas pequeñas, provocando grandes errores en los estimadores, por lo que son necesarios otros estimadores que nos permitan lidiar con este problema.

Existen diversas clases de estimadores que tratan de resolver el problema anterior como pueden ser los estimadores sintéticos pero nosotros destacaremos los estimadores mixtos, que, a diferencia de los estimadores directos, tienen en cuenta el efecto aleatorio del área correspondiente, puesto que se basan en los modelos mixtos que describimos en las anteriores secciones de este capítulo. Para más información acerca de estos, puede consultarse *Pfeffermann (2013)*, *Rao (2003)* o *Herrador el al. (2009)*, en los que se pueden ver otros tipos de estimadores habituales en el contexto SAE.

Estimadores basados en modelos de Fay-Herriot.

En este trabajo nos centraremos en los modelos mixtos de Fay-Herriot, por lo que nos centraremos en los estimadores que podemos obtener aplicando dichos modelos para el total y la media de la variable

de interés. Asumimos que:

$$y_d = \mu_d + \epsilon_d$$

donde y_d denota el estimador de la media de la población del área d y $\epsilon_d \sim N(0, \sigma_d^2)$, con σ_d^2 conocido. Luego suponemos que μ_d está linealmente relacionado con las variables auxiliares, es decir:

$$\mu_d = \mathbf{x'}_d\boldsymbol{\beta} + u_d$$

con $u_d \sim N(0, \sigma_u^2)$ y ϵ_d independientes, siendo σ_u^2 desconocido. Entonces, tal y como se comenta en *Herrador et al. (2009)*, ajustando el modelo por máxima verosimilitud, obtenemos el estimador EBLUP de la media en el área d basado en el modelo de Fay-Herriot, siendo este

$$\hat{\mu}_d = \mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{u}_d.$$

donde $\hat{\mu}_d$ es el estimador EBLUP de la media.

2.3.1. Estimadores basados en modelos de Fay-Herriot con efecto temporal.

El procedimiento que hemos visto para los modelos de Fay-Herriot puede aplicarse a otras clases de modelos mixtos, como por ejemplo el modelo con efecto temporal que hemos visto en la sección anterior. En este caso asumimos que:

$$y_{dt} = \mu_{dt} + \epsilon_{dt}$$

siendo y_{dt} el estimador de la media poblacional en el área d y el instante t, $\epsilon_{dt} \sim N(0, \sigma_{\epsilon,dt}^2)$, con $\sigma_{\epsilon,dt}^2$ conocida. A continuación suponemos que se sigue el modelo con efecto temporal, es decir:

$$\mu_{dt} = \mathbf{x'}_{dt}\boldsymbol{\beta} + u_{1,d} + u_{2,dt}$$

donde $u_{1,d} \sim N(0, \sigma_{u1}^2)$ y $(u_{2,d1}, ..., u_{2,dT_d})$ es i.i.d. AR(1) con misma varianza σ_2^2, siendo σ_1^2 y σ_2^2 desconocidas. Entonces, al igual que hacíamos para los modelos de Fay-Herriot, ajustando por máxima verosimilitud obtenemos el estimador EBLUP de la media, que denotaremos por $\hat{\mu}_{dt}$ y es de la forma:

$$\hat{\mu}_{dt} = \mathbf{x'}_{dt}\hat{\boldsymbol{\beta}} + \hat{u}_{1,d} + \hat{u}_{2,dt}$$

2.3.2. Error cuadrático medio de los estimadores mixtos.

El cálculo del error cuadrático medio (MSE) de un estimador es algo vital, puesto que nos indica la precisión de las estimaciones que hacemos con el mismo. En este trabajo nos centraremos en los modelos mixtos de Fay-Herriot, en concreto aquellos con efecto temporal, por lo que mostraremos a continuación como calcular el MSE en estos modelos.

Estimador de Prasad y Rao.

Un estimador del MSE muy utilizado en este ámbito es el estimador de Prasad y Rao, que se trata de una aproximación de segundo orden del MSE del modelo de Fay-Herriot vista por primera vez en *Prasad y Rao (1990)*.

Tal y como se dice en dicha referencia, la aproximación de segundo orden del MSE del modelo de Fay-Herriot para el área i puede estimarse por:

$$\widehat{MSE} = g_{1i} + g_{2i} + 2g_{3i}$$

donde g_1, g_2 y g_3 son funciones, cuyo cálculo detallado puede verse en *Prasad y Rao (1990)*, de la forma:

$$g_{1i}(\sigma_u^2) = \frac{\sigma_u^2\sigma_{\epsilon_i}^2}{\sigma_u^2 + \sigma_{\epsilon_i}^2}$$

$$g_{2i}(\sigma_u^2) = \sigma_{\epsilon_i}^4(\sigma_u^2 + \sigma_{\epsilon_i}^2)^2 x_i'(X'V^{-1}X)^{-1}x_i$$

$$g_{3i} = \sigma_{\epsilon_i}^4(\sigma_u^2 + \sigma_{\epsilon_i}^2)^3 var(\hat{\sigma}_u^2)$$

Estimación por remuestreo.

Además del estimador de Prasad y Rao, es posible estimar el MSE utilizando técnicas de remuestreo. Un ejemplo de estos métodos es el bootstrap paramétrico, que puede verse aplicado a multitud de modelos en algunos artículos, como en *Maruhenda et al. (2013)*, donde se aplica a modelos de Fay-Herriot con efecto temporal y espacial (considerando que estos no son independientes entre ellos). En este trabajo nos centraremos en los modelos de Fay-Herriot con efecto temporal y efectos de área independientes, que no es el mismo que el mostrado en la referencia, aunque puede adaptarse de forma análoga. El procedimiento entonces es el siguiente:

1. Estimamos los parámetros del modelo.
2. Repetimos B veces ($b = 1, ..., B$) los siguientes pasos:
   (a) Generamos la parte aleatoria del modelo $u^*_{1,d} \sim N(0, \hat{\sigma}^2_1)$, $u^*_{2,dt}$ como un $AR(1)$ para cada $d = 1, ..., D$ y $\epsilon^*_{dt} \sim N(0, \sigma^2_{d,t})$ independientes para cada $d = 1, ..., D$, con $\sigma^2_{d,t}$ conocido, construyendo el modelo bootstrap $y^{*(b)}_{dt} = \mu^{*(b)}_{dt} + \varepsilon^{*(b)}_{dt}$, con $\mu^{*(b)}_{dt} = \mathbf{x'}_{dt}\hat{\boldsymbol{\beta}}^{(b)} + u^{*(b)}_{1,d} + u^{*(b)}_{2,dt}$.
   (b) Para cada muestra bootstrap $\{y^{*(b)}_{dt}, \mathbf{x}_{dt}\}$ calculamos $\hat{\mu}^{*(b)}_{dt} = \mathbf{x'}_{dt}\hat{\boldsymbol{\beta}}^{(b)} + \hat{u}^{*(b)}_{1,d} + \hat{u}^{*(b)}_{2,dt}$.
3. Obtenemos el estimador del MSE:

$$MSE(\hat{\mu}_{dt}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}^{(b)}_{dt} - \mu^{(b)}_{dt})^2.$$

2.4. Métodos de selección en SAE.

En las Secciones 2.1 y 2.2 hemos propuesto varios modelos que pueden utilizarse para analizar nuestros datos, pero no todos ellos de la misma manera, de forma que algunos modelos se ajusten mejor a nuestros datos que otros modelos, por lo que necesitaremos criterios para saber que modelo es más adecuado.

Uno de los criterios más utilizados es el *Akaike information criterion*(AIC) (*Akaike (1973)*). Este criterio se basa en la distancia de Kullback-Leibler entre la verdadera densidad f generadora de la variable y y su aproximación dada por el modelo ajustado, M, definida como

$$I(f, M) = E_f \log f(y) - E_f \log M(y)$$

donde $I(f, M)$ denota la distancia de Kullback-Leibler entre la densidad f y su aproximación dada por el modelo M, E_f representa la esperanza respecto de la densidad f, de forma que una aproximación es mejor cuanto menor es la distancia. En la práctica los parámetros del modelo M son estimados utilizando los datos y (que denotaremos por $\hat{M}$). Teniendo en cuenta lo anterior, la calidad de la aproximación puede medirse utilizando:

$$E_f I(f, \hat{M}) = E_{f(y^*)} \log f(y^*) - E_{f(y)} E_{f(y^*)} \log \hat{M}(y^*)$$

donde y^* son obsevaciones independientes de y. En *Vaida y Blanchard (2005)* se comenta que al comparar distintas clases de modelos el termino $E_{f(y^*)} \log f(y^*)$ puede ser ignorado, por lo tanto podemos utilizar como criterio el AI (Akaike Information), que es de la forma:

$$AI = -2E_{f(y)} E_{f(y^*)} \log \hat{M}(y^*)$$

Obviamente la distribución f tampoco es conocida, por lo que deberemos estimar el AI. En *Vaida y Blanchard (2005)* se comenta que el AIC es un estimador del AI, y viene dado por:

$$\text{AIC} = -2 \log \hat{M}(y) + 2K$$

donde K es el número de parámetros libres del modelo. El criterio AIC nos permite comparar multitud de modelos clásicos como pueden ser los modelos de regresión más habituales pero no puede ser utilizado directamente para los modelos lineales mixtos. Esto último es debido a que, tal y como se comenta en la referencia, la propia definición del AIC no es sencilla cuando tenemos presente efectos aleatorios, puesto que no sabemos que verosimilitud utilizar ni tampoco si los efectos aleatorios son parámetros. A lo largo de esta sección veremos distintas formas de lidiar con estos problemas, de forma que obtendremos criterios similares al AIC aplicados en el contexto de áreas pequeñas.

2.4.1. AIC condicional y marginal.

Para solventar los problemas mencionados anteriormente, en *Vaida y Blanchard (2005)* se sugiere utilizar variantes del AIC según los intereses que tengamos. La idea es que en un modelo mixto el interés puede recaer en la población global o en un área en particular.

Consideremos el modelo general

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

donde $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_y)$ siendo $\mathbf{V}_y = var(\mathbf{y})$, $\mathbf{y}|\mathbf{Z}\mathbf{u} \sim N(\boldsymbol{\mu}, \mathbf{V}_{y|u})$, siendo $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}+\mathbf{u}$, $\mathbf{V}_{y|u} = var(\mathbf{y}|\mathbf{u})$ y $\mathbf{Z}$ es una matriz conocida.

AIC marginal.

Cuando nos centramos en toda la población, podemos considerar los efectos aleatorios $\mathbf{u}$ como una herramienta para modelar la correlación dentro del área d, por lo que el modelo lineal mixto (que consideraremos gaussiano por simplicidad) se trataría del siguiente modelo lineal:

$$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{Z}\boldsymbol{V_u}\boldsymbol{Z'} + \boldsymbol{V}_\varepsilon). \tag{2.7}$$

donde $\mathbf{V}_u = var(\boldsymbol{u})$ y $\mathbf{V}_\varepsilon = var(\boldsymbol{\varepsilon})$.

Debido a este planteamiento,como se comenta en *Vaida y Blanchard (2005)*, para el modelo lineal mixto se utiliza el llamado AIC marginal,

$$\text{mAIC} = -2\log(l_m(M)) + 2K$$

donde $\log(l_m(M))$ denota la log-verosimilitud marginal para un modelo mixto M como el mostrado en la Ecuación (2.7), siendo:

$$\log(l_m(M)) = \frac{1}{2}D\log(2\pi) - \frac{1}{2}|\mathbf{V}_y| - \frac{1}{2}(y - \boldsymbol{X}\boldsymbol{\beta})'\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})$$

donde K el número de parámetros a estimar en el modelo.

AIC condicional.

Como ya mencionábamos anteriormente, nuestro interés puede no centrarse en el global de la población, si no en cada una de las áreas. En este caso estamos interesados en las predicciones en cada una de las áreas y los u_d actúan como parámetros. Por ello la log-verosimilitud de interés es la condicional, que denotaremos por $\log(l_c(M))$ para un modelo M. Considerando esto *Vaida y Blanchard (2005)* sugieren entonces el AIC condicionado (cAIC) que consiste en :

$$\text{cAIC} = -2\log(l_c(M)) + 2K$$

donde, si consideramos un modelo mixto como el de la Ecuación (2.7) ,

$$\log(l_c(M)) = \frac{1}{2}D\log(2\pi) - \frac{1}{2}|\mathbf{V}_{y|u}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}_{y|u}^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

es la verosimiliud condicionada a $\mathbf{u}$, K son los grados de libertad, $\mathbf{V}_{y|u}$ es la matriz de covarianzas de $\mathbf{y}$, cuando está condicionado a los efectos aleaorios $\mathbf{u}$, es decir cuando es condicionado a las distintas áreas y $E[\mathbf{y}|\mathbf{u}] = \boldsymbol{\mu} = \boldsymbol{X\beta} + \mathbf{Zu}$.

Uno de los principales problemas del cAIC es determinar el parámetro K, teniendo varios valores distintos para el REML y el MLE, incluso variando entre los distintos autores. Por ejemplo *Vaida y Blanchard (2005)* sugiere para el MLE

$$K_{MLE} = \frac{N(N-p-1)}{(N-p)(N-p-2)}(\rho+1) + \frac{N(p+1)}{(N-p)(N-p-2)}$$

y para el REML

$$K_{REML} = \frac{N-p-1}{N-p-2}(\rho+1) + \frac{p+1}{N-p-2}$$

donde N es el total de observaciones, p es el número de efectos aleatorios $\rho = tr(\mathbf{H})$, siendo $\mathbf{H}$ la matriz *hat* de forma que $\hat{\mathbf{y}} = \mathbf{Hy}$, con $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}$.

Los estimadores anteriores tienen ciertos problemas, puesto que realizan ciertas suposiciones sobre el conocimiento de las matrices de covarianzas de los efectos aleatorios y de los errores (ver *Vaida y Blanchard (2005)* para más detalles). Por ello otros autores sugieren alternativas evitando dichas suposiciones, por ejemplo en *Liang et al.(2008)* se propone

$$K = \sum_{i=1}^{N} \frac{\partial \hat{y}_i}{\partial y_i}$$

siendo $\hat{y}_i$ e y_i la observación i esperada por el modelo y la observación i respectivamente. Cabe destacar que cada término $\frac{\partial \hat{y}_i}{\partial y_i}$ puede ser calculado directamente o numéricamente utilizando $\frac{\hat{y}_i(y+he_i)-\hat{y}_i}{h}$ donde h es un número pequeño y e_i es el vector de dimensión N (siendo este el número de áreas) con un 1 en la posición i y ceros en las demás. Cabe destacar que en este caso el método de estimación de los parámetros no está especificado. En *Han (2013)* se muestran aproximaciones considerando diferentes estimadores, entre los que destacaremos el caso del MLE y el REML, siendo entonces de la forma:

$$K = \rho - 2\mathbf{A}^{-1}r_R\hat{\mathbf{V}}_y^{-1}\hat{\mathbf{P}}^*\hat{\mathbf{V}}_\varepsilon\hat{\mathbf{V}}_y^{-1}\hat{\mathbf{P}}^*r_R$$

donde ρ es la traza de la matriz *hat*, A es de la siguiente forma para el MLE:

$$\mathbf{A} = tr((\hat{\mathbf{V}}_y^{-1}\hat{\mathbf{P}}^*)^2) - 2r_R^{'}\hat{\mathbf{V}}_y^{-1}\hat{\mathbf{P}}^*r_R$$

y para el REML:

$$\mathbf{A} = tr((\hat{\mathbf{V}}_y^{-2}) - 2r_R^{'}\hat{\mathbf{V}}_y^{-1}\hat{\mathbf{P}}^*r_R)$$

siendo

$$\begin{aligned} r_R &= \hat{\mathbf{V}}_y^{-1}\hat{\mathbf{P}}^*\mathbf{y} \\ \hat{\mathbf{P}}^* &= (\mathbf{I}-\mathbf{X})(\mathbf{X}'\hat{\mathbf{V}}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_y^{-1} \end{aligned} \tag{2.8}$$

donde $\hat{\mathbf{V}}_y$ es la estimación de la varianza de la variable respuesta y en el modelo, es decir $\hat{\mathbf{V}}_y = \hat{\mathbf{V}}_u + \mathbf{V}_\varepsilon$.

2.4.2. xGAIC.

Los coeficientes mAIC y cAIC tienen diversos inconvenientes, pues en primer lugar requiere de nuestro criterio el uso de uno o otro, según si le damos más importancia a los parámetros $\boldsymbol{\beta}$ o a los efectos aleatorios de las áreas, si bien en el contexto de SAE se suele optar por el cAIC. Otra cuestión es el valor K, ya que en este contexto es difícil de definir y determinar. En *Lombardía et al.(2017)* se sugiere una solución a esto, mediante el uso de otra versión del AIC pensada en un contexto general, denominado xGAIC.

El xGAIC, manteniendo la estructura general del AIC, no sigue ni un enfoque condicional ni uno marginal, si no que combina ambos, utilizando una cuasi-verosimilitud en lugar de las verosimilitudes marginal y condicional. Este criterio también se basa en la definición de grados de libertad generalizados que puede encontrarse en *You et al. (2016)*, de forma que se establece de forma más clara el valor de los grados de libertad. De esta forma el xGAIC se define de la siguiente forma:

$$x\text{GAIC} = -2\log(l_x(\hat{M})) + x\widehat{GDF}$$

donde $log(l_x(\hat{M}))$ es la cuasi log-verosimilitud, que se define, considerando el modelo general antes definido para D observaciones, como:

$$\log(l_x(M)) = \frac{1}{2}D\log(2\pi) - \frac{1}{2}|\mathbf{V}_y| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}_y^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

$x\widehat{GDF}$ se trata de la estimación del xGDF, una extensión del concepto de GDF de *You et al. (2016)* de modelos a nivel de unidad a modelos a nivel de área (ver *Lombardía et al.(2017)* para más detalles), que se define como:

$$x\text{GDF} = \sum_{d=1}^{D}\frac{\partial E_y(\hat{\mu}_d)}{\partial\theta_d} = \sum_{d=1}^{D}\sum_{i=1}^{D}V_y^{di}\text{cov}_y(\hat{\mu}_d, y_i)$$

donde $\hat{\mu}_d$ es el estimador de μ_d y V_y^{di} es el elemento (d,i) de la matriz $\mathbf{V}_y^{-1}$. Observamos que el xGDF es una medida de la sensibilidad de la estimación de la respuesta respecto a sus respectivas medias.

En general xGDF no es fácil de calcular de forma analítica, por lo que utilizamos una estimación $x\widehat{GDF}$. En *Lombardía et al.(2017)* sugieren obtener esta estimación utilizando un bootstrap paramétrico:

1. Estimamos los parámetros del modelo mediante el procedimiento EBLUP.
2. Repetimos B veces $(b = 1, ..., B)$ los siguientes pasos:
   (a) Generamos la parte aleatoria del modelo $u_d^* \sim N(0, \hat{\sigma}_u^2)$ y $\epsilon_d^* \sim N(0, \sigma_{\boldsymbol{\varepsilon}}^2)$ independientes para cada $d = 1, ..., D$, construyendo el modelo bootstrap $y_d^{*(b)} = \mu_d^{*(b)} + \varepsilon_d^{*(b)}$, con $\mu_d^{*(b)} = \mathbf{x}'_d\hat{\boldsymbol{\beta}} + u_d^{*(b)}$ y la matriz de covarianzas $\hat{\mathbf{V}}_y$.
   (b) Para cada muestra bootstrap $\{y_d^{*(b)}, \mathbf{x}_d\}$ calculamos $\hat{\mu}_d^{*(b)} = \mathbf{x}'_d\hat{\boldsymbol{\beta}}^{*(b)} + \hat{u}_d^{*(b)}$.
3. Aproximamos xGDF mediante muestreo Monte Carlo, es decir:

$$x\widehat{\text{GDF}} = \sum_{d=1}^{D}\sum_{i=1}^{D}\frac{1}{B-1}\sum_{b=1}^{B}V_y^{*(b),di}(\hat{\mu}_d^{*(b)} - \bar{\hat{\mu}}_d^*)(y_i^{*(b)} - \bar{y}_i^*)$$

donde $\bar{\hat{\mu}}_i^* = (1/B)\sum_{b=1}^{B}\hat{\mu}_i^{*(b)}$, $\bar{\hat{y}}_i^* = (1/B)\sum_{b=1}^{B}y_i^{*(b)}$ y $V_y^{*(b),di}$ es el elemento (d,i) de la inversa de la matriz $\mathbf{V}_y^{*(b)}$.

Como ya comentamos, este criterio tiene la ventaja de no partir de un enfoque inicial, siendo entonces más general que los criterios $mAIC$ y el $cAIC$. Además de lo anterior, en *Lombardía et al.(2017)* se muestra mediante simulación que el xGAIC muestra un mejor comportamiento que los otros criterios ya mencionados.

2.4.3. cYMO.

En *You et al. (2016)* se plantea otro criterio de selección. Dicho criterio es similar al AIC, puesto que mantiene su estructura general, combinando en este caso el enfoque condicional, con la definición de grados de libertad que se realiza en dicha referencia, es decir, los GDF. Teniendo en cuenta lo anterior, dicho criterio (que denotaremos por cYMO) aplicado a nuestros modelos es de la forma:

$$cYMO = -2log(l_c(M)) + x\widehat{GDF}$$

donde $log(l_c(M))$ es la log-verosimisitud condicional para el modelo M.

2.5. Paquetes en R para el ajuste.

Como ya comentamos anteriormente, en este trabajo aplicaremos lo mostrado a lo largo de este capítulo en un caso con datos reales. Para poder hacer esto nos serviremos del software estadístico R (*R core team (2018)*). En concreto, en R podemos encontrar diversos paquetes orientados al contexto SAE. Algunos ejemplos son el paquete *sae*(*Molina y Marhuenda (2015)*) y el paquete *saery*(*Esteban et al. (2014)*). Si bien el primero es seguramente al más utilizado, nosotros nos centraremos en el uso del segundo.

El *saery* es un paquete centrado en los modelos de Fay-Herriot con efecto temporal como el visto en la Sección 2.2, permitiéndonos ajustar modelos de dicha clase, así como de estimar el MSE de las estimaciones realizadas por el modelo. Además de lo anterior, el paquete permite considerar otros tipos de efecto temporal además del AR(1), en concreto permite efectos temporales MA(1) o independientes.

Respecto a los criterios de selección adaptados para esta clase de modelos, estos no están integrados en el paquete ni (hasta donde conocemos) en ningún otro , por lo que los criterios que utilizaremos tuvieron que ser programados. El código correspondiente a ellos puede verse en el Apéndice A.

Capítulo 3

Descriptiva de los datos.

En el Capítulo 1 comentábamos los objetivos de la EEH, así como los datos que utilizaremos. Como ya hemos comentado en dicho capítulo, nuestro objetivo es mejorar las estimaciones realizadas por los estimadores directos respecto de los distintos tipos de ingresos, utilizando la metodología SAE.

Para ello, nos serviremos de los datos de la EEH entre 2007 y 2016, de forma que las variables que estudiaremos serán:

- **ITOT**. Se trata de la media de los ingresos totales de cada uno de los hogares en cada área y para cada año.
- **IAJ**. Se trata de la media de los ingresos por cuenta ajena de cada uno de los hogares que poseen ingresos de este tipo en cada área y para cada año.
- **IPROP**. Se trata de la media de los ingresos por cuenta propia de cada uno de los hogares que poseen ingresos de este tipo en cada área y para cada año.
- **ICON**. Se trata de la media de los ingresos por prestaciones contributivas de cada uno de los hogares que poseen ingresos de este tipo en cada área y para cada año.

Además de estas variables, que actuarán como respuestas en los modelos que plantearemos en las secciones posteriores, son necesarias otras variables que nos servirán como explicativas, siendo las siguientes:

- **Grupos de edad**. Se trata del porcentaje de personas de una cierta edad en cada área y para cada año. Distinguimos 3 grupos:

  PM18: Porcentaje de personas menores de 18 años.

  P18a65: Porcentaje de personas menores de 65 años y mayores de 18.

  PM65: Porcentaje de personas mayores de 65 años.

- **Nacionalidad**. Porcentaje de personas en cada área según su nacionalidad y para cada año. Distinguimos 2 grupos

  PEXTR: Porcentaje de extranjeros.

  PNEXTR: Porcentaje de españoles.

- **Nivel de estudios**. Porcentaje de personas en cada área según su nivel de estudios y para cada año. Distinguimos 4 categorías:

  PENA: Porcentaje de personas menores de 16 años.

  PEPRIM: Porcentaje de personas con estudios primarios.

PESEC: Porcentaje de personas con estudios secundarios.

PESUP: Porcentaje de personas con estudios superiores.

- **Umbral de pobreza**. Porcentaje de hogares bajo o sobre el umbral de pobreza (60 % de la mediana de los ingresos totales) en cada área y para cada año.

  PFBLIM: Porcentaje de hogares bajo el umbral de pobreza.

  PFSLIM: Porcentaje de hogares sobre el umbral de pobreza.

- **Tipología de los hogares**. Porcentaje de personas según la tipología del hogar en cada área y para cada año. Distinguimos 7 categorías:

  PFTIPO1: Porcentaje de hogares unipersonales.

  PFTIPO2: Porcentaje de hogares en los que no hay un núcleo familiar.

  PFTIPO3: Porcentaje de hogares conformados por una pareja y sus hijos.

  PFTIPO4: Porcentaje de hogares conformados por parejas sin hijos.

  PFTIPO5: Porcentaje de hogares monoparentales.

  PFTIPO6: Porcentaje de hogares conformados por un núcleo familiar y personas ajenas al mismo.

  PFTIPO7: Porcentaje de hogares conformados por varios núcleos.

- **REND**: Rendimiento medio declarado en cada área entre los años 2007 y 2016.

- **PENMEAN**: Pensión media en cada área entre los años 2007 y 2016.

Considerando las variables que acabamos de presentar, en este capítulo realizaremos un análisis exploratorio de las mismas, en el que comprobaremos si se ajustan a las hipótesis de los modelos presentados en el capítulo anterior y buscaremos cuales son las variables auxiliares que deberemos tener en cuenta para los modelos que ajustaremos en los siguientes capítulos.

3.1. Ingresos totales medios.

En esta sección estudiaremos la variable ITOT, estudiando si cumple los requisitos de normalidad y seleccionando entre las posibles variables explicativas aquellas que estén más relacionadas con la variable. Lo primero que veremos es si es necesario aplicar la metodología SAE para obtener buenos estimadores.
En la Tabla 3.1 podemos ver el mínimo, el máximo y la mediana del tamaño muestral de la EEH en las áreas en el período 2007-2016. En ella podemos ver una gran diferencia entre los mínimos y los máximos para alguas áreas tenemos tamaños muestrales pequeños, puesto que no hay mucha diferencia entre los mínimos y los primeros cuartiles lo que puede provocar que los estimadores directos sean poco precisos. De esta forma, es razonable plantear el uso de la metodología SAE.

En la Figura 3.1 podemos ver la evolución de la variable ITOT entre los años 2007 y 2016. Observamos que algunas estimaciones sufren cambios bruscos al pasar de un año a otro. Un ejemplo de esto es el caso de Pontevedra sur (área 2 de Pontevedra), donde observamos cambios bruscos entre los años 2010 y 2012.

Teniendo en cuenta lo anterior, nuestro objetivo es reducir esas diferencias en las estimaciones. Para ello utilizaremos la metodología SAE, en concreto, utilizando modelos mixtos de Fay-Herriot con efectos temporales.

ITOT				
Año	Mínimo	Primer cuartil	Mediana	Máximo
2007	271	315.2	383.5	1271
2008	287	319.5	395	1221
2009	287	288	343	1062
2010	285	288	342	1060
2011	287	288	341.5	1060
2012	287	288	341.5	1058
2013	284	287.5	341	1059
2014	288	288	341.5	1056
2015	287	288	341	1058
2016	288	288	341	1059

Tabla 3.1: Mínimos, máximos, mediana y primer cuartil de las áreas en el período 2007-2016 para la variable ITOT.

Es conveniente que la variable ITOT cumpla ciertos requisitos. Uno de ellos es que los coeficientes de variación de las estimaciones realizadas por el estimador directo deben mantenerse hasta cierto punto estables a lo largo del tiempo, puesto que en caso contrario podría ser que no fuéramos capaces de reducir lo suficiente la varianza de las estimaciones. Esto lo podemos ver en las Figuras 3.2 y 3.3. En la primera podemos ver dos mapas de Galicia con los coeficientes de variación de las estimaciones de la variable ITOT en cada área en el período 2007-2016. En ellos podemos observar que hay variaciones en las áreas, aunque la diferencia es ligera. Esto podemos verlo también en la Figura 3.3, que muestra los boxplots de los coeficientes de variación en cada año. En ella podemos observar que no se muestra una gran diferencia en la distribución de los coeficientes de variación en los distintos años.

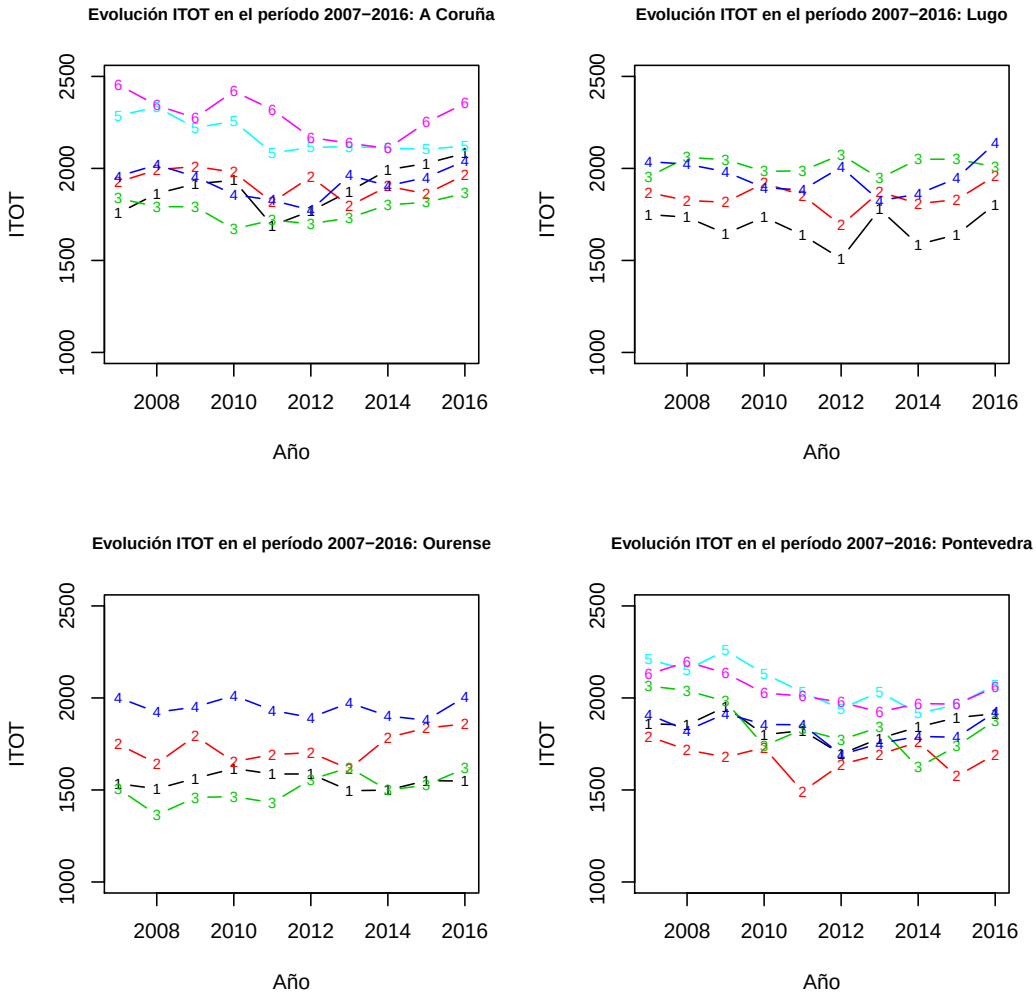

Figura 3.1: Evolución de la variable ITOT en las distintas áreas en el período 2007-2016. Se muestran las estimaciones según la provincia y su área dentro de la misma (indicada por su numeración dentro de su correspondiente provincia).

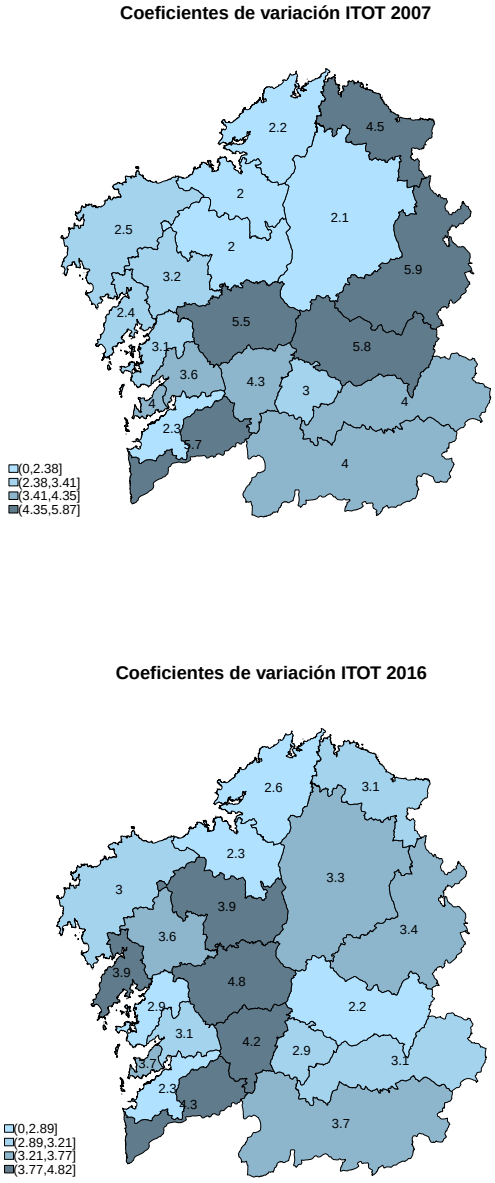

Figura 3.2: Mapas de los coeficientes de variación correspondientes a la variable ITOT por área en 2007 y 2016.

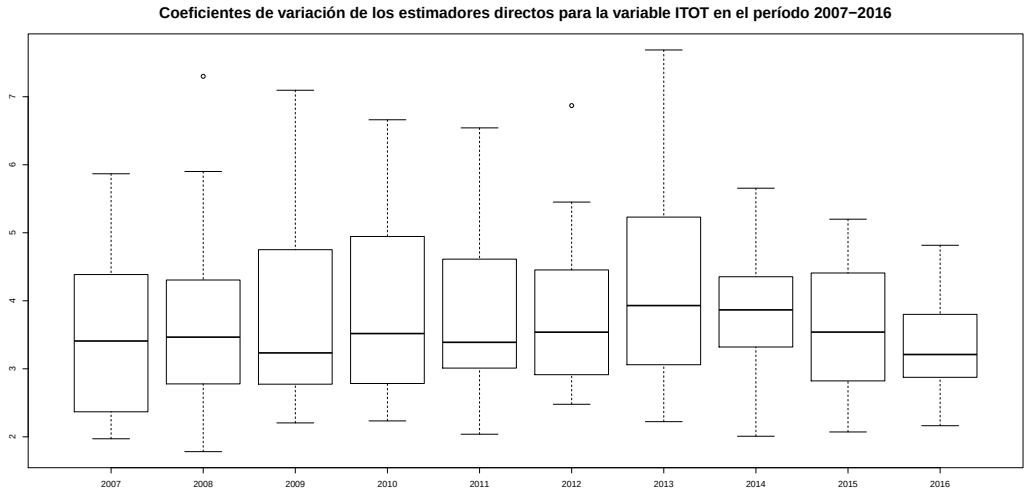

Figura 3.3: Boxplots de los coeficientes de variación de la variable ITOT en el período 2007-2016.

El otro requisito que es conveniente que cumpla es que su distribución sea normal, puesto que, como ya hemos visto en el Capítulo 2, gran parte de la metodología planteada se basa en la normalidad. En la Figura 3.4 podemos ver el histograma y el qqplot de la variable en cuestión. En ellos observamos que parece cumplir la hipótesis de normalidad, algo que se confirma al observar los p-valores obtenidos en el test de Kolmogorov-Smirnov-Lilliefors para el que se obtiene un valor de 0.93, por lo que parece razonable asumir la normalidad de la variable.

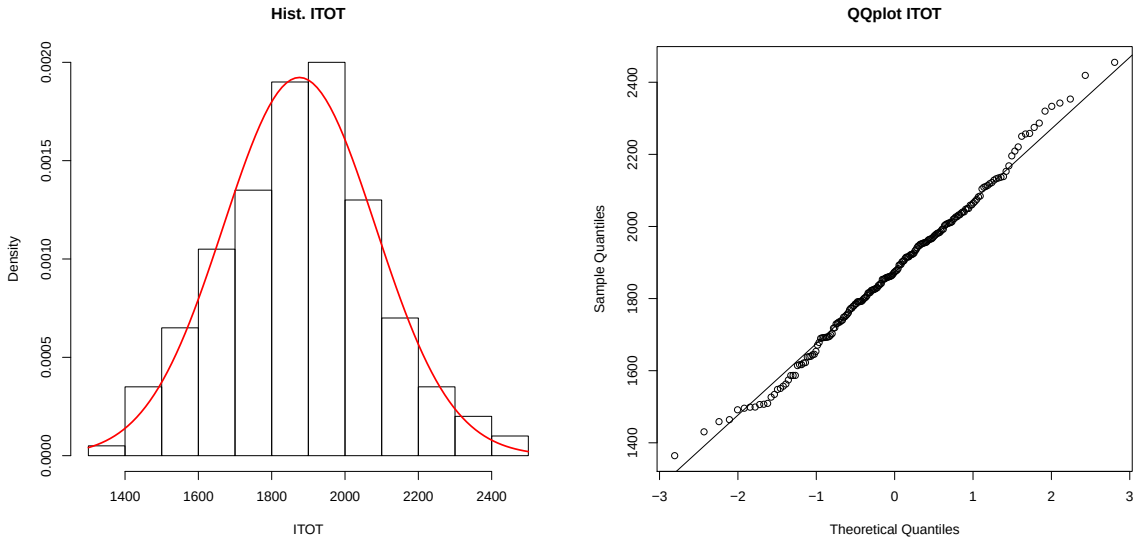

Figura 3.4: Histograma y QQplot de los ingresos totales medios.

El siguiente paso es encontrar que variables auxiliares de las que hemos comentado al principio del capítulo son más razonables para explicar la variable ITOT. Como los modelos que planeamos utilizar son modelos lineales, parece entonces razonable escoger aquellas que tengan una fuerte relación lineal

con la variable ITOT. Esta relación la podemos ver utilizando el coeficiente de correlación lineal de Pearson. En la Tabla 3.2 pueden verse las correlaciones de aquellas con una correlación relativamente alta con la variable ITOT (en este caso superior a 0.65). Observamos que aparentemente los niveles de estudios PEPRIM y PSUP junto con la variable REND son las que tienen una correlación más fuerte, seguidas de los grupos de edad P18a65 y PM65 y la variable PFBLIM. Cabe destacar que en las categoría de porcentaje, la suma de las variables es 1, razón por la que no incluimos una de las categorías, puesto que estarían relacionadas linealmente con las demás del mismo tipo, como podría pasar de incluirse la variable PFSLIM.

Si bien la correlación de Pearson suele indicar si dos variables están linealmente relacionadas, esto no

	ITOT	P18a65	PM65	PEPRIM	PESUP	PFBLIM	REND
ITOT	1.00	0.68	-0.63	-0.74	0.74	-0.67	0.76
P18a65	0.68	1.00	-0.97	-0.65	0.49	-0.31	0.74
PM65	-0.63	-0.97	1.00	0.67	-0.50	0.23	-0.74
PEPRIM	-0.74	-0.65	0.67	1.00	-0.93	0.48	-0.88
PESUP	0.74	0.49	-0.50	-0.93	1.00	-0.53	0.81
PFBLIM	-0.67	-0.31	0.23	0.48	-0.53	1.00	-0.38
REND	0.76	0.74	-0.74	-0.88	0.81	-0.38	1.00

Tabla 3.2: Correlaciones de ITOT con las variables explicativas (con correlación superior a 0.6)

siempre es así, por lo que es recomendable representar los diagramas de dispersión para ver si efectivamente hay una relación lineal. En las Figuras 3.5, 3.6 y 3.7 podemos ver que efectivamente tenemos una relación lineal entre las variables, por lo que parece razonable incluirlas en el modelo.

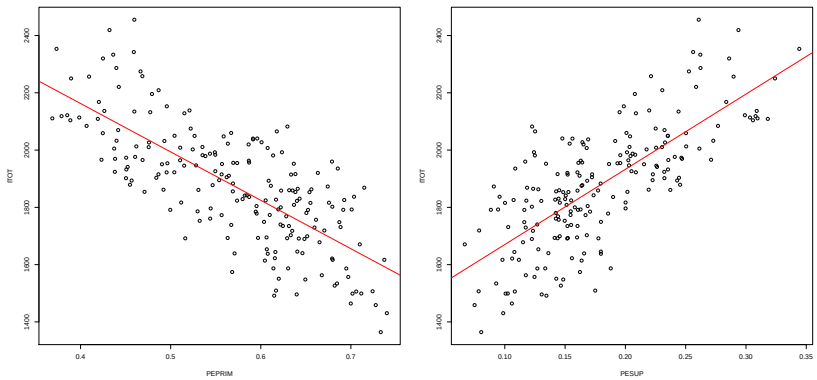

Figura 3.5: Diagramas de dispersión de los ingresos totales medios contra los distintos niveles de estudios.

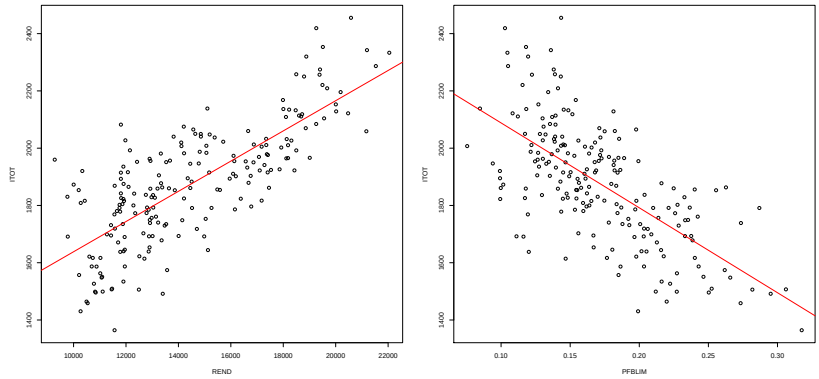

Figura 3.6: Diagramas de dispersión de los ingresos totales medios contra el rendimiento y los hogares bajo el umbral de pobreza.

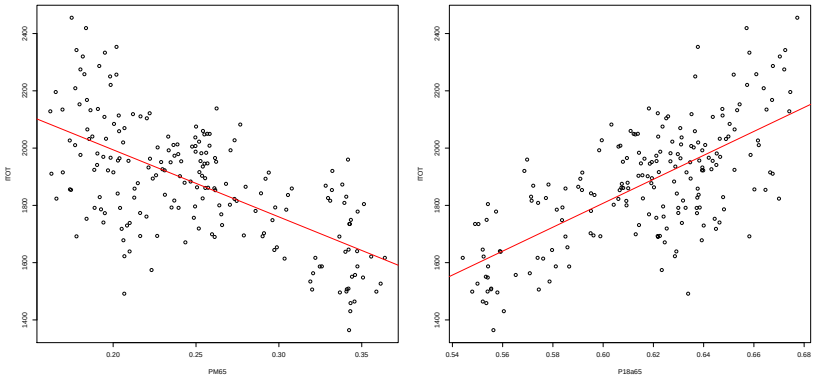

Figura 3.7: Diagramas de dispersión de los ingresos totales medios contra los distintos grupos de edad.

3.2. Ingresos por cuenta ajena medios.

En esta sección estudiaremos la variable IAJ, de forma similar a lo realizado en la sección anterior para la variable ITOT. En la Figura 3.8 podemos ver la evolución de la variable IAJ entre 2007 y 2016, en la que, al igual que ocurría para la variable ITOT, observamos cambios drásticos en las estimaciones, por lo que trataremos de disminuir esas diferencias.
En la Tabla 3.3 podemos ver el mínimo, el máximo y la mediana del tamaño muestral de la EEH en

IAJ				
Año	Mínimo	Primer cuartil	Mediana	Máximo
2007	112	139.8	193	821
2008	121	150.2	193.5	993
2009	109	136	172	644
2010	114	132.5	167.5	683
2011	91	130.2	171	660
2012	90	115.8	157	624
2013	98	123	153	621
2014	90	127	157.5	648
2015	91	118.5	156	624
2016	96	127	162	618

Tabla 3.3: Mínimos, máximos, mediana y primer cuartil de las áreas en el período 2007-2016 para la variable IAJ.

las áreas en el período 2007-2016 para la variable IAJ. En ella, al igual que para la variable ITOT, podemos ver que para algunas áreas tenemos tamaños muestrales pequeños, puesto que no hay mucha diferencia entre los mínimos y los primeros cuartiles, además de, nuevamente, tener una gran diferencia entre los mínimos y los máximos. Todo eso puede provocar que los estimadores directos sean poco precisos, por lo que es razonable plantear el uso de la metodología SAE.

En los mapas de la Figura 3.9 observamos que si bien en determinadas áreas hay ciertos cambios en los coeficientes de variación de las estimaciones, los cambios no suelen ser muy marcados, algo que puede verse en los boxplots de la Figura 3.10, puesto que en general las distribuciones de los errores cada año no difieren demasiado entre ellas, por lo que parece razonable decir que se mantienen estables.

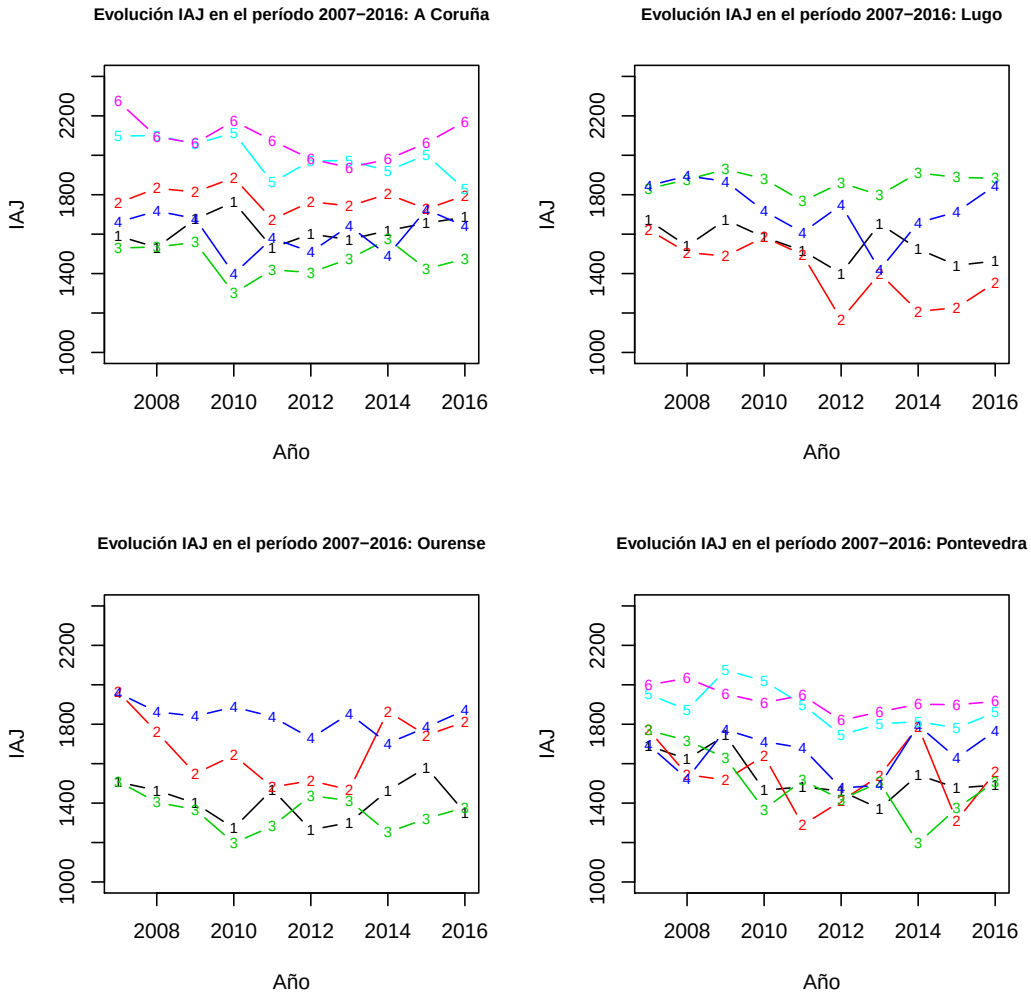

Figura 3.8: Evolución de los ingresos por cuenta ajena medios en las distintas áreas en el período 2007-2016. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

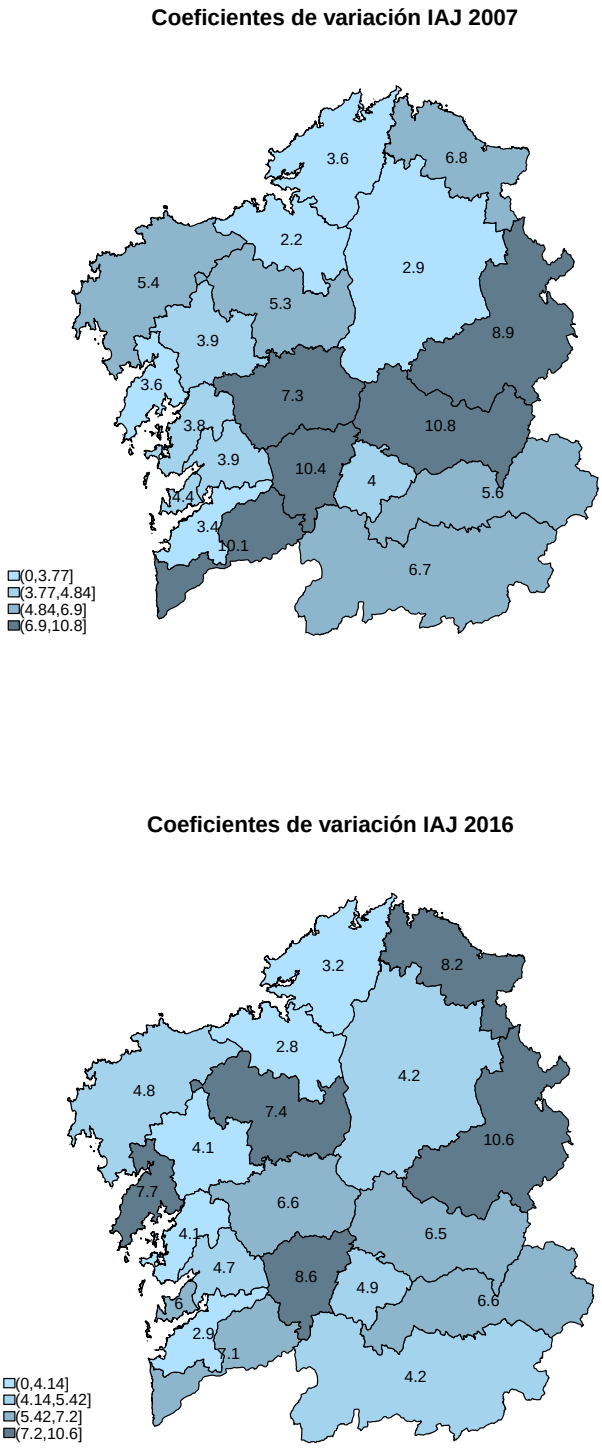

Figura 3.9: Mapas de los coeficientes de variación de las estimaciones de la variable IAJ por área en los años 2007 y 2016.

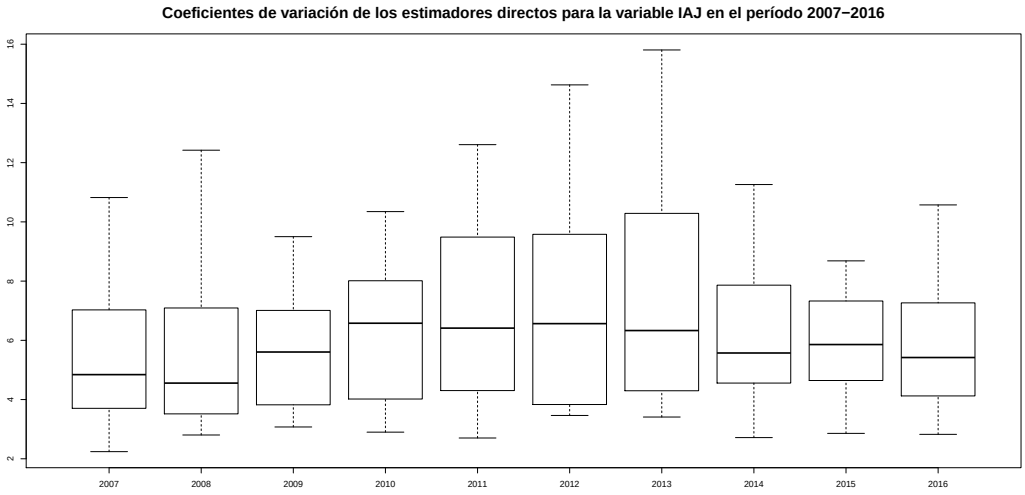

Figura 3.10: Boxplots de los coeficientes de variación de las estimaciones de la variable IAJ en el período 2007-2016.

Al igual que hicimos para la variable ITOT, debemos comprobar si la variable IAJ sigue una distribución normal. En la Figura 3.11 podemos ver el histograma y el qqplot de la variable IAJ, donde el primero si parece indicar que sigue una distribución normal mientras el qqplot parece mostrar una cierta desviación respecto de la normal. Algo parecido sucede con el contraste de Komogorov-Smirnov-Lilliefors, puesto que si bien se rechaza al 5 % ya que se obtiene un p-valor de 0,01,no se rechaza al 1 %. Al tratarse de datos reales no es raro que esto suceda aunque la variable sea normal, por lo que consideraremos que la variable IAJ es normal.

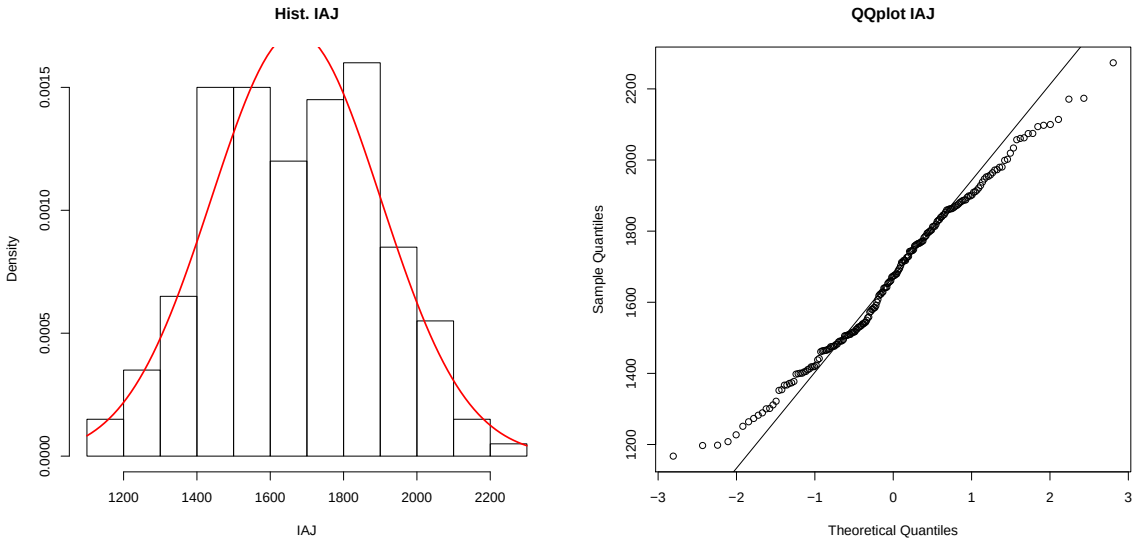

Figura 3.11: Histograma y QQplot de la variable IAJ.

Nos corresponde ahora escoger que variables explicativas utilizaremos en el modelo que vamos a plantear. Para ello procedemos de igual forma que para la variable ITOT, seleccionando nuevamente

aquellas que tengan una mayor correlación con la variable IAJ, que en este caso podemos considerar aquellas con una correlación en valor absoluto igual o superior a 0.55.

	IAJ	P18a65	PM65	PEPRIM	PESUP	REND
IAJ	1.00	0.61	-0.58	-0.77	0.77	0.85
P18a65	0.61	1.00	-0.97	-0.65	0.49	0.74
PM65	-0.58	-0.97	1.00	0.67	-0.50	-0.74
PEPRIM	-0.77	-0.65	0.67	1.00	-0.93	-0.88
PESUP	0.77	0.49	-0.50	-0.93	1.00	0.81
REND	0.85	0.74	-0.74	-0.88	0.81	1.00

Tabla 3.4: Correlaciones entre la variable IAJ y las variables explicativas (superiores a 0.55)

En la Tabla 3.4 observamos que REND, PESUP y PEPRIM son las que muestran una correlación más alta (especialmente el rendimiento), mientras que P18a65 y PM65 parecen mostrar una relación más débil. Estos resultados los confirmamos con los diagramas de dispersión de las Figuras 3.12, 3.13 y 3.14, donde podemos ver que la relación lineal es parecida a la descrita por el coeficiente de correlación.

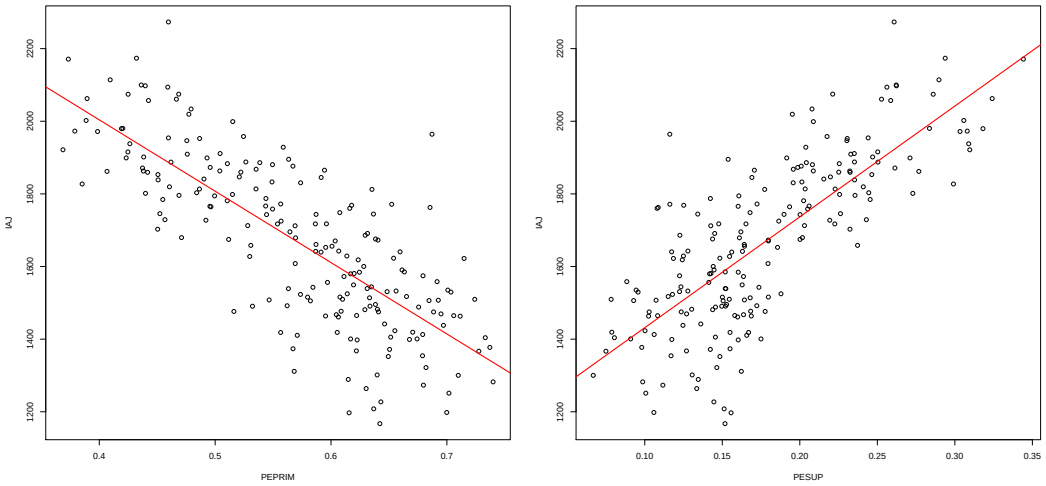

Figura 3.12: Diagramas de dispersión de la variable IAJ contra los distintos niveles de estudios.

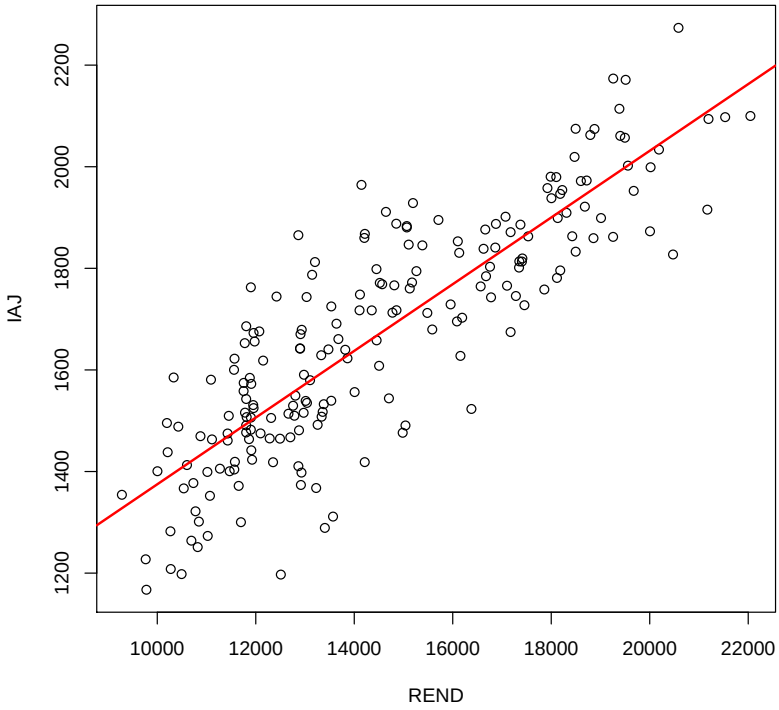

Figura 3.13: Diagramas de dispersión de la variable IAJ contra el rendimiento.

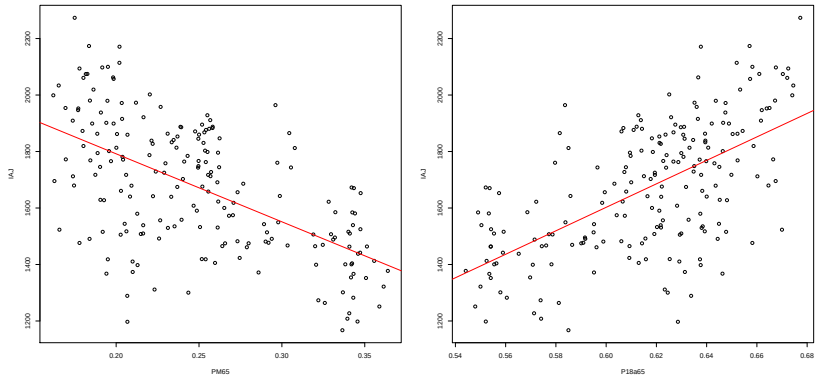

Figura 3.14: Diagramas de dispersión de la variable IAJ contra los distintos grupos de edad.

3.3. Ingresos por cuenta propia medios.

En esta sección estudiaremos la variable IPROP de forma análoga a lo hecho en las anteriores secciones. En la Figura 3.15 podemos ver la evolución de la variable IPROP con el tiempo, mostrándonos una vez más que tenemos cambios bruscos entre los distintos años, en este caso más causados que en las variables de las dos secciones anteriores.

En la Tabla 3.5 podemos ver el mínimo, el máximo y la mediana del tamaño muestral de la EEH en

IPROP				
Año	Mínimo	Primer cuartil	Mediana	Máximo
2007	37	73	96.5	229
2008	45	70.25	94.5	213
2009	48	59	87	187
2010	44	71	85.5	200
2011	42	61	77	189
2012	37	60.75	79	170
2013	37	55.75	80	186
2014	36	54	75	181
2015	33	56	64.5	182
2016	35	58	77.5	165

Tabla 3.5: Mínimos, máximos, mediana y primer cuartil de las áreas en el período 2007-2016 para la variable IPROP.

las áreas en el período 2007-2016 para la variable IPROP. En ella podemos ver que, al igual que ara

la variable IAJ, tenemos algunas áreas con tamaños muestrales pequeños, puesto que no hay mucha diferencia entre los mínimos y los primeros cuartiles, de forma que es razonable plantear el uso de la metodología SAE.

En cuanto a los coeficientes de variación de las estimaciones, en los mapas de la Figura 3.16 observamos que en algunas áreas como Lugo oriental hay diferencias notables, pero en cambio en los boxplots de la Figura 3.17 observamos que si bien en 2016 parecen ser algo menores, la distribución de los coeficientes de variación sigue pareciendo similar todos los años.

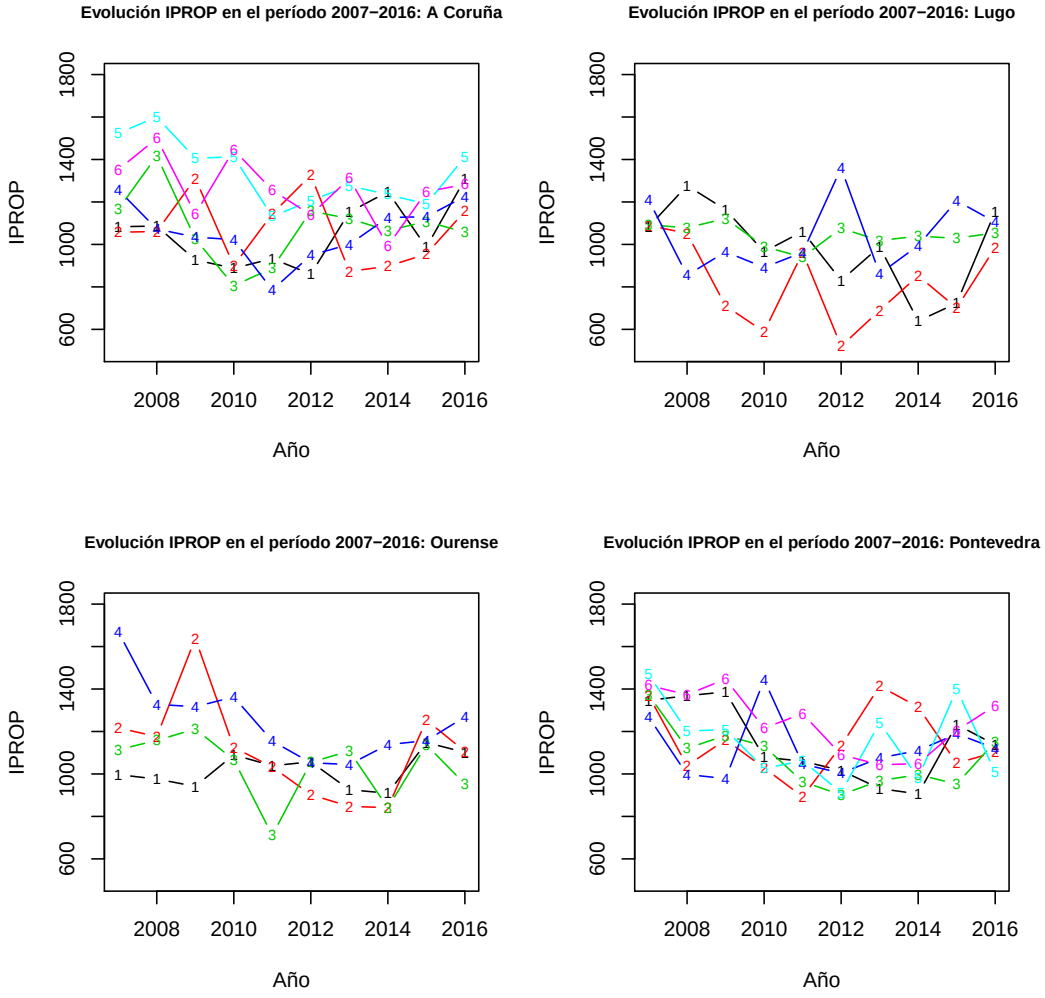

Figura 3.15: Evolución de la variable IPROP en las distintas áreas en el período 2007-2016. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

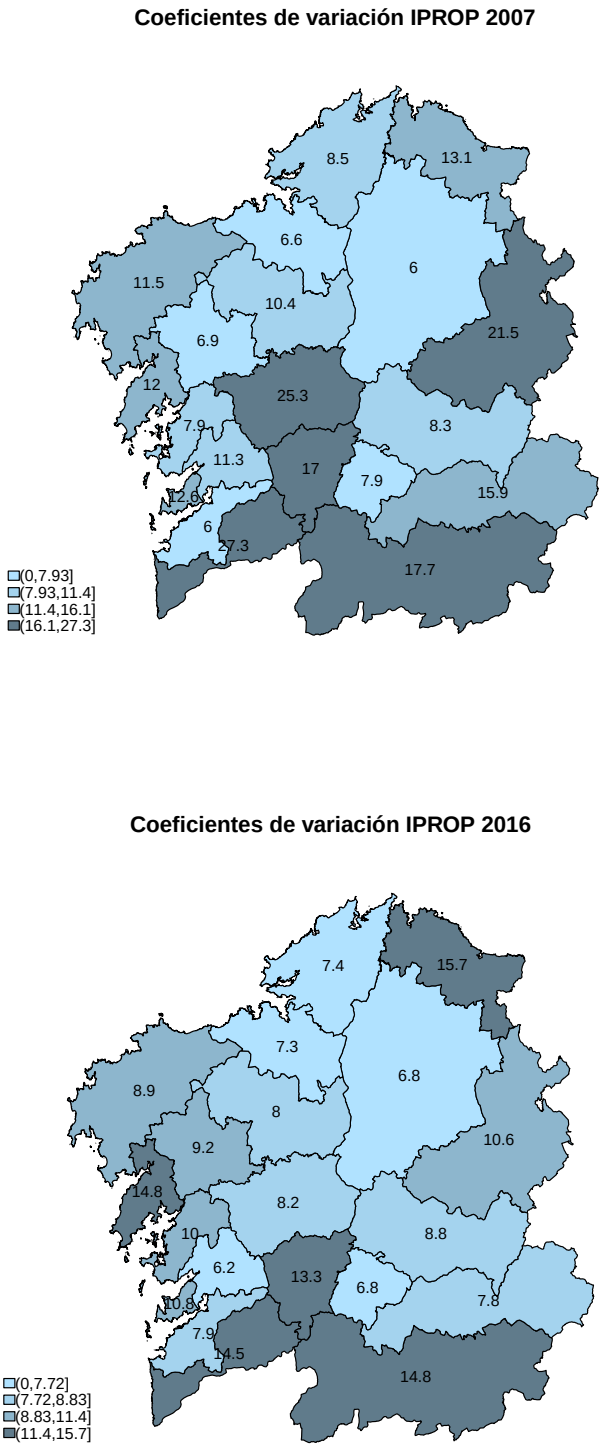

Figura 3.16: Mapas de los coeficienes de variación correspondientes a la variable IPROP en cada área en 2007 y 2016.

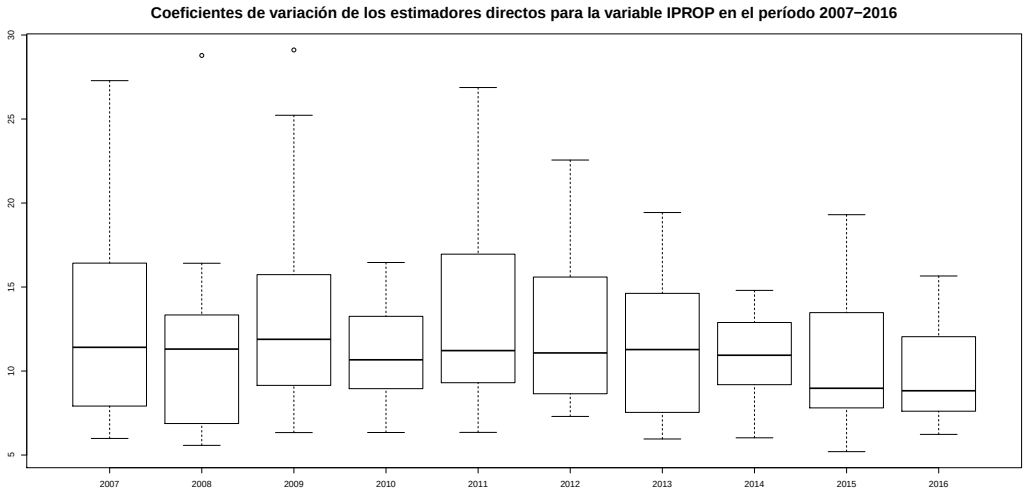

Figura 3.17: Boxplots de los coeficientes de variación de la variable IPROP en el período 2007-2016.

Al igual que sucedía con la variable IAJ, el histograma de la Figura 3.18 de la variable IPROP parece mostrar que sigue una distribución normal, mientras que el qqplot de la misma Figura parece mostrar cierta desviación respecto de la normal. En el contraste de normalidad sucede algo similar a lo visto para la variable IAJ, puesto que obtenemos un p-valor de 0,02, lo que significa que tenemos significación al 5 % pero no al 1 %, algo que como ya comentábamos para la variable IAJ, puede suceder cuando se trabaja con datos reales, por lo que supondremos que la variable IPROP es normal.

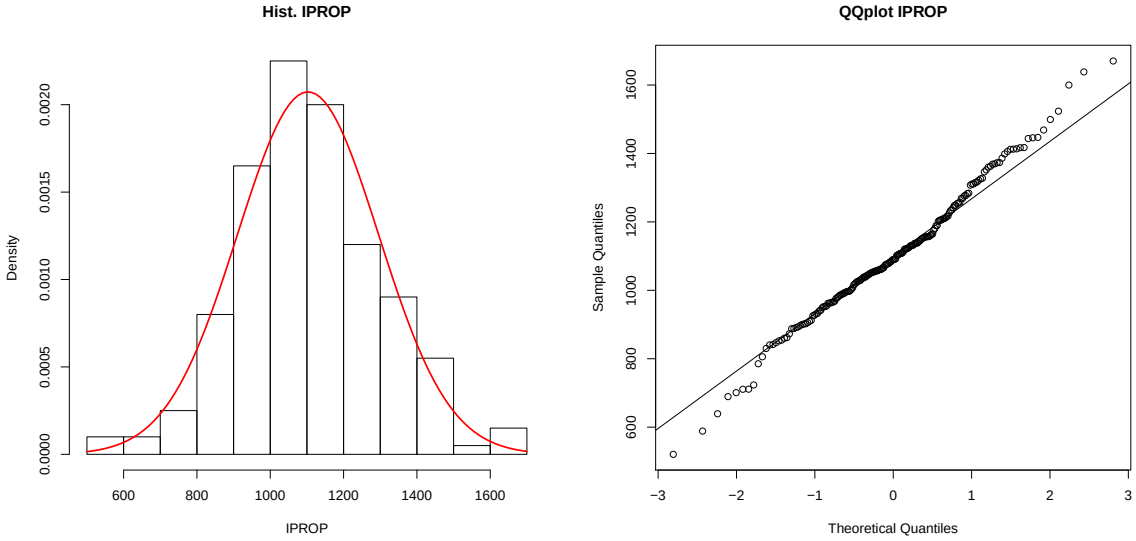

Figura 3.18: Histograma y QQplot de la variable IPROP.

Comprobamos ahora las correlaciones entre IPROP y las demás variables. Observamos que en este caso las correlaciones son mucho menores a las vistas en los casos anteriores, por lo que seleccionaremos aquellas con una correlación superior a 0,4. En la Tabla 3.6 podemos ver las correlaciones entre las distintas variables seleccionadas, las cuales son en este caso P18a65, PM65, PFTIPO3 y REND, siendo esta última la que en principio está más relacionada. Nuevamente, los diagramas de dispersión de las Figuras 3.19 y 3.20 nos confirman la existencia de una relación lineal en todos los casos.

	IPROP	P18a65	PM65	PFTIPO3	REND
IPROP	1.00	0.40	-0.41	0.44	0.54
P18a65	0.40	1.00	-0.97	0.83	0.74
PM65	-0.41	-0.97	1.00	-0.87	-0.74
PFTIPO3	0.44	0.83	-0.87	1.00	0.63
REND	0.54	0.74	-0.74	0.63	1.00

Tabla 3.6: Correlaciones entre la variable IPROP y las variables explicativas (superiores a 0.4)

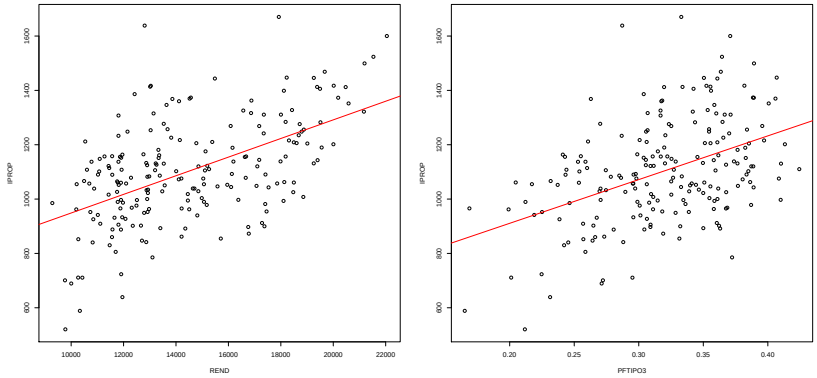

Figura 3.19: Diagramas de dispersión de la variable IPROP contra el rendimiento y los hogares consistentes en parejas con hijos.

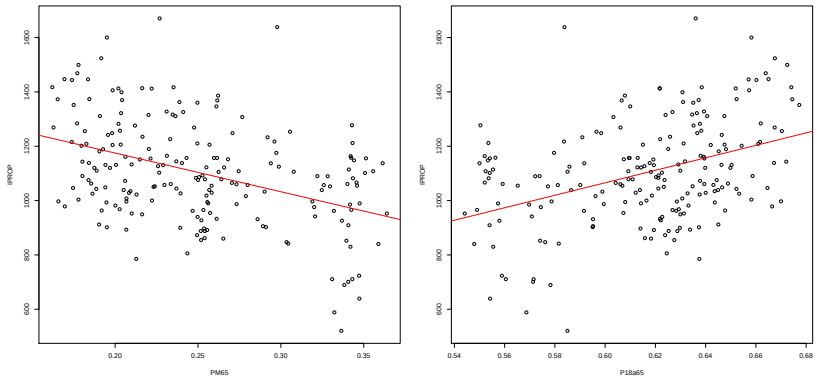

Figura 3.20: Diagramas de dispersión de la variable IPROP contra los distintos grupos de edad.

3.4. Ingresos por prestaciones contributivas medios.

En esta sección estudiaremos la variable ICON, de igual forma a lo visto en las secciones anteriores. En la Figura 3.21 podemos ver la evolución de los estimadores de la variable ICON entre 2007 y 2016 en cada área, observando otra vez que hay cambios abruptos en los estimadores de distintos años, por lo que también corresponde plantear un modelo para reducir dichos cambios. En cuanto a los coeficientes de variación de los estimadores ocurre algo similar a lo visto para ITOT, pues tal y como podemos observar en los mapas de la Figura 3.22, los coeficientes de variación de los estimadores varían ligeramente en las áreas, algo que se puede observar también en los boxplots de la Figura 3.23.

ICON				
Año	Mínimo	Primer cuartil	Mediana	Máximo
2007	143	204.2	229	629
2008	161	199	1229	605
2009	158	182.5	199.5	533
2010	140	193.8	213	482
2011	146	189.8	207	502
2012	141	184.8	213	463
2013	152	189.8	207	496
2014	162	190.5	209.5	501
2015	163	185	217.5	514
2016	158	182.8	211.5	511

Tabla 3.7: Mínimos, máximos, mediana y primer cuartil de las áreas en el período 2007-2016 para la variable ICON.

En la Tabla 3.7 podemos ver el mínimo, el máximo y la mediana del tamaño muestral de la EEH en las áreas en el período 2007-2016 para la variable ICON. Parece que nuevamente tenemos algunas áreas con tamaños muestrales pequeños, puesto que, una vez más, no hay mucha diferencia entre los mínimos y los primeros cuartiles y, al igual que para el resto de ingresos, hay mucha diferencia entre los mínimos y los máximos, lo que puede provocar que los estimadores directos sean poco precisos. Basándonos en ello, consideramos adecuado el uso de la metodología SAE.

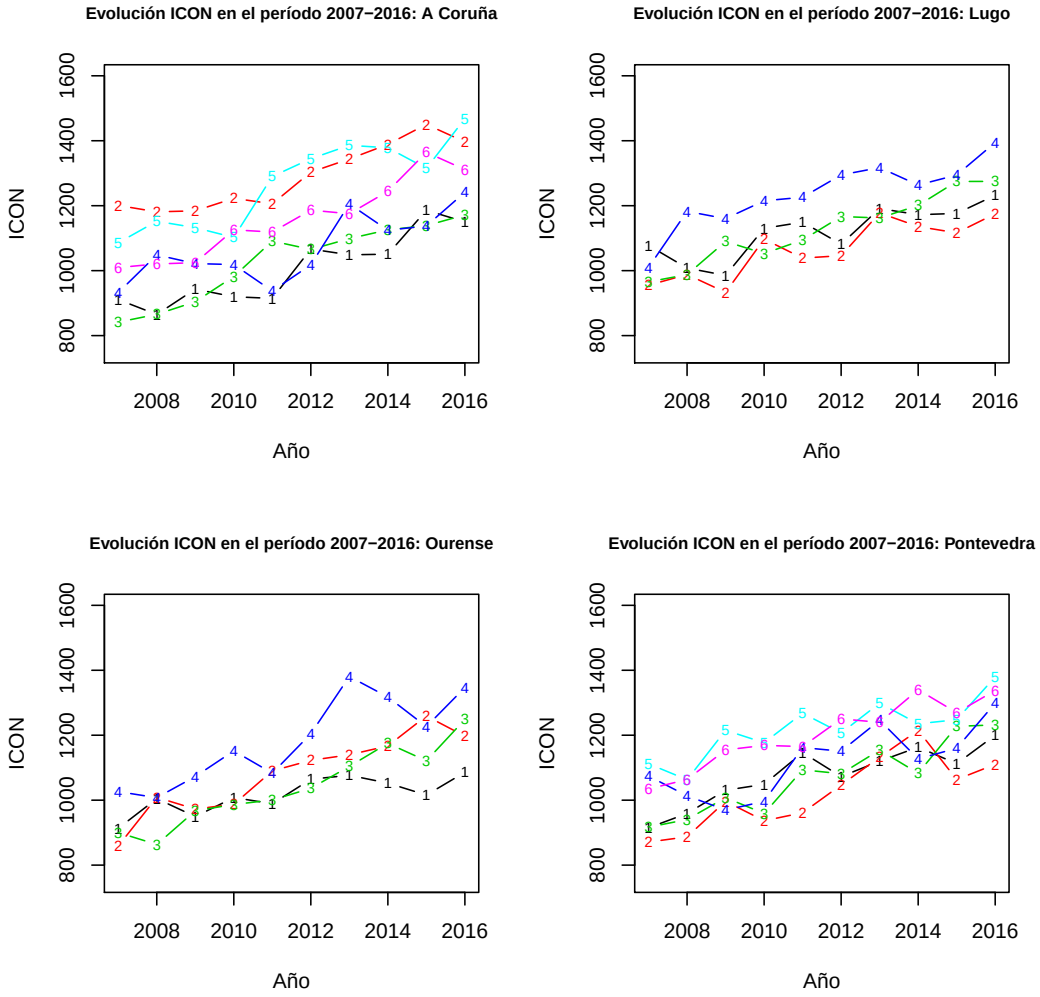

Figura 3.21: Evolución de la variable ICON en cada área en el período 2007-2016. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

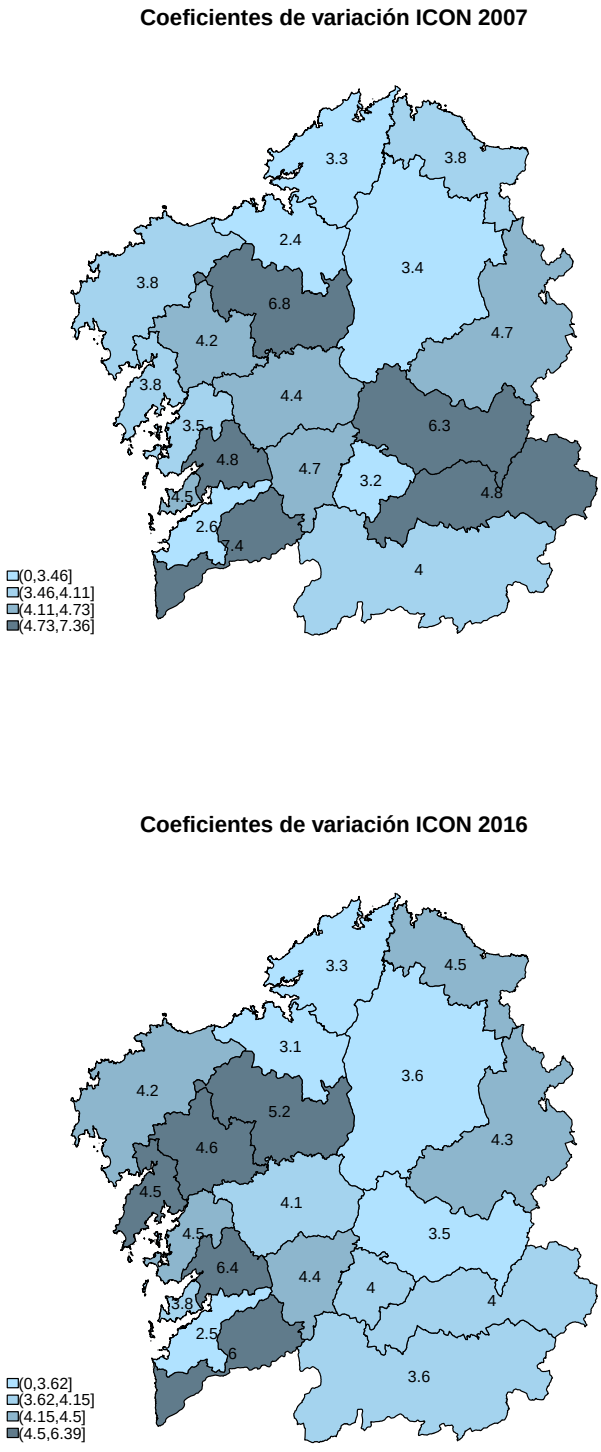

Figura 3.22: Mapas de los coeficientes de variación correspondientes a la variable ICON en cada área en 2007 y 2016.

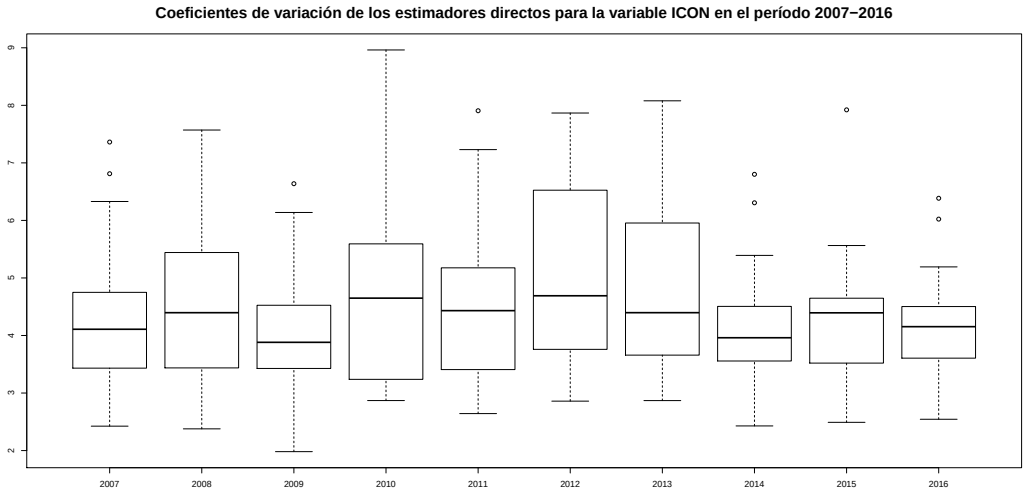

Figura 3.23: Boxplots de los coeficientes de variación de la variable ICON en el período 2007-2016.

La normalidad también parece cumplirse puesto que tanto en el histograma de la variable ICON como su qqplot (pueden verse en la Figura 3.24) parecen mostrar el comportamiento propio de una variable normal, algo que se confirma con los resultados del test de normalidad, puesto que se obtiene un p-valor de 0.68 por lo que podemos no tenemos pruebas significativas par rechazar la normalidad de la variable ICON.

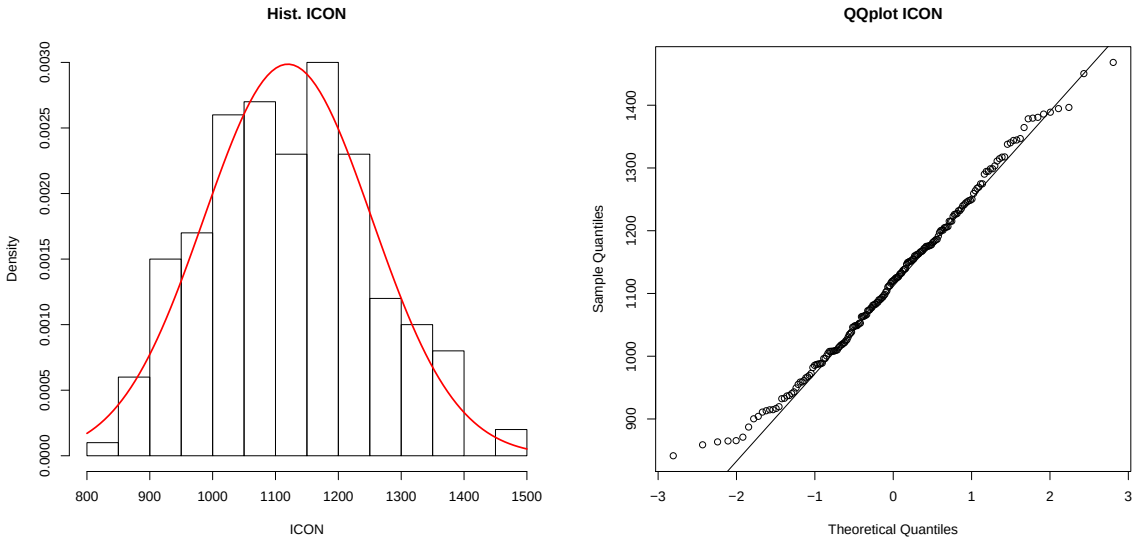

Figura 3.24: Histograma y QQplot de la variable ICON.

Finalmente escogeremos como variables explicativas aquellas con un alta correlación lineal con la variable ICON. A igual que con la variable IPROP, las variables explicativas (salvo la variable PENMEAN), por lo que en este caso escogeremos aquellas variable con una correlación en valor absoluto superior a 0.5. En la Tabla 3.8 podemos ver las correlaciones entre ICON y aquellas variables con una correlación superior en valor absoluto a 0.5. Observamos que en este caso la variable con una mayor correlación es PENMEAN seguida por una gran diferencia de las variables PEPRIM, PNESUP y PNESEC en ese orden, por lo que se concluye qe los principales factores que influyen son las pensiones y el nivel de estudios. Estas relaciones se confirman con los diagramas de dispersión de las Figuras 3.24 y 3.25.

	ICON	PEPRIM	PESEC	PESUP	PENMEAN
ICON	1.00	-0.64	0.56	0.62	0.83
PEPRIM	-0.64	1.00	-0.76	-0.93	-0.77
PESEC	0.56	-0.76	1.00	0.56	0.65
PESUP	0.62	-0.93	0.56	1.00	0.68
PENMEAN	0.83	-0.77	0.65	0.68	1.00

Tabla 3.8: Correlaciones entre la variable ICON y las variables explicativas (superiores a 0.5)

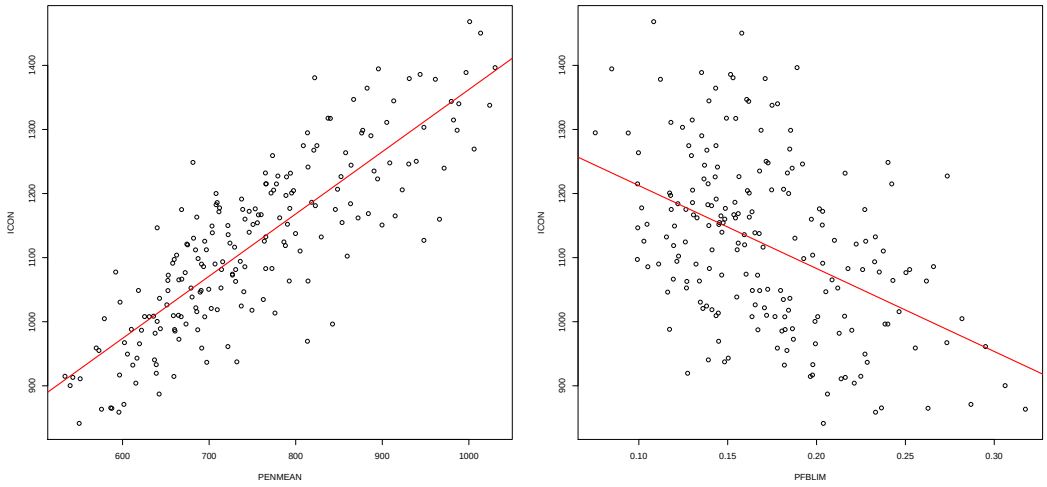

Figura 3.25: Diagramas de dispersión de los ingresos por prestaciones contributivas medios contra las pensiones y los hogares bajo el umbral de pobreza.

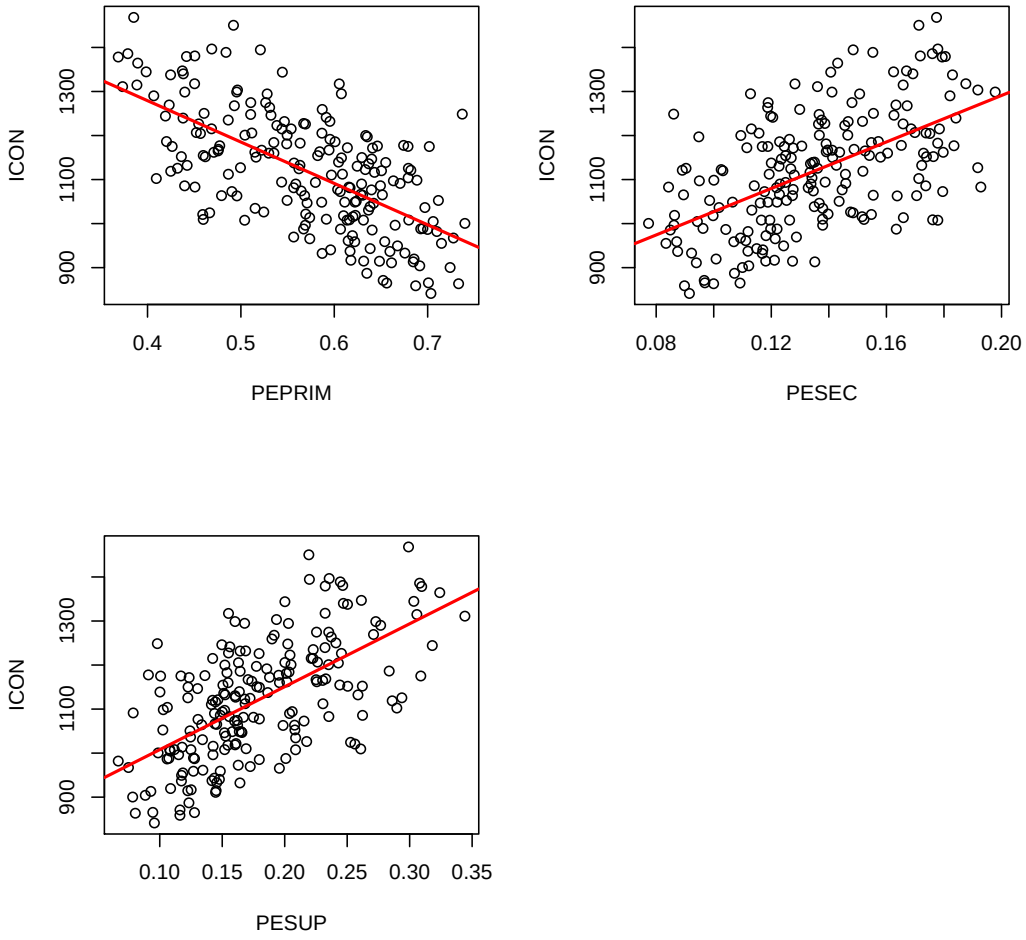

Figura 3.26: Diagramas de dispersión de los ingresos por prestaciones contributivas medios contra los distintos niveles de estudios.

Capítulo 4

Ajuste de modelos.

En el capítulo anterior hemos encontrado diversas variables que estaban linealmente relacionadas con los distintos tipos de ingresos descritos en dicho capítulo. Por ello, dichas variables pueden ser adecuadas para explicar los diferentes tipos de ingresos ajustando modelos de Fay-Herriot, que ya vimos en el Capítulo 2.

Concretamente, en este capítulo, aprovechando que disponemos de datos de varios años de cada variable, ajustaremos modelos de Fay-Herriot con efecto temporal para cada tipo de ingreso, ayudándonos del paquete *saery* (*Esteban M. et al (2014)*) del software estadístico R (*R Core Team (2018)*). Dicho paquete nos permite ajustar, además del modelo ya mostrado en el capítulo 2, modelos con un efecto temporales MA(1) de varianza σ^2 y media θ o independientes (que se ajustan de forma análoga al ya visto). Para simplificar, denomiraremos estos modelos como modelo AR, modelo Indep y modelo MA, según la clase de efecto temporal que tengan para cada tipo de ingreso.

4.1. Ingresos totales medios.

En el capítulo anterior vimos que aparentemente las variables P18a65, PM65, PEPRIM, PESUP, PFBLIM y REND parecen relacionadas de forma lineal con la variable ITOT, por lo que serían adecuadas para utilizarlas como covariales en los modelos de Fay-Herriot con efecto temporal. Aún así no incluiremos a todas en los modelos que ajustaremos, puesto que, como podemos ver en la Tabla 3.2, existe una gran correlación entre las variables PEPRIM y PESUP. Esto podría ocasionar problemas en el ajuste de los modelos, por lo que estableceremos como criterio general que si dos variables tienen una correlación en valor absoluto entre ellas superior a 0,9, solo incluiremos una de ellas. En este caso aplicaremos este criterio a las variables P18a65, PM65, PEPRIM y PESUP, escogiendo como covariables a PM65 y PESUP debido a la alta correlación que tienen, respectivamente, con las variables P18a65 y PEPRIM. Como esto también ocurre para las covariables del resto de tipos de ingreso, aplicaremos este criterio en todos los casos.

Como comentábamos al inicio de este capítulo, ajustaremos modelos cuyos efectos temporales para cada área siguen un AR(1), un MA(1) o son independientes. En todos ellos nos encontramos con problemas computacionales por parte del paquete *saery* debidos a la escala de la variable ITOT y sobre todo de las varianzas de sus estimadores directos. Esto es algo que, como veremos a lo largo de este capítulo le sucede a todas las variables de ingresos. En ninguno de los casos funcionan las transformaciones clásicas (como la logarítmica), por lo que haremos será dividir esta variables por cantidades que permitan el ajuste y a la vez sea estable para el cálculo del MSE mediante el bootstrap paramétrico. Siguiendo este razonamiento, realizamos un cambio de unidades para la variable ITOT y la varianza

de su error, lo que en este caso hacemos dividiéndolas por $\sqrt{4000}$ y 4000, respectivamente. Hay otros problemas computacionales con el modelo con el AR(1), puesto que las estimaciones del MSE son muy inestables si consideramos todas las covariables, aunque esto se soluciona si solo consideramos como variables explicativas REND y PFBLIM, por lo que solo consideraremos estas. En el modelo con el MA(1) también nos encontramos con problemas computacionales, pues el algoritmo del Fisher Scoring no es convergente considerando todas las variables. Como además de lo anterior, obtiene resultados muy pobres (en términos de su MSE) si consideramos otras combinaciones de variables, descartaremos los modelos de este tipo para la variable ITOT. El modelo con tiempos independientes no muestra problemas añadidos al cambio de unidades.

Basándonos en lo que hemos comentado anteriormente, en este caso ajustaremos entonces dos modelos, un modelo AR con REND y PFBLIM como covariables y un modelo Indep con las covariables PM65, PESUP, PFBLIM y REND, además de considerar una constante en ambos casos (a la que denotaremos como α). En la Tabla 4.1 podemos ver los resultados de dichos ajustes. Observamos que en el modelo AR todas las variables son claramente significativas, mientras que en el modelo independiente nos encontramos con que las variables PM65 y PESUP no son significativas, lo que puede deberse a que están muy correlacionadas con la variable REND, por lo que los efectos fijos de los dos modelos son muy similares. Si bien podríamos probar a ajustar el modelo sin las variables no significativas, como buscamos comprobar el comportamiento de los criterios de selección, dejaremos que sean estos los que decidan si quitamos o no las variables. Como nuestro objetivo es evitar los cambios bruscos en a evolución de los ingresos, una forma razonable de comparar los modelos es comparar sus coeficientes de variación.

ITOT	α	PM65	PESUP	PFBLIM	REND	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\rho}/\hat{\theta}$
Modelo AR	22,5($<$ 0,001)			$-20{,}79$($<$ 0,001)	$7{,}34e-04$($<$ 0,001)	1,51	0,29	0,59
Modelo Indep	23,76($<$ 0,001)	$-5{,}24$(0,33)	1,07(0,76)	$-20{,}92$($<$ 0,001)	$7{,}24e-04$($<$ 0,001)	1,57	0,30	

Tabla 4.1: Coeficientes de los modelos ajustados para la variable ITOT (transformada), con sus respectivos niveles de significación (entre paréntesis).

En la Figura 4.1 podemos ver esto último, entre los estimadores directos, el modelo Indep y el modelo AR, donde observamos que los modelos reducen en gran medida la variación de los estimadores directos, siendo similar en los dos modelos planteados. Esto puede verse también en la Figura 4.2, donde podemos ver que utilizando el modelo AR, conseguimos que las estimaciones no cambien de forma tan brusca de un año para otro, obteniendo resultados similares con el modelo Indep.

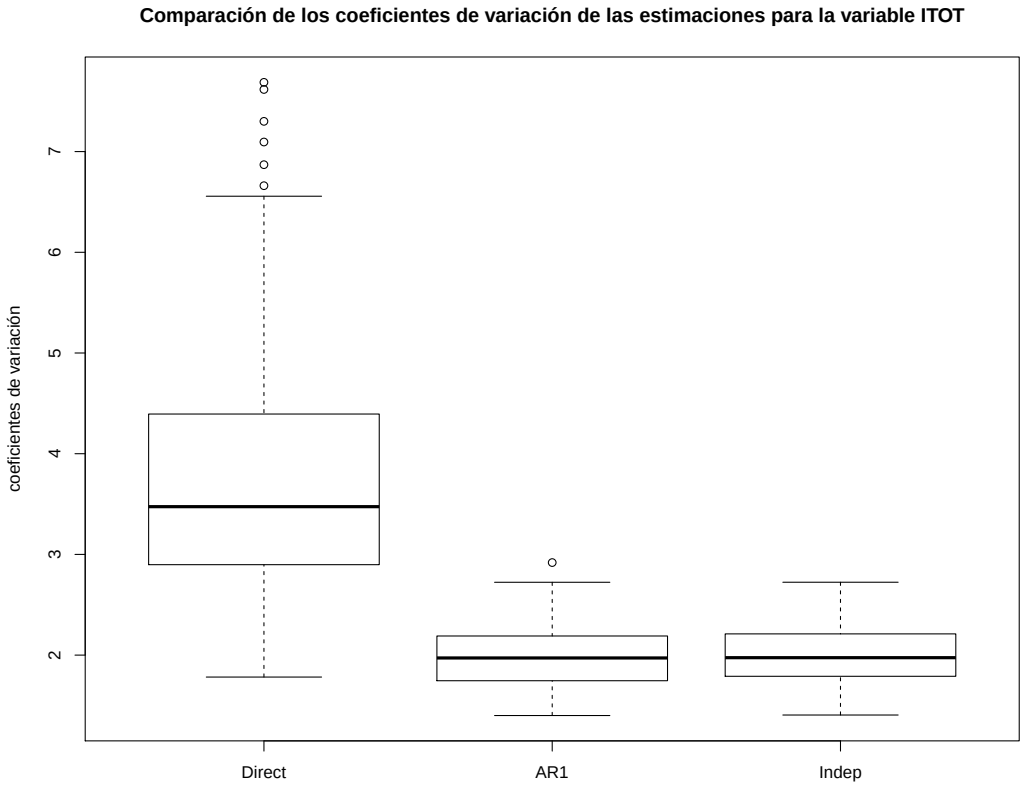

Figura 4.1: Boxplot de los coeficientes de variación del modelo AR, modelo Indep y estimadores directos para la variable ITOT, donde el color indica la provincia y el número el área de dicha provincia.

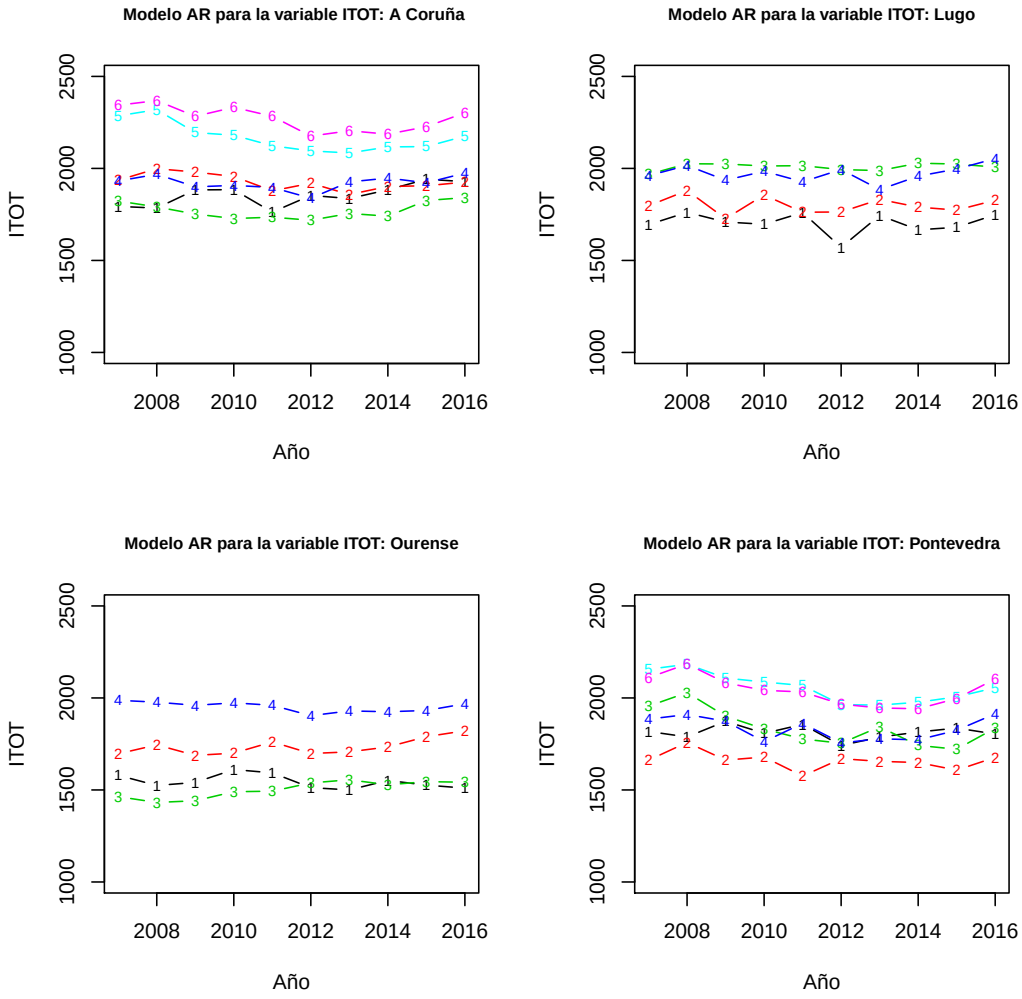

Figura 4.2: Evolución de los ingresos medios totales, estimados con el modelo AR ajustado para la variable ITOT. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

4.2. Ingresos por cuenta ajena medios.

Al igual que hicimos para la variable ITOT en la sección anterior, haremos algo análogo para la variable IAJ. Esto quiere decir que omitiremos las variables PEPRIM y P18a65 y trataremos de ajustar modelos con los distintos tipos de efecto temporal utilizando el resto de variables vistas en el análisis exploratorio.

Al realizar los ajustes nos encontramos con problemas computacionales debido a la escala en la que se encuentran, por lo que para evitarlo deberemos dividir, como ya hicimos de forma similar para la variable ITOT, la variable IAJ entre $\sqrt{3500}$ y la varianza de sus errores entre 3500. El ajuste tiene además otras clases de problemas computacionales, similares a los vistos para la variable ITOT, puesto que la variable correspondiente a la constante impide realizar el ajuste para el modelo AR, por lo que no la consideraremos en dicho ajuste. Los modelos Indep y MA no muestran problemas añadidos.

IAJ	α	PM65	PESUP	REND	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\rho}/\hat{\theta}$
Modelo AR		25,70($<$ 0,001)	11,50(0,03)	$1{,}3e-03$($<$ 0,001)	1,43	0,88	0,55
Modelo Indep	14,48($<$ 0,001)	$-4{,}59$(0,4)	6,96(0,46)	$9{,}3e-04$($<$ 0,001)	1,36	0,62	
Modelo MA	14,28($<$ 0,001)	$-3{,}78$(0,55)	8,38(0,07)	$9{,}1e-04$($<$ 0,001)	1,25	0,60	$-0{,}36$

Tabla 4.2: Coeficientes de los modelos ajustados para la variable IAJ (transformada), con sus respectivos niveles de significación (entre paréntesis)

En la Tabla 4.2 podemos ver los resultados. En ella podemos observar que, salvo para el modelo AR, las variables PM65 y PESUP no son significativas, algo nuevamente esperado puesto que están altamente correlacionadas con la variable REND. Nuevamente, dejaremos que los criterios de selección decidan si mantenemos o no estas variables (lo veremos en el Capítulo 5).

En la Figura 4.3 podemos ver una comparación de los coeficientes de variación entre los modelos y los estimadores directos, observando que, si bien todos parecen mejorar mucho los estimadores directos, los modelos Indep y MA parecen mostrar un mejor resultado que el modelo AR. Un ejemplo de esta mejoría se puede ver en la Figura 4.4, donde podemos ver que el modelo Indep es capaz de eliminar los cambios bruscos que podíamos encontrar en los datos originales.

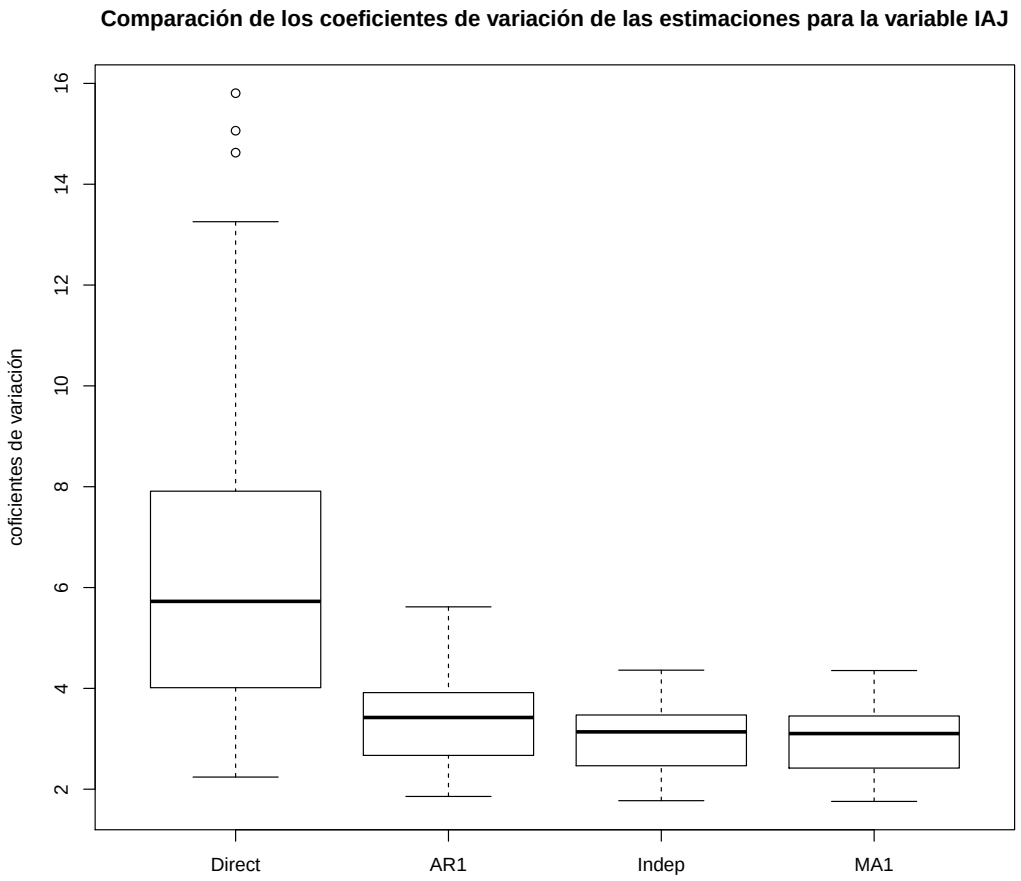

Figura 4.3: Boxplots de los coeficientes de variación de los estimadores directos y los modelos ajustados para la variable IAJ.

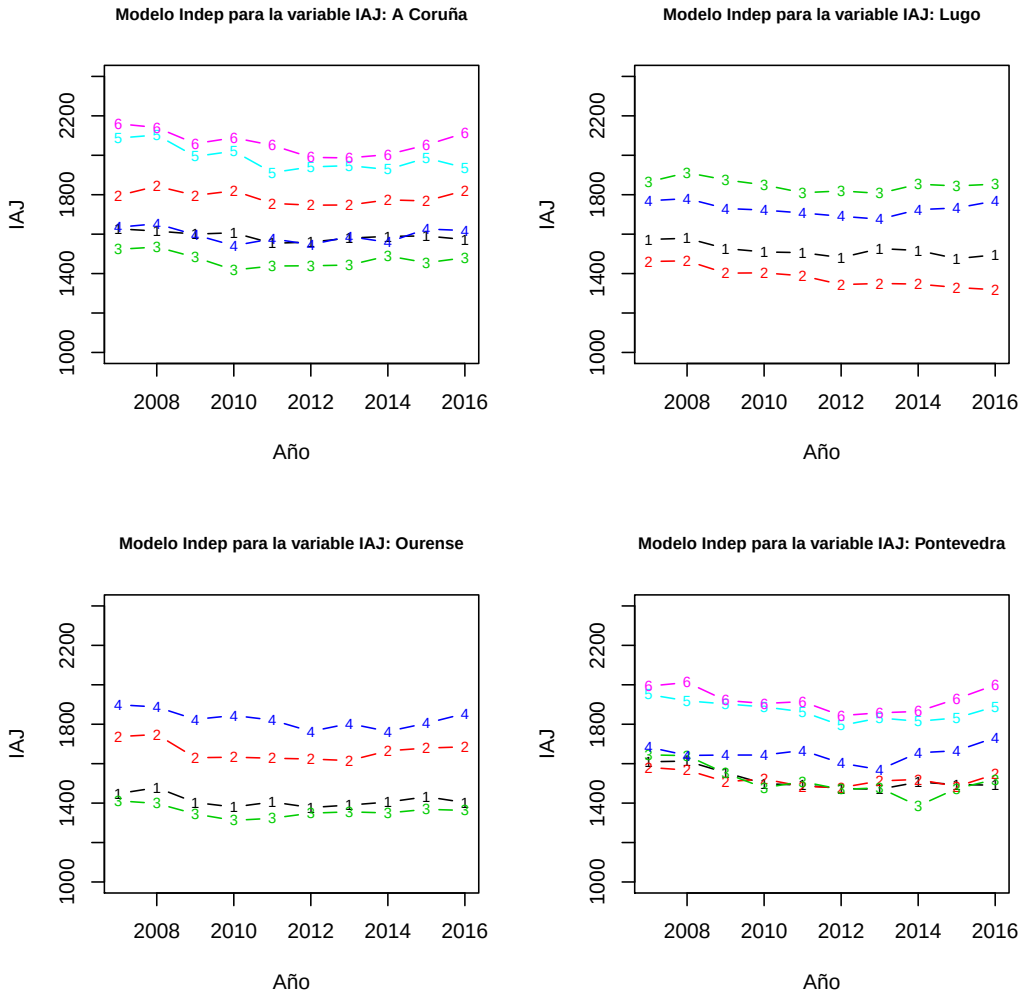

Figura 4.4: Estimación de los ingresos por cuenta ajena medios utilizando el modelo Indep ajustado para la variable IAJ. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

4.3. Ingresos por cuenta propia medios.

Siguiendo la metodología que hemos utilizado para ITOT e IAJ, por lo que ajustaremos modelos para la variable IPROP utilizando las variables PM65, REND y PFTIPO3 como covariables.

Al igual que para IAJ e ITOT, deberemos cambiar las unidades de la variable IPROP, dividiendo entre $\sqrt{3500}$ y entre 3500 la varianza de sus errores. En este caso no aparecen problemas de ningún tipo (además del ya mencionado) en los ajustes, pudiendo ver sus resultados en la Tabla 4.3.

IPROP	α	PM65	REND	PFTIPO3	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\rho}/\hat{\theta}$
Modelo AR	$-5{,}6(0{,}23)$	$21{,}82(0{,}01)$	$6{,}9e-04(<0{,}001)$	$26{,}52(<0{,}001)$	1,39	2,31	0,13
Modelo Indep	$-6{,}43(0{,}21)$	$22{,}75(0{,}01)$	$7{,}1e-04(<0{,}001)$	$27{,}11(<0{,}001)$	1,49	2,10	
Modelo MA	$-4{,}12(0{,}36)$	$19{,}86(0{,}01)$	$6{,}5e-04(<0{,}001)$	$25{,}10(<0{,}001)$	1,43	2,35	$-0{,}45$

Tabla 4.3: Coeficientes de los modelos ajustados para la variable IPROP (transformada), con sus respectivos niveles de significación (entre paréntesis)

En dicha tabla observamos que todas las variables son significaivas, salvo la constante. En la Figura 4.5 tenemos una comparación de los coeficientes de variación entre los modelos y los estimadores directos, donde podemos ver que todos los modelos mejoran sustancialmente a los estimadores directos, obteniendo los modelos resultados similares. La mejora que estos modelos ofrecen puede verse la Figura 4.6, donde podemos ver que el modelo MA mejora notablemente la estimación dada por los estimadores directos, aunque aún mantienen una cierta inestabilidad (también debida a que este tipo de ingreso puede ser muy inestable).

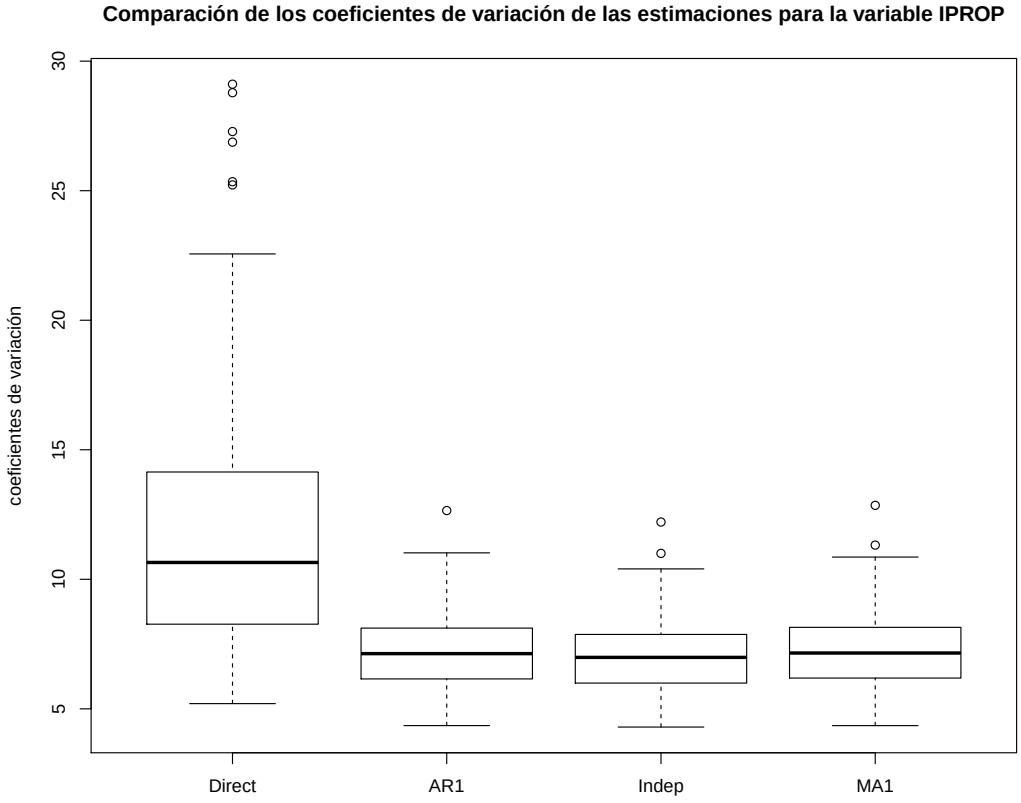

Figura 4.5: Boxplots de los coeficientes de variación de las estimaciones realizadas con los estimadores directos y con los modelos ajustados para la variable IPROP.

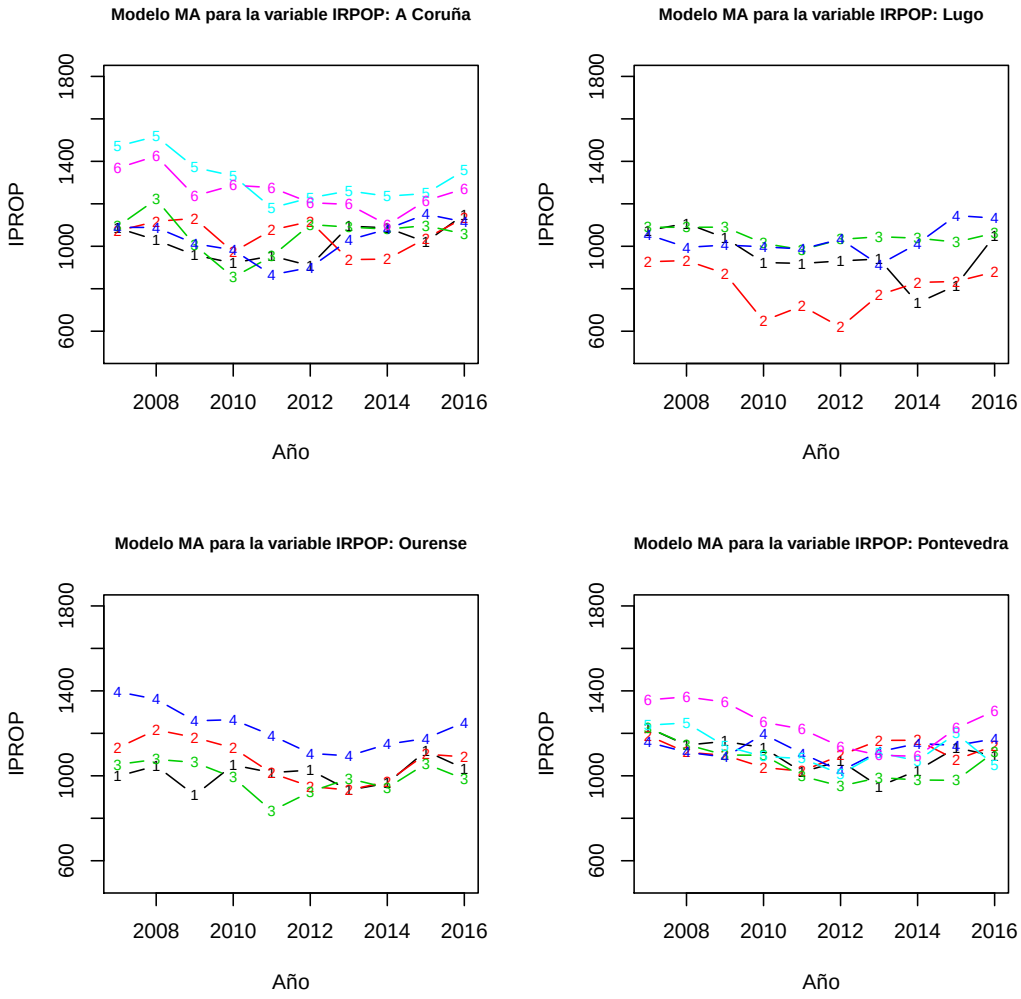

Figura 4.6: Estimación de la evolución de los ingresos por cuenta propia medios, realizada por el modelo MA ajustado para la variable IPROP. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

4.4. Ingresos por prestaciones contributivas medios.

Al igual que para los ingresos vistos en las secciones anteriores de este mismo capítulo, deberemos cambiar las unidades de la variable ICON, dividiendo entre $\sqrt{3500}$ y entre 3500 la varianza de sus errores. En este caso vuelve a suceder lo mismo que le ocurría a la variable IAJ, teniendo que suprimir la variable PEPRIM de todos los modelos por estar muy correlacionada con PESUP, y la constante de los modelos AR y MA por problemas computacionales. Los resultados de los ajustes pueden verse en la tabla 4.4, donde observamos que para los modelos AR y MA la variable PESUP no es significativa mientras PENMEAN y PESEC son significativas al 5 %. Esto no ocurre con el modelo Indep, pues en dicho modelo tanto la variable PESUP como la variable PESEC no son significativas.

ICON	α	PENMEAN	PESUP	PESEC	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\rho}/\hat{\theta}$
Modelo AR		0,02($<$ 0,001)	6,08(0,08)	9,90(0,008)	1,74	0,18	0,74
Modelo Indep	1,74(0,027)	0,02($<$ 0,001)	5,61(0,11)	5,79(0,14)	1,37	0,24	
Modelo MA		0,02($<$ 0,001)	6,57(0,066)	8,89(0,018)	1,88	0,18	$-$0,70

Tabla 4.4: Coeficientes de los modelos ajustados para la variable ICON (transformada), con sus respectivos nivel de significación (entre paréntesis)

En la Figura 4.7 podemos ver una comparación de los coeficientes de variación obtenidos por los modelos, con los obtenidos de los estimadores directos, observando que los modelos mejoran notablemente a los estimadores directos, aunque aparecen algunos inusualmente grandes, especialmente en el modelo AR, siendo los resultados similares para todos los modelos.

En la Figura 4.8 podemos ver la evolución de los ingresos medios por prestaciones contributivas, estimados por el modelo Indep. En ella podemos ver que en efecto muestra cierta mejoría respecto a los datos originales.

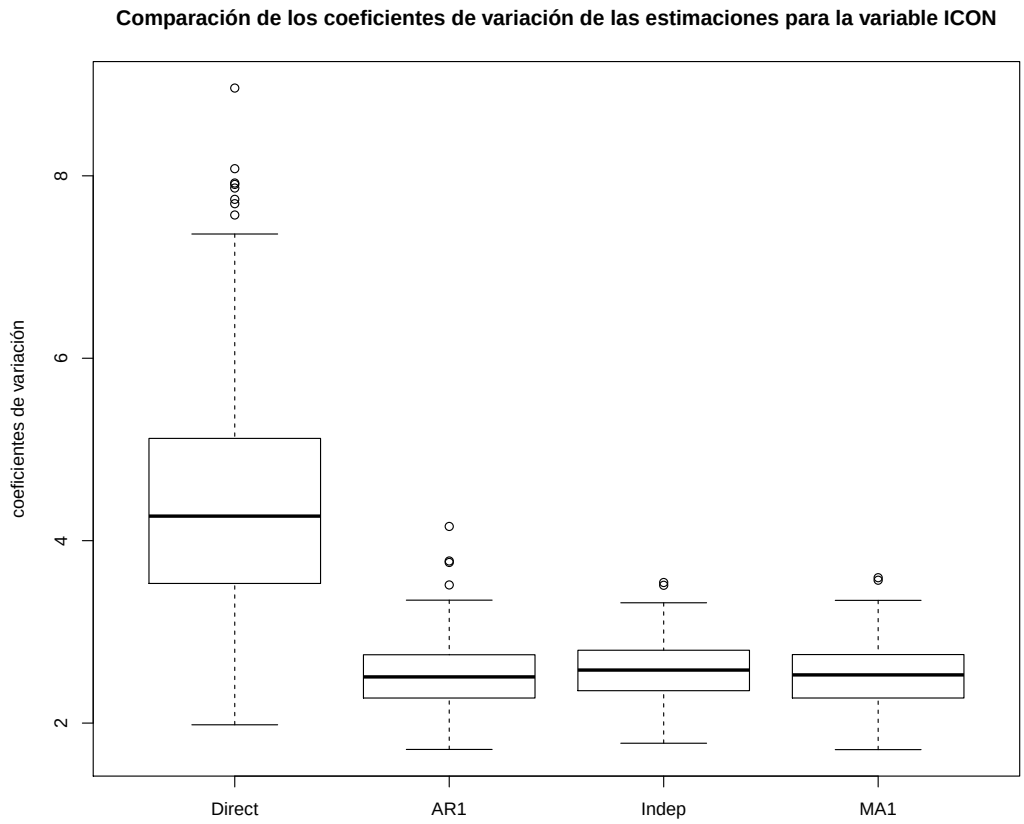

Figura 4.7: Boxplots de los coeficientes de variación de las estimaciones realizadas con los estimadores directos y con los modelos ajustados para la variable ICON.

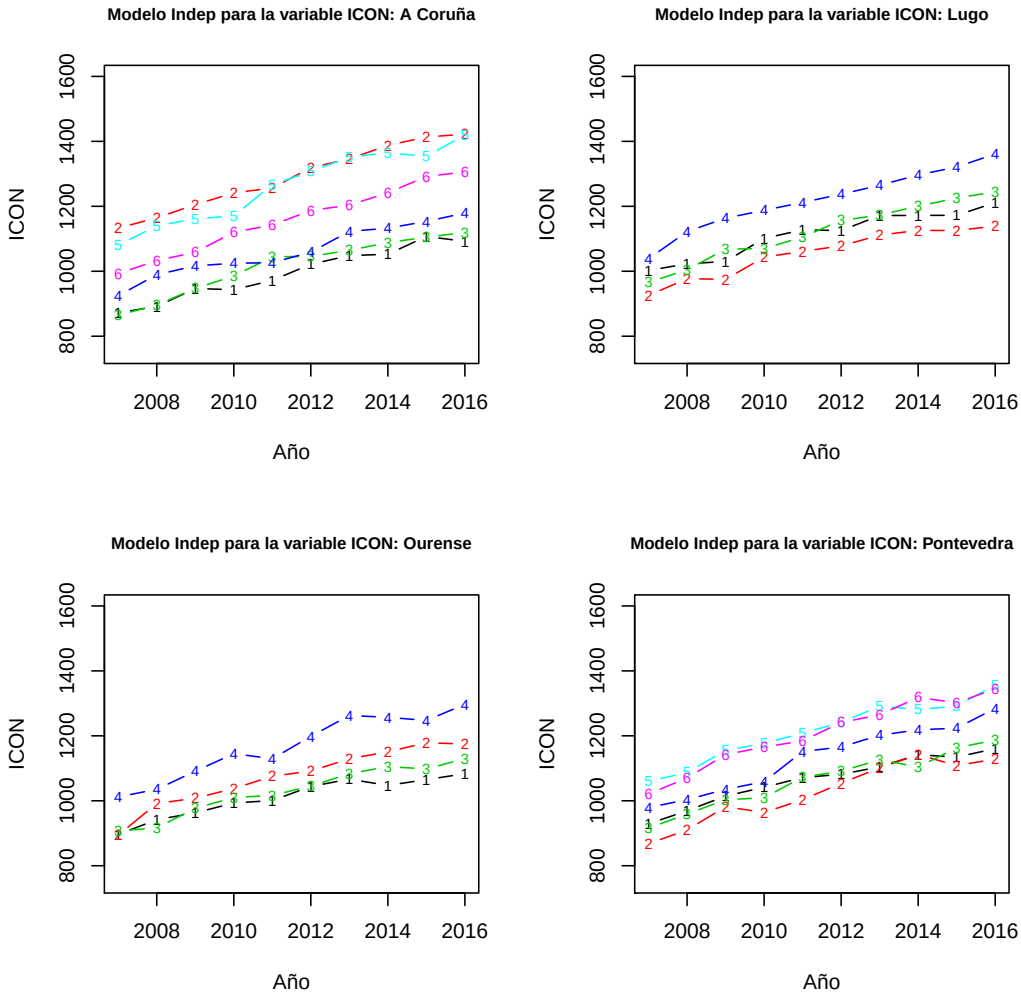

Figura 4.8: Estimación de la evolución de los ingresos por prestaciones contributivas, realizada por el modelo Indep ajustado para la variable ICON. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

Capítulo 5

Selección de modelos.

En el Capítulo 4 hemos ajustado diversos modelos para los distintos tipos de ingresos, obteniendo estimaciones mejores que las dadas por los estimadores directos. Dichos modelos sin embargo no distinguimos claramente cual es el mejor para cada tipo de ingreso e incluso no tienen que ser necesariamente los mejores. Esto puede ser debido a que, como vimos en el Capítulo 4, algunas variables no eran significativas, por lo que podría mejorarse el modelo al no considerarlas en el ajuste.

Para seleccionar el mejor modelo para cada tipo de ingreso lo que haremos será utilizar el criterio AIC (descrito en el Capítulo 2), considerando las diferentes combinaciones de variables de cada uno de los modelos vistos en el capítulo anterior. En esto exceptuamos la constante, que se mantendrá igual que en dichos modelos, es decir, para un modelo con constante se considerarán todas las combinaciones de covariables que incluyan la misma. A dichos modelos les calcularemos su cAIC (tanto la versión de *Han (2013)* (que denotaremos por cAICH) como la de *Vaida y Blanchard (2005)* (que denotaremos por cAICVyB), el xGAIC y la versión de *You et al (2016)* (que denotaremos por cYMO).

Al realizar el procedimiento encontramos que el xGAIC suele escoger modelos razonables, en los cuales aparecen las variables más importantes de los modelos que vimos en el Capítulo 4 y que si bien no son los que se escogen utilizando el cAIC (tanto el de cAICH como el cAICVyB, que suelen escoger el mismo modelo), los resultados tanto en las variables que se seleccionan como en la propia escala de los dos criterios, son similares. El cYMO suele en cambio dar resultados muchos peores, seleccionando modelos en los que se pierden algunas de las variables más importantes.

Nuevamente existieron múltiples problemas debidos a la inestabilidad de la estimación del *Fisher Scoring*, lo que provoca que la estimación de los GDF del xGAIC sea incluso absurda, obteniendo resultados gigantescos o incluso negativos. Para solucionar esto se optó por controlar las remuestras, de forma que, en caso de que el Fisher Scoring no sea convergente para una remuestra, repitamos la misma hasta que obtengamos una sí convergente, descartando directamente aquellos que requieran 1000 réplicas más de las teóricas, es decir, que necesite más de 1500 remuestras para obtener 500 réplicas válidas.

Para simplificar la notación denominaremos a los modelos según el criterio que los selecciona, por ejemplo, nombraremos como "modelo xGAIC"al modelo seleccionado por el criterio xGAIC. Haremos una excepción con los criterios cAICVyB y cAICH, puesto que, salvo para la variable ICON, coinciden en la elección, por lo que, salvo en dicho caso, los modelos que seleccionen dichos modelos los denominaremos por "modelo cAIC". Esto lo realizaremos para cada uno de los tipos de ingreso.

5.1. Ingresos totales medios.

En la Tabla 5.1 podemos ver los resultados del procedimiento descrito anteriormente para la variable ITOT. En ella podemos ver que las dos versiones del cAIC escogen el mismo modelo (que coincide con el modelo AR ajustado en la Sección 4.1) mientras que tanto el xGAIC (que coincide con el modelo Indep ajustado en la Sección 4.1) como el cYMO escogen otros modelos distintos. Observamos que si bien no coinciden en la clase de efecto temporal, si lo hacen en las variables escogidas, de hecho el modelo escogido por el cAIC tiene un xGAIC de 648,79, un valor muy cercano al que tiene el modelo seleccionado por el criterio xGAIC. Esto no ocurre con el cYMO, que nos lleva a elegir un modelo que prescinde de la variable REND, que es una de las variables más importantes para los demás criterios, sumándose además el hecho de que el ajuste del modelo ni siquiera es convergente, obteniendo una estimación de σ_1^2 con un valor excesivo.

ITOT							AIC			
MODELO	EFECTO T.	α	PFBLIM	REND	PESUP	PM65	cAICVyB	cAICH	xGAIC	cYMO
Modelo cAIC	AR(1)	✓	✓	✓			**632.83**	**634.00**	653.20	572.90
Modelo xGAIC	Indep	✓	✓	✓	✓	✓	641.09	642.15	**648.79**	576.41
Modelo cYMO	AR(1)	✓	✓				648.19	651.38	798.94	**553.57**

Tabla 5.1: Modelos para la variable ITOT seleccionados por los distintos criterios (indicando en su nombre el criterio que lo seleccionó).

Tenemos entonces entre dos modelos en principio similares en cuanto resultados (como ya vimos en el capítulo anterior), sin embargo nos decidiremos por el modelol xGAIC, puesto que se mostró muy estable en las remuestras (sin necesidad de repetir ninguna), a diferencia del escogido por las dos versiones del cAIC que necesitó repetir más de 100 remuestras. En la Figura 5.1 podemos ver la estimación de la evolución de los ingresos totales medios.

En la Figura 5.2 podemos ver los coeficientes de variación en cada área de las estimaciones basadas en el modelo xGAIC, en los años 2007 y 2016. Si la comparamos con lo visto en la Sección 3.1, observamos que las estimaciones basadas en el modelo mejoran en gran medida las realizadas por los estimadores directos en todas las áreas.

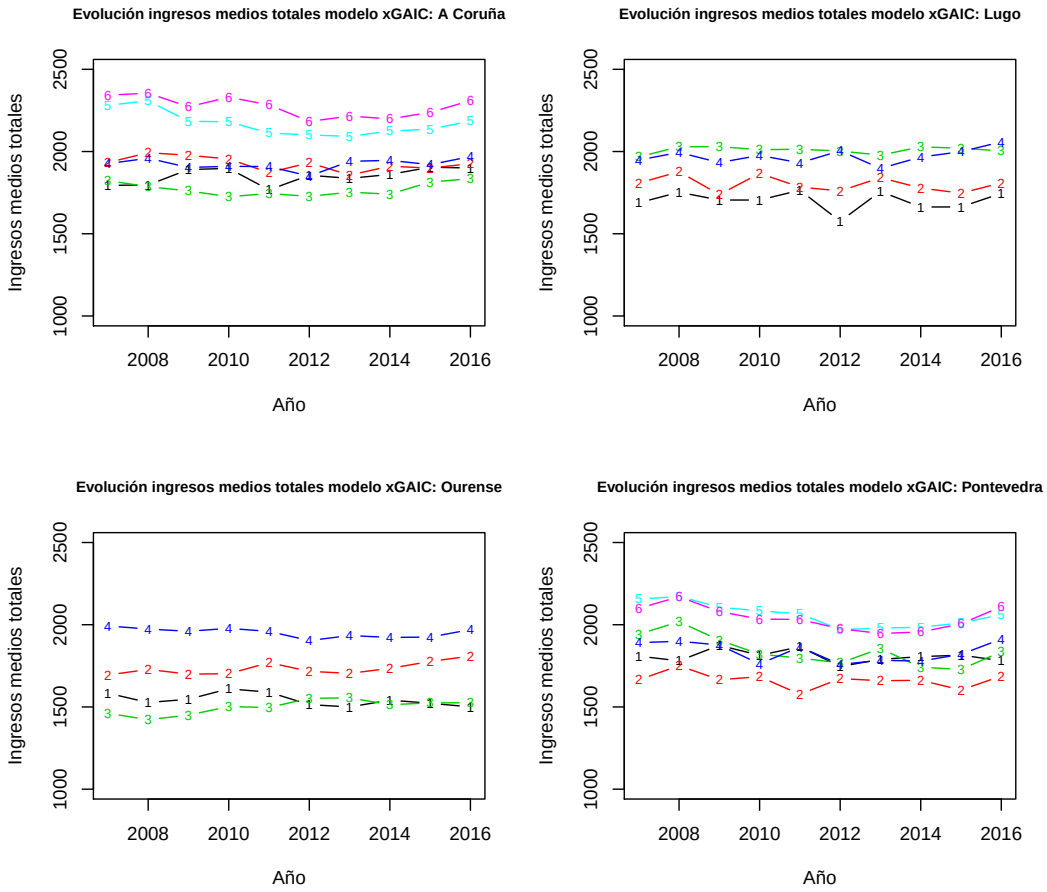

Figura 5.1: Evolución de los ingresos medios totales, estimados con el modelo seleccionado por el xGAIC para la variable ITOT. Se muestran las estimaciones según la provincia y su área (indicada por su númeración dentro de su correspondiente provincia).

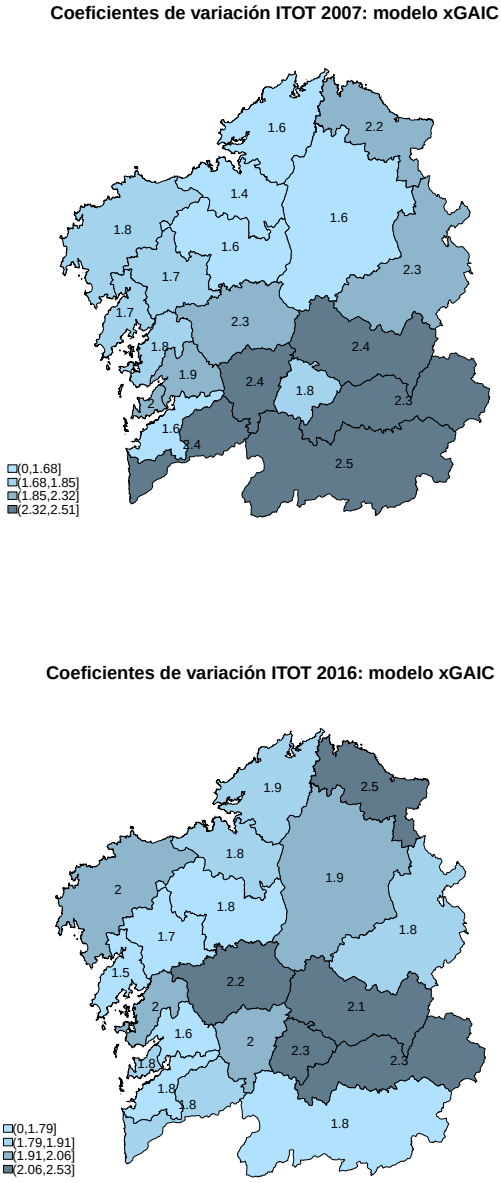

Figura 5.2: Mapas de los coeficientes de variación de las estimaciones basadas en el modelo seleccionado por el xGAIC, correspondientes a la variable ITOT por área en 2007 y 2016.

5.2. Ingresos por cuenta ajena medios.

IAJ						AIC			
MODELO	EFECTO T.	α	PM65	REND	PESUP	cAICVyB	cAICH	xGAIC	cYMO
Modelo cAIC	MA(1)	✓	✓	✓		**792.55**	**793.48**	800.94	740.99
Modelo xGAIC	Indep	✓		✓	✓	795.01	796.00	**789.35**	734.14
Modelo cYMO	AR(1)		✓	✓		805.74	806.562	811.31	**730.87**

Tabla 5.2: Modelos para la variable IAJ seleccionados por los distintos criterios (indicando en su nombre el criterio que lo seleccionó).

En la Tabla 5.2 podemos ver los modelos seleccionados por los criterios y el valor de los mismos para cada uno. En ella observamos que, al igual que sucedió con la variable ITOT, los distintos criterios de selección escogen modelos diferentes , si bien a diferencia de lo que vimos en la sección anterior en este caso todos los modelos seleccionados parecen razonables en cuanto a las variables explicativas de los mismos, destacando entre ellas la variable REND, presente en todos los modelos.

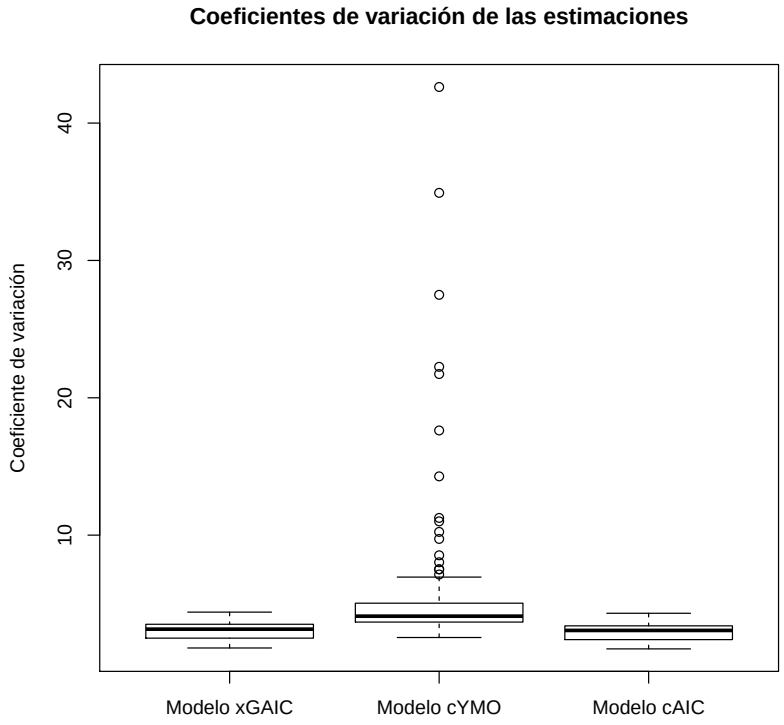

Figura 5.3: Boxplots de los coeficientes de variación de los modelos seleccionados por los distintos criterios para la variable IAJ.

En la Figura 5.3 podemos ver una comparación de los boxplots correspondientes a los coeficientes

de variación de las estimaciones realizadas por los modelos, en la que observamos que los resultados son muy similares para los modelos seleccionados por el xGAIC y el cAIC, existiendo una diferencia muy ligera entre ellos, por lo que sería razonable utilizar cualquiera de los dos modelos, mientras que el seleccionado por el cYMO se muestra muy inestable, puesto que tiene una gran cantidad de estimaciones atípicas, por lo que en principio lo descartaremos. En la Figura 5.4 y la Figura 5.5, podemos ver las estimaciones realizadas por los modelos xGAIC y cAIC respectivamente, observando que ambas son similares. Cabe destacar (como se puede ver en el Apéndice B) que los dos modelos tienen valores similares bajo el cAIC, mientras que se aprecia una mayor diferencia bajo el xGAIC, lo que puede ser provocado por la inestabilidad del método de ajuste, pues nuevamente se necesitaron repetir 150 remuestras para poder realizar el bootstrap con 500 réplicas.

IAJ	EFECTO T.	α	PM65	PESUP	REND	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\theta}$
Modelo xGAIC	Indep	12,59(< 0,001)		7,38(0,1)	$9{,}79e-04$(< 0,001)	1,16	0,65	
Modelo cAIC	MA(1)	15,30(< 0,001)	−5,38(0,42)		$9{,}74e-04$(< 0,001)	1,68	0,60	−0,18

Tabla 5.3: Coeficientes de los modelos seleccionados por el xGAIC y los criterios cAICVyB y cAICH para la variable IAJ (transformada), con sus respectivos niveles de significación (entre paréntesis)

En la Tabla 5.3 podemos ver los coeficientes de los modelos seleccionados y su significación. En ella observamos que los valores y niveles de significación de los coeficientes son similares a los vistos para los modelos ajustados en el Capítulo 4.

En la Figura 5.6 podemos ver los coeficientes de variación en cada área de las estimaciones basadas en el modelo xGAIC, en los años 2007 y 2016. Al igual que con la variable ITOT, si la comparamos con lo visto en la Sección 3.2, observamos que las estimaciones basadas en el modelo mejoran en gran medida las realizadas por los estimadores directos en todas las áreas.

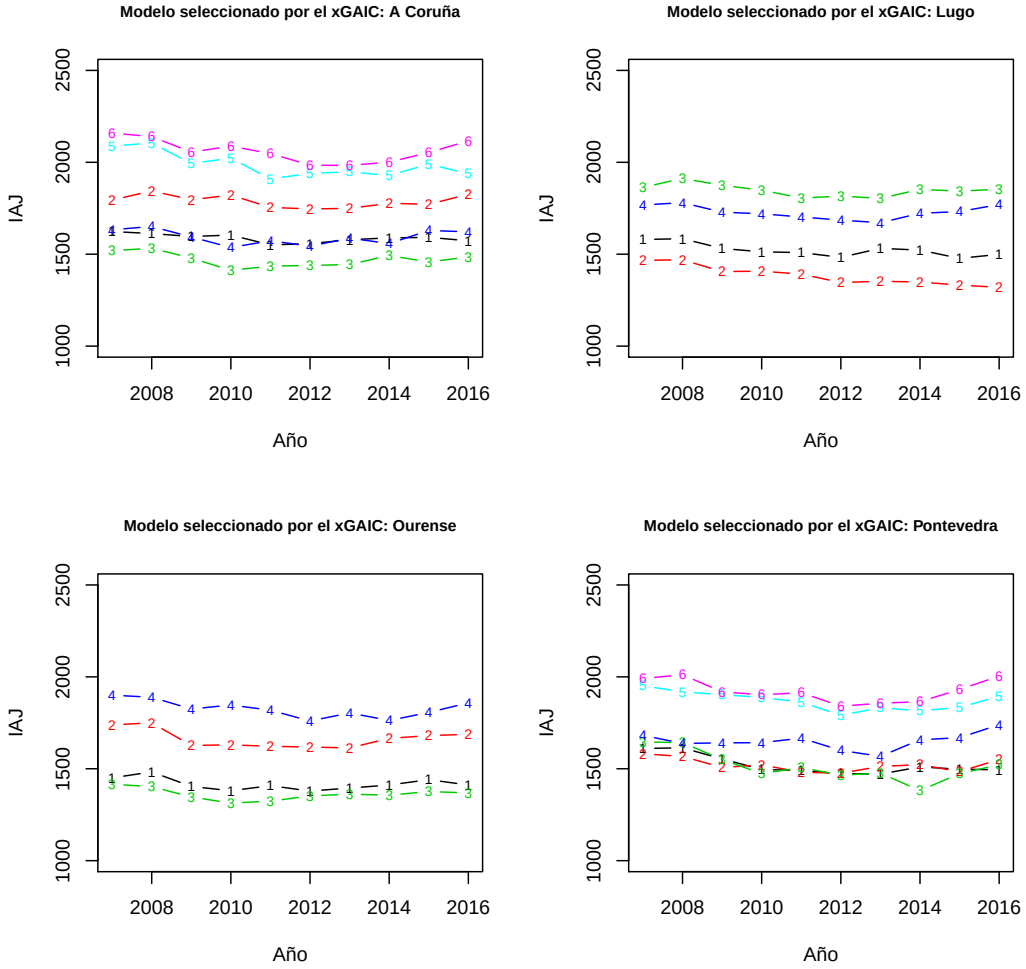

Figura 5.4: Evolución estimada de la variable IAJ realizada por el modelo seleccionado por el xGAIC. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

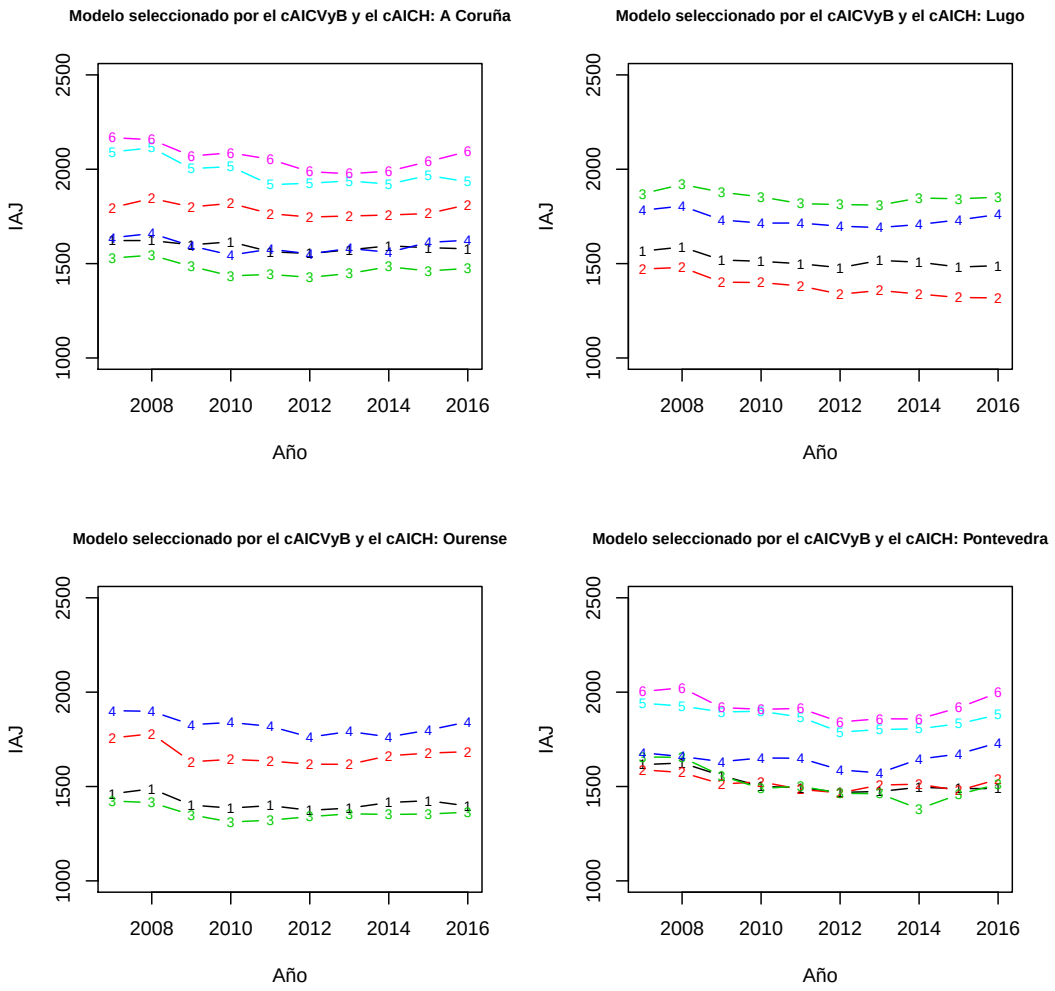

Figura 5.5: Evolución estimada de la variable IAJ realizada por el modelo seleccionado por el cAICVyB y el cAICH. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

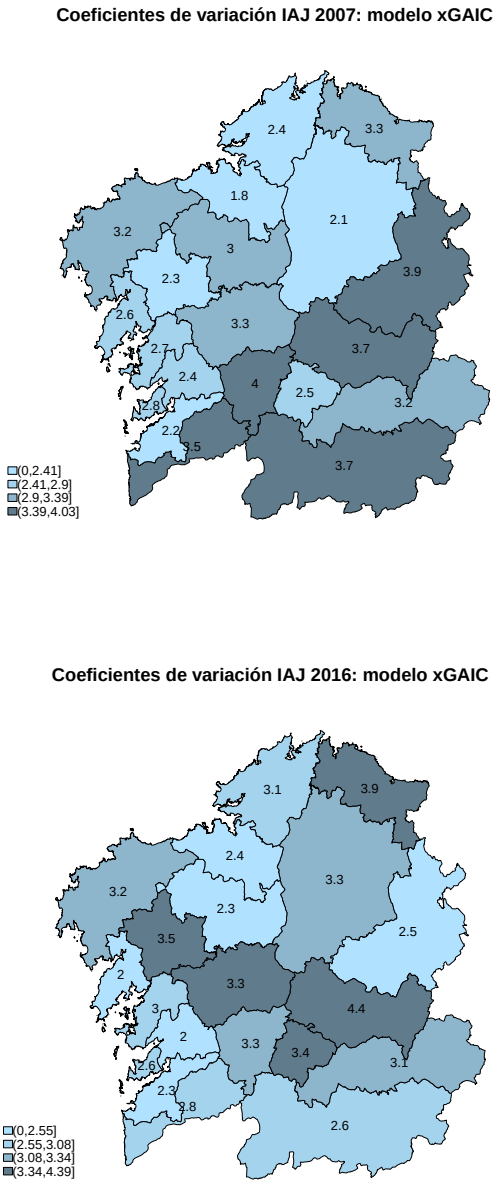

Figura 5.6: Mapas de los coeficientes de variación de las estimaciones basadas en el modelo seleccionado por el xGAIC, correspondientes a la variable IAJ por área en 2007 y 2016.

5.3. Ingresos por cuenta propia medios.

IPROP						AIC			
MODELO	EFECTO T.	α	PM65	REND	PFTIPO3	cAICVyB	cAICH	xGAIC	cYMO
Modelo cAIC	AR(1)	✓	✓	✓	✓	**890.26**	**891.00**	891.65	806.99
Modelo xGAIC	Indep	✓	✓	✓	✓	893.06	893.95	**890.32**	808.27
Modelo cYMO	AR(1)	✓	✓			899.62	**9**00.54	922.72	**800.08**

Tabla 5.4: Modelos para la variable IPROP seleccionados por los distintos criterios (indicando en su nombre el criterio que lo seleccionó).

En la Tabla 5.4 podemos ver los modelos seleccionados por los criterios y el valor de los mismos para cada uno. En ella observamos que, al igual que sucedía con la variable IAJ e ITOT, las versiones del cAIC coinciden en seleccionar el mismo modelo, mientras que los criterios xGAIC y cYMO seleccionan distintos modelos. Tanto bajo el cAIC como el xGAIC, se seleccionan todas las variables, siendo distinto el efecto temporal, mientras el criterio cYMO escoge el modelo que considera únicamente la variable PM65. Esto último es, al igual que sucedía con la variable ITOT, una elección muy mala, puesto que el modelo nuevamente no es ni siquiera convergente, obteniendo estimaciones de las componentes de la varianza malas y en consecuencia, predicciones de baja calidad, por lo que lo descartamos. Este problema se podría arreglar aumentando el peso de los GDF en el criterio, lo que debería resultar en la selección de modelos similares a los dados por los otros criterios.

En la Figura 5.7 podemos ver una comparación del coeficiente de variación de las estimaciones realizadas por los modelos. En ella observamos que la distribución de los coeficientes de variación de los modelos seleccionados por el xGAIC y el seleccionado por el cAICH y el cAICVyB son muy similares.

En la Figura 5.8 y la Figura 5.9 pueden verse los resultados correspondientes a los modelos xGAIC y cAIC respectivamente, observando que en efecto las estimaciones son similares. También observamos que las estimaciones siguen variando mucho entre un año y otro en una misma área, aunque esto puede deberse a que este tipo de ingreso es de naturaleza inestable.

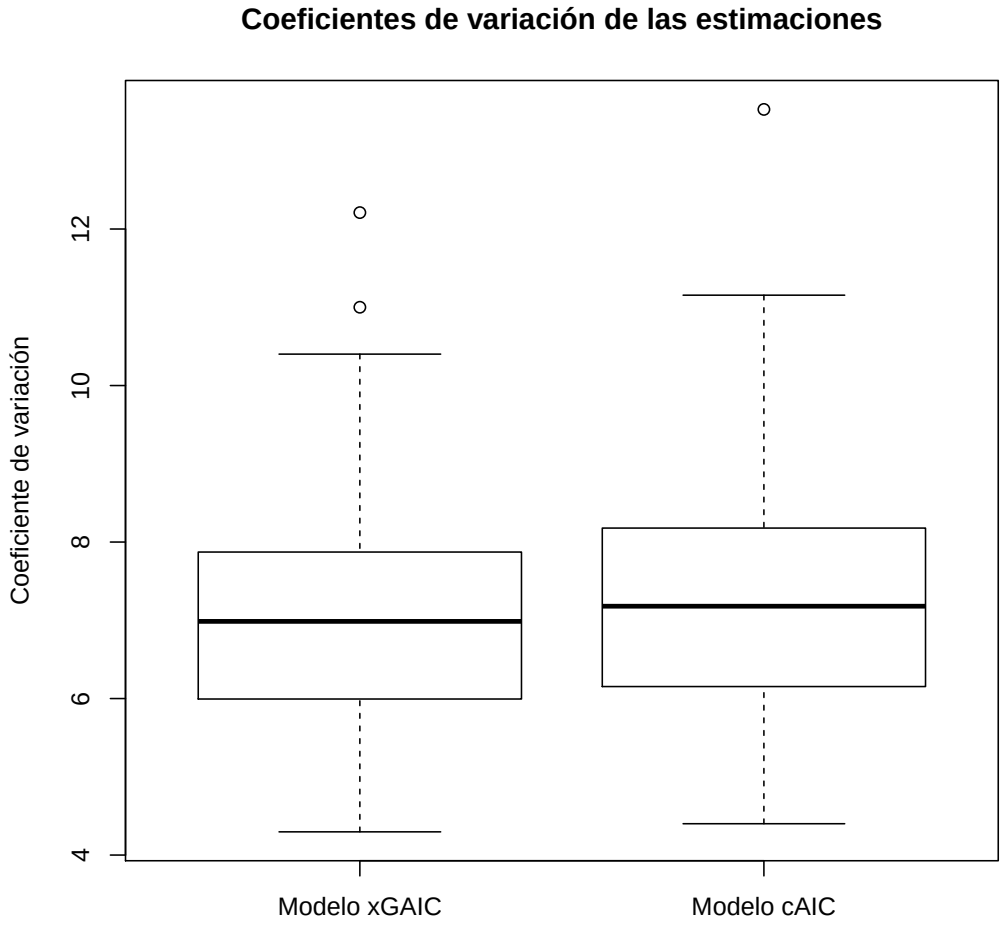

Figura 5.7: Boxplots de los coeficientes de variación de los modelos ajustados para la variable IPROP seleccionados por los distintos criterios.

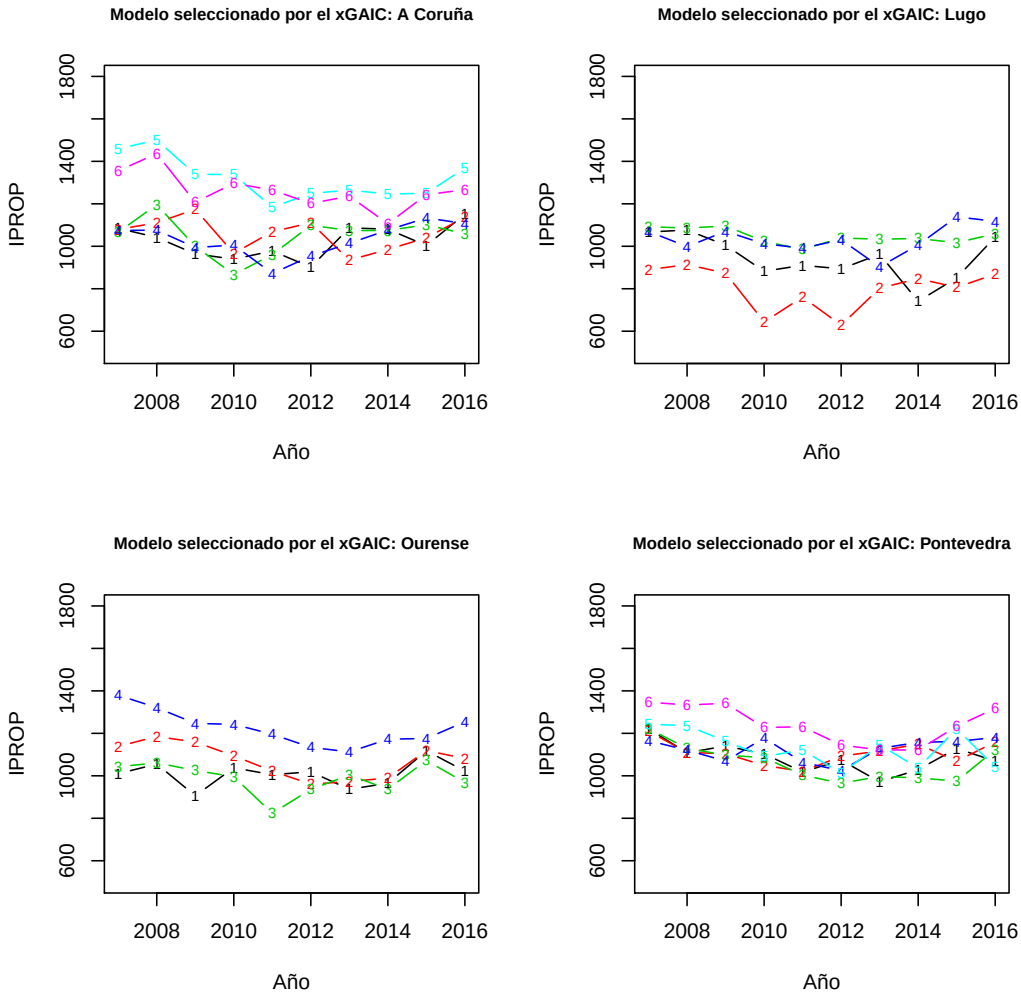

Figura 5.8: Evolución estimada de los ingresos por cuenta ajena medios realizada por el modelo seleccionado por el xGAIC para la variable IPROP. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

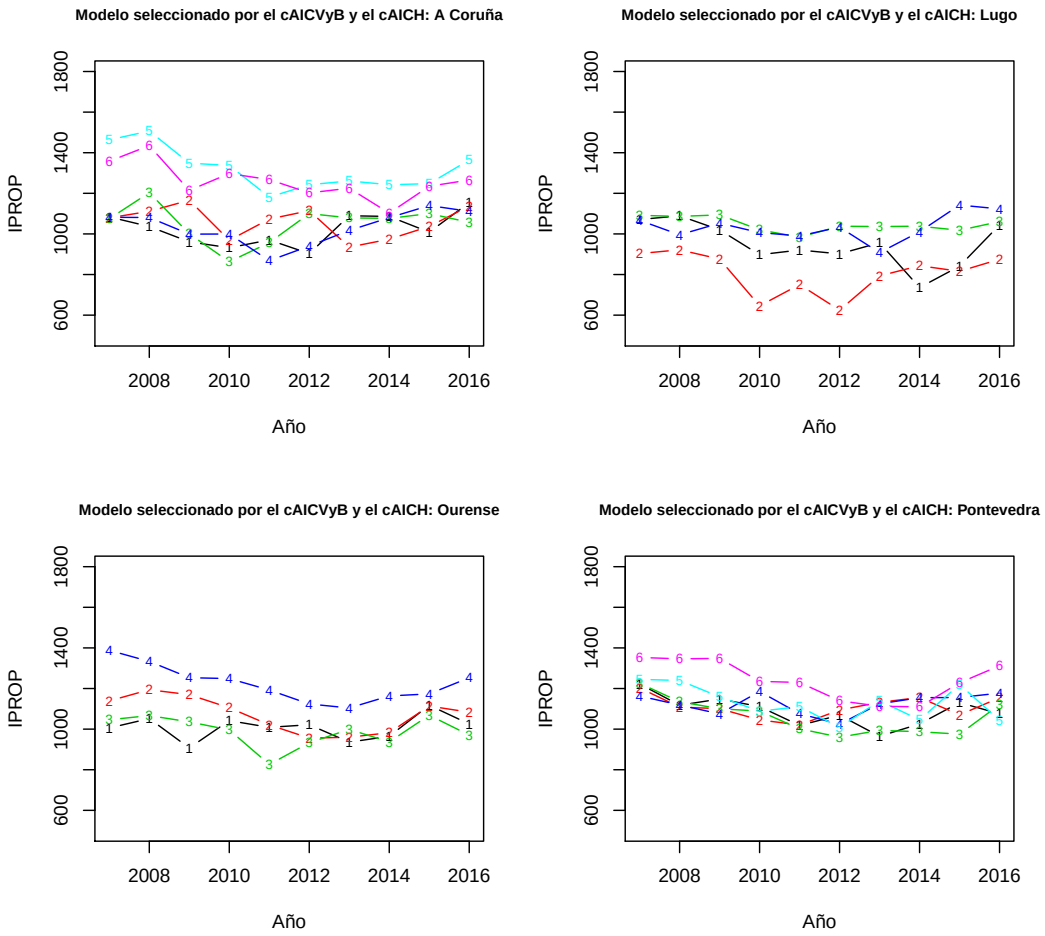

Figura 5.9: Evolución estimada de los ingresos por cuenta ajena medios realizada por el modelo seleccionado cAIC para la variable IPROP. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

En la Figura 5.10 podemos ver los coeficientes de variación en cada área de las estimaciones basadas en el modelo xGAIC, en los años 2007 y 2016. Observamos que en este caso el modelo no mejora todas las estimaciones, puesto que en Lugo-Central y en Lugo Sur (respectivamente, áreas 3 y 1 de la provincia de Lugo) empeora ligeramente. Esto no es un gran problema, puesto que en el resto de áreas mejoran, en algunos casos mucho. Un ejemplo de esto es el área de Barbanza-Noia (área 4 de la provincia de A Coruña), en la que en 2016 pasamos de un coeficiente de variación del 14.8 del estimador directo, a uno de 5.2 del estimador basado en el modelo, por lo que se sigue recomendando el uso del modelo xGAIC para la estimación.

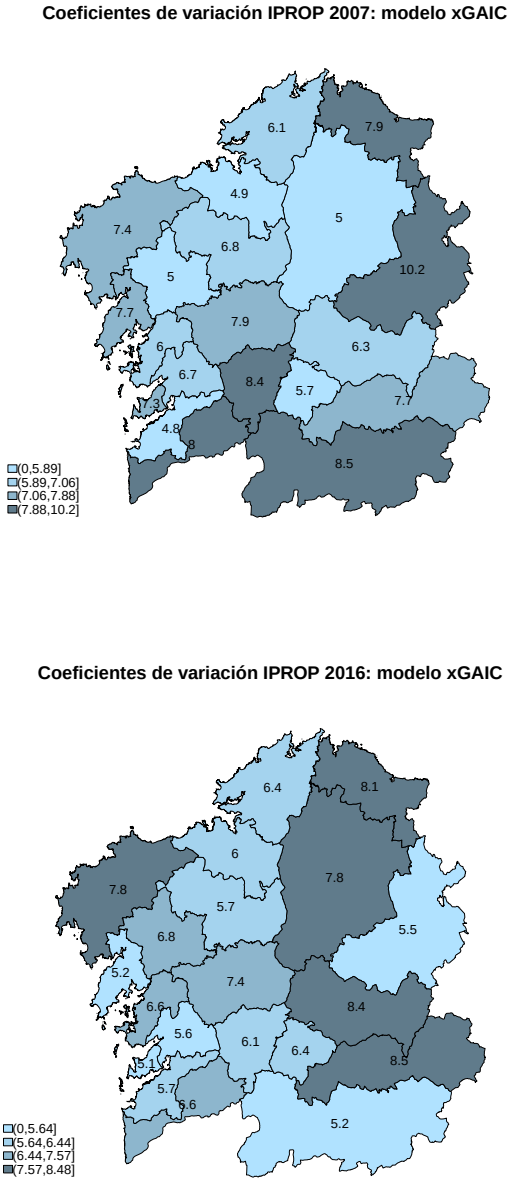

Figura 5.10: Mapas de los coeficientes de variación de las estimaciones basadas en el modelo seleccionado por el xGAIC, correspondientes a la variable ICON por área en 2007 y 2016.

5.4. Ingresos por prestaciones contributivas medios.

ICON						AIC			
MODELO	EFECTO T.	α	PENMEAN	PESUP	PESEC	cAICVyB	cAICH	xGAIC	cYMO
Modelo cAICH	AR(1)		✓	✓	✓	560.80	**561.49**	596.64	496.10
Modelo cAICVyB	AR(1)				✓	**560.52**	561.81	NA	NA
Modelo xGAIC	Indep	✓	✓			575.09	576.16	**588.76**	499.03
Modelo cYMO	AR(1)		✓	✓		565.62	566.51	597.96	**495.63**

Tabla 5.5: Modelos para la variable ICON seleccionados por los distintos criterios (indicando en su nombre el criterio que lo seleccionó).

En la Tabla 5.5 podemos ver los resultados de los criterios. En ella observamos que, por primera vez, el cAICVyB y el cAICH no coinciden en el modelo que seleccionan, teniendo seleccionados en este caso cuatro modelos diferentes. De entre las variables explicativas que puede tener el modelo, la que en principio debería ser más importante es la variable PENMEAN, por ello destaca el modelo seleccionado por el criterio cAICVyB, el cuál no cuenta con dicha variable. Esto nuevamente es una mala selección de modelo, como ya sucedió con otros tipos de ingresos, puesto que el modelo no es convergente, por lo que las estimaciones de las componentes de la varianza vuelven a ser desastrosas incluso absurdas, por lo que descartamos este modelo. El resto de modelos sí parecen razonables en cuanto a las variables que seleccionan, siendo además todos ellos convergentes.

En la Figura 5.11 podemos ver una comparación de los coeficientes de variación de las estimaciones realizadas por los tres modelos seleccionados. En ella observamos que, si bien la media es similar para los tres, el modelo seleccionado por el criterio xGAIC tiene muchos menos valores atípicos que los otros dos modelos, por lo que seleccionaremos dicho modelo. En la Figura 5.12 podemos ver las estimaciones realizadas por el modelo xGAIC.

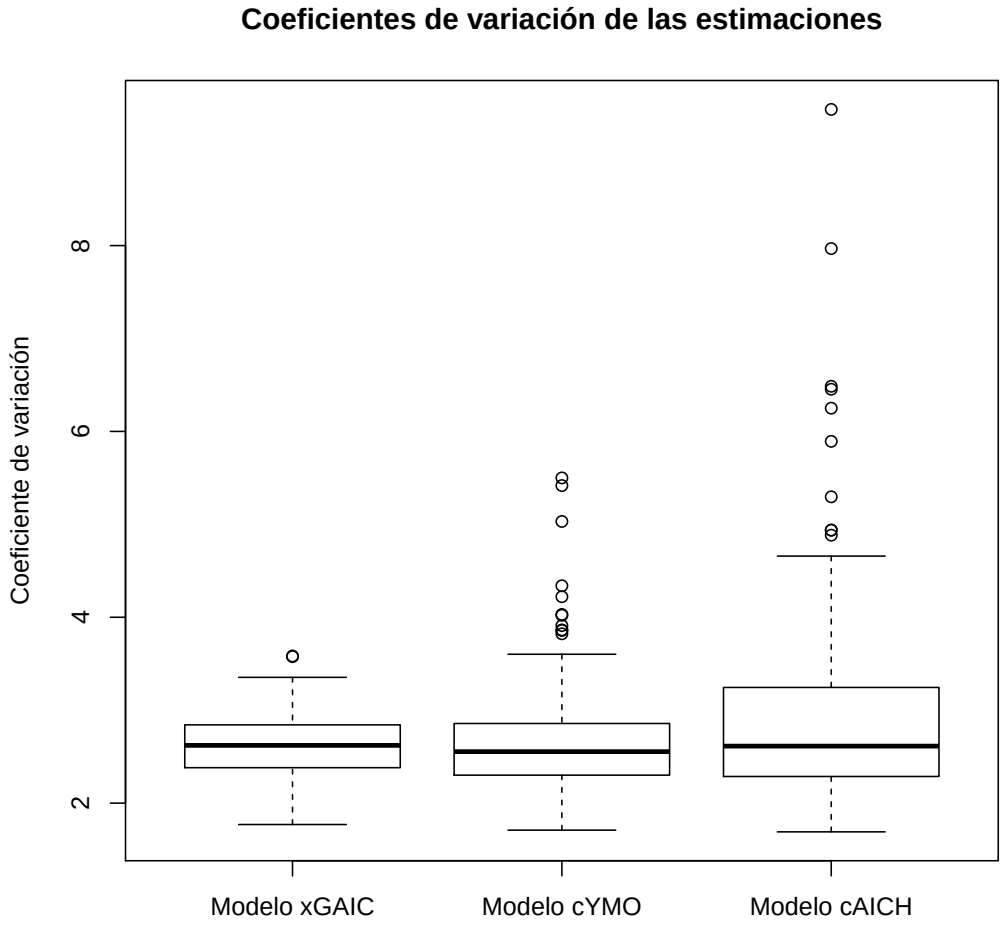

Figura 5.11: Boxplots de los coeficientes de variación de los modelos ajustados para la variable ICON seleccionados por los distintos criterios.

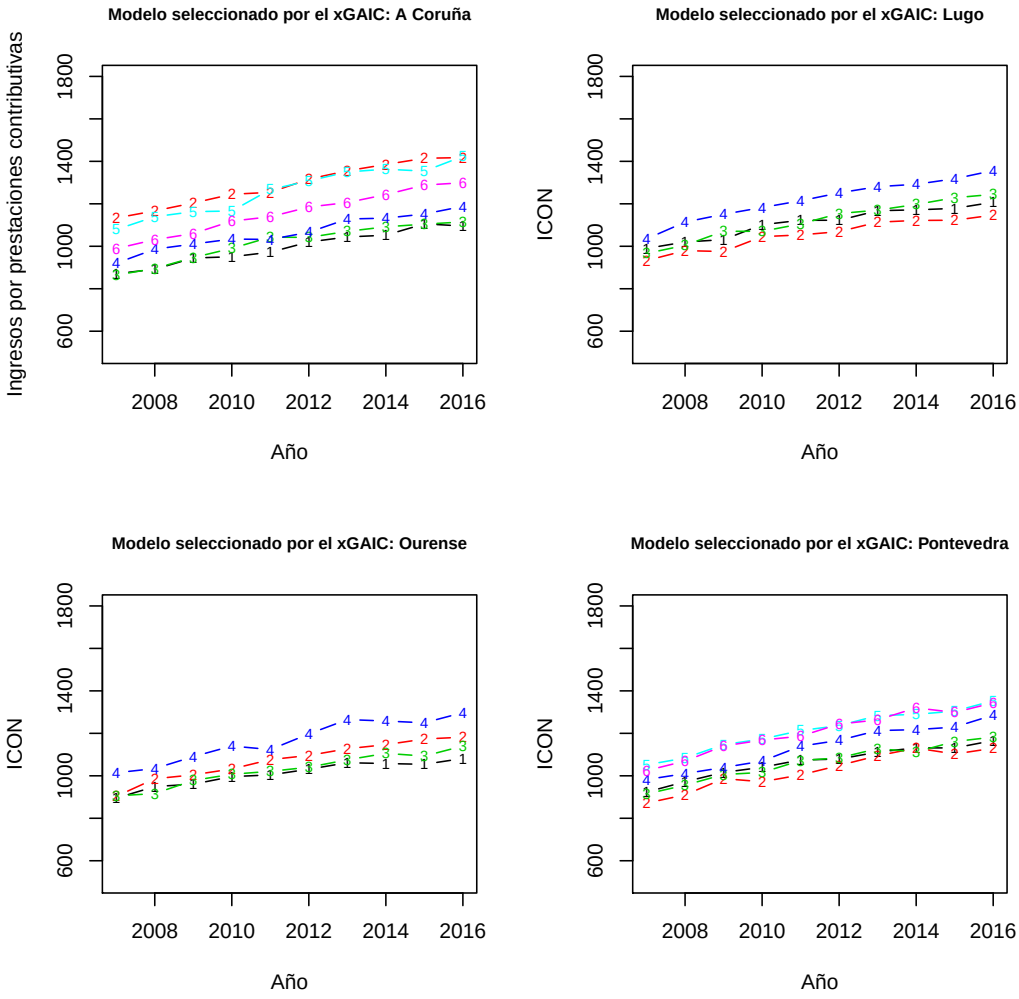

Figura 5.12: Evolución estimada de los ingresos por prestaciones contributivas medios realizada por el modelo seleccionado por el xGAIC para la variable ICON. Se muestran las estimaciones según la provincia y su área (indicada por su numeración dentro de su correspondiente provincia).

Cabe destacar que si bien el cAICH selecciona un modelo razonable, como podemos ver en el Apéndice B que apenas hay diferencia entre el resultado para dicho modelo y el modelo seleccionado por el cAICVyB, lo que indica que no se trata de un problema de ese criterio en particular. Eso no le sucede ni al xGAIC ni al cYMO, puesto que no somos capaces de generar remuestras del mismo.

En la Tabla 5.6 podemos ver el valor de los coeficientes del modelo seleccionado por el xGAIC. En ella observamos que el modelo es muy similar al modelo Indep ajustado en el Capítulo 4 para la variable ICON en cuanto a los valores y niveles de significación de la constante y del coeficiente de la variable PENMEAN. La principal diferencia es que el xGAIC descarta las variables PESUP y PESEC, que no son significativas en el modelo Indep.

	ICON	EFECTO T.	α	PENMEAN	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
Modelo xGAIC		Indep	2,49($<$ 0,001)	0,02($<$ 0,001)	1,05	0,26

Tabla 5.6: Coeficientes del modelo seleccionado por el xGAIC para la variable ICON (transformada), con sus respectivos nivel de significación (entre paréntesis)

En la Figura 5.13 podemos ver los coeficientes de variación en cada área de las estimaciones basadas en el modelo xGAIC, en los años 2007 y 2016. En este caso, al igual que para las variables ITOT e IAJ, los estimadores basados en el modelo mejoran los resultados de los estimadores directos en todas las áreas, tanto en 2007 como en 2016.

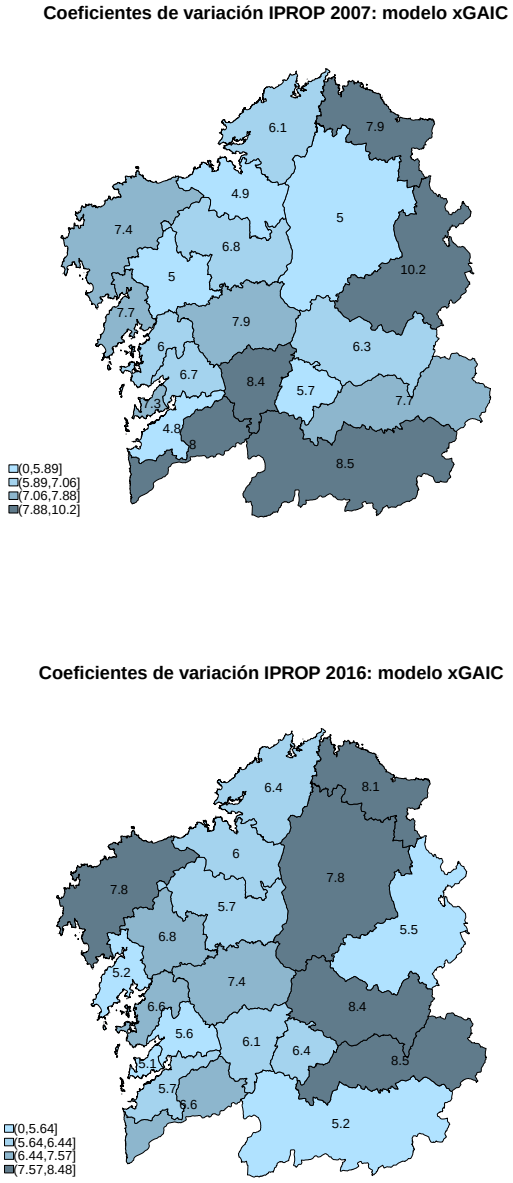

Figura 5.13: Mapas de los coeficientes de variación de las estimaciones basadas en el modelo seleccionado por el xGAIC, correspondientes a la variable IPROP por área en 2007 y 2016.

Capítulo 6

Conclusiones.

En este trabajo hemos estudiado como resolver los problemas que surgen en la estimación mediante estimadores directos de los ingresos medios de los hogares según su tipología, utilizando los datos de la EEH, realizada por el IGE.

En el Capítulo 4 realizamos el ajuste de diversos modelos de Fay-Herriot con efecto temporal para cada tipo de ingreso, considerando distintos tipos de efecto temporal y utilizando diversas variables socioeconómicas para explicar dichos ingresos. En todos los casos, los estimadores basados en los modelos se muestran mucho más precisos que los estimadores directos en todas las áreas. Esto puede ser de utilidad para el IGE, puesto que podrían utilizar estos estimadores en futuras ediciones de la EEH, de forma que mejorarían de forma notable las estimaciones que se realizaban hasta el momento para los distintos tipos de ingreso en el hogar.

Si bien los modelos mejoran los resultados dados por los estimadores directos, una de las tareas que nos propusimos en este trabajo fue escoger el mejor modelo entre los propuestos, así como las covariables del mismo. Para esta tarea se utilizó el AIC, adaptado a este contexto, concretamente utilizamos el criterio xGAIC y el cAIC considerando los modelos surgidos de las diversas combinaciones de covariables. Estos criterios nos proporcionaron criterios objetivos para seleccionar variables y modelos de forma que nos permitieron estudiar una gran cantidad de modelos sin necesidad de estudiarlos de forma individual, obteniendo grandes resultados en algunos de los criterios.

Respecto a los modelos seleccionados por los criterios, en general observamos que seleccionan modelos razonables, aunque podemos ver que el cAIC (tanto el cAICVyB como el cAICH) y en mayor medida el cYMO pueden llegar a seleccionar modelos que no son siquiera convergentes, lo que resulta en estimaciones que pueden llegar incluso a empeorar la realizada por los estimadores directos. Esto no le sucede al xGAIC, puesto que en todos los casos opta por modelos bastante buenos (rechazando claramente los no convergentes), mostrando cierta tendencia a tomar los modelos con efectos temporales independientes. Teniendo en cuenta que en la mayor parte de los casos los grados de libertad de los criterios coinciden, es posible que se deba a la verosimilitud tomada, concretamente a la matriz de covarianzas considerada puesto que, a diferencia de los otros criterios, tiene en cuenta las varianzas de los efectos aleatorios al no ser esta condicionada a $\mathbf{u}$. Si bien en algunas ocasiones sí hay una gran diferencia en los grados de libertad, en otros, como para las variables ITOT e IPROP, tenemos modelos con efectos temporales AR(1) con valores muy cercanos al modelo seleccionado, por lo que la diferencia puede deberse a que el criterio lo considere peor por su verosimilitud, aunque esto nos indica que en caso de encontrar un modelo con efectos temporales AR(1) o MA(1) lo suficientemente bueno, este podría ser seleccionado por el xGAIC. En cualquier caso, para todos los tipos de ingreso, ninguno de los modelos seleccionados por los otros criterios parecen superar claramente a los escogidos por el xGAIC. Por todo lo anterior consideramos que el criterio más fiable es el xGAIC y por lo tanto nos

decantaremos por los modelos que este selecciona.

Cabe destacar que surgieron muchos problemas computacionales tanto en el ajuste de los modelos como en la selección de los mismos. Los primeros parecen debidos al software utilizado, que parece mostrar problemas con la magnitud de las variables o incluso con la presencia de algunas de ellas, como es el caso de la constante. Estos problemas deberían poder solucionarse utilizando otro software o realizando algunos ajustes en las funciones del mismo. Las complicaciones surgidas en la selección son más problemáticas, puesto que, además de posibles problemas del software, parecen vinculadas al algoritmo *Fisher Scoring*. Este algoritmo parece ser muy sensible a los pequeños cambios que sufre la muestra original en las remuestras, lo que puede llevar a no obtener convergencia en el mismo, llevando en ocasiones al algoritmo a obtener estimaciones absurdas de la varianza. Esto último provoca que el bootstrap utilizado para el cálculo de los xGDF deba repetir muchas de las remuestras, puesto que la presencia de las no convergentes provocan graves perturbaciones en sus resultados, llegando incluso a obtener valores negativos para algunas de las varianzas.

En conclusión, en este trabajo hemos encontrado modelos que mejoran mucho los resultados dados por los estimadores directos. Dichos modelos pueden permitir al IGE en próximas ediciones de la EEH obtener estimadores mucho más precisos que los estimadores directos que se utilizan actualmente. Además hemos comprobado el funcionamiento de algunos criterios de selección en el contexto de áreas pequeñas. Entre ellos destacó el xGAIC, que mostró no solo ser más robusto que los criterios probados en este trabajo, si no que los modelos que este seleccionó superaron en varias ocasiones a los elegidos por el resto de criterios probados, sin llegar nunca a ser superados. Por ello, consideramos el xGAIC el criterio más recomendado, ya que mostró ser fiable tanto en la selección de las variables del modelo, como en el efecto temporal del mismo.

Apéndice A

Código R

En este apándice mostraremos las funciones programadas en R, que se utilizan para el cálculo de los criterios de selección en modelos de Fay-Herriot con efecto temporal. Para cada uno de ellos tenemos entonces tres funciones distintas según el efecto temporal. Los parámetros de entrada de las mismas son los siguientes:

- **fit**. Lista de igual forma que la que se obtiene en la salida de la función *fit.saery* del paquete *saery*.
- **X**. Matriz del diseño del modelo.
- **Y**. Vector con los valores de la variable respuesta.
- **sigma2edi**. Vector con las varianzas de los errores.
- **B**. Solo presente en las funciones referentes al xGAIC y al cYMO. Se trata del número de réplicas del bootstrap utilizado para estimar los xGDF. Su valor por defecto es 500.

y su salida es una lista con la siguiente información (adaptada al criterio utilizado) :

- El valor del criterio.
- Los grados de libertad del modelo.
- Valor de la log-verosimilitud
- Número de remuestras que tuvieron que repetirse. Esto es excclusivo de las funciones referentes al xGAIC y al cYMO.

De esta forma, este apéndice estará dividido en cuatro secciones (una por criterio), en las cuales aparecerán las funciones programadas para cada uno, indicando previamente el efecto temporal que en ellas se considera. Cabe destacar que el código presentado está especialmente diseñado para el problema, por lo que ya se consideran 20 áreas y 10 instantes temporales para cada una.

A.1. Funciones en R para el cálculo del cAICVyB

```
##############################################
# cAIC de Vaida y Blanchard
##############################################
```

```
cAICVyB_AR<-function(fit,X,Y,sigma2edi){


Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
rho<-fit$SIGMA[3,1]
Omega<-matrix(0,nrow=200,ncol=200)

for (i in 1:20){
Omegadaux<-matrix(0,nrow=10,ncol=10)

for (j in 1:10){
Omegadaux[j,j:10]<-rho^(0:(10-j))
Omegad<-t(Omegadaux)+Omegadaux

}
diag(Omegad)<-rep(1,10)
Omegad<-Omegad*(1/(1-(rho^2)))

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad

}

X<-X
Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve



Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY
```

```
mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*%
solve(V) %*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)



traza<- sum(diag(H))
N<-200
p<-dim(X)[2]
K<-((N-p-1)/(N-p-2))*(traza+1)+((p+1)/(N-p-2))

cAIC<- -2*(Verosimilitud)+2*K

return(list(cAIC=cAIC,Verosi=Verosimilitud,K=K))
}


###############
#Independiente
###############

cAICVyB_Indep<-function(fit,X,Y,sigma2edi){


Ve<-diag(sigma2edi)

sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]

Omega<-diag(1,200)

X<-X
Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve

Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)
```

```
Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
 Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY

traza<- sum(diag(H))
mu<-X %*% beta_noY %*% Y+Z %*% Vu %*% t(Z) %*% solve(V) %*% (Y-X %*% beta_noY %*% Y)
Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)



N<-200
p<-dim(X)[2]
K<-((N-p-1)/(N-p-2))*(traza+1)+((p+1)/(N-p-2))
cAIC<- -2*(Verosimilitud)+2*K

return(list(cAIC=cAIC,Verosi=Verosimilitud,K=K))
}




###################
# MA1
###################
cAICVyB_MA<-function(fit,X,Y,sigma2edi){


Ve<-diag(sigma2edi)

sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
theta<-fit$SIGMA[3,1]

Omega<-matrix(0,200,200)
Omegad <- matrix(0,nrow=10,ncol=10)
for(j in 2:10){
Omegad[j,j-1]<-(-theta)
Omegad[j-1,j]<-(-theta)
}
diag(Omegad) <- 1+theta^2
for (i in 1:20){

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad
}

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
```

```
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve

Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
 Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY #Falta por terminar
traza<- sum(diag(H))
mu<-X %*% beta_noY %*% Y+Z %*% Vu %*% t(Z) %*% solve(V)
 %*% (Y-X %*% beta_noY %*% Y)
Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)


N<-200
p<-dim(X)[2]
K<-((N-p-1)/(N-p-2))*(traza+1)+((p+1)/(N-p-2))
cAIC<- -2*(Verosimilitud)+2*K

return(list(cAIC=cAIC,Verosi=Verosimilitud,K=K))
}
```

A.2. Funciones en R para el cálculo del cAICH

```
##############################################
# cAIC de Han
##############################################

cAICH_AR<-function(fit,X,Y,sigma2edi){


Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
rho<-fit$SIGMA[3,1]
Omega<-matrix(0,nrow=200,ncol=200)
```

```
for (i in 1:20){
Omegadaux<-matrix(0,nrow=10,ncol=10)

for (j in 1:10){
Omegadaux[j,j:10]<-rho^(0:(10-j))
Omegad<-t(Omegadaux)+Omegadaux

}
diag(Omegad)<-rep(1,10)
Omegad<-Omegad*(1/(1-(rho^2)))

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad

}

X<-X
Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve

Vc<-sigma2*Omega+Ve


Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
 Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY


mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
 %*% (Y-X %*% beta_noY %*% Y)
Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)

traza<- sum(diag(H))
N<-200
p<-dim(X)[2]
I<-diag(1,200)
```

```
P<-I-X %*% solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
rs<-solve(V) %*% P %*% Y
stau<-sum(diag((solve(V) %*% P) %*% ((solve(V) %*% P))))-2*t(rs) %*% solve(V) %*% rs
K<-traza-2*stau^(-1)*t(rs) %*% solve(V) %*% P %*% rs
cAIC<- -2*(Verosimilitud)+2*K

return(list(cAIC=cAIC,Verosi=Verosimilitud,K=K))
}
#############
#Indep
#############
cAICH_Indep<-function(fit,X,Y,sigma2edi){


Ve<-diag(sigma2edi)

sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]

Omega<-diag(1,200)

X<-X
Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1%*%t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve

Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)

H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
 Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY

traza<- sum(diag(H))

mu<-X %*% beta_noY %*% Y+Z %*% Vu %*% t(Z )%*% solve(V)
%*% (Y-X %*% beta_noY %*% Y)
```

```
Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)

N<-200
p<-dim(X)[2]
I<-diag(1,200)
P<-I-X %*% solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
rs<-solve(V) %*% P %*% Y
stau<-sum(diag((solve(V) %*% P) %*% ((solve(V) %*% P))))-2*t(rs) %*% solve(V) %*% rs
K<-traza-2*stau^(-1)*t(rs) %*% solve(V) %*% P %*% rs
cAIC<- -2*(Verosimilitud)+2*K

return(list(cAIC=cAIC,Verosi=Verosimilitud,K=K))
}
##############
# MA1
##############
cAICH_MA<-function(fit,X,Y,sigma2edi){

Ve<-diag(sigma2edi)

sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
theta<-fit$SIGMA[3,1]

Omega<-matrix(0,200,200)
Omegad <- matrix(0,nrow=10,ncol=10)
for(j in 2:10){
Omegad[j,j-1]<-(-theta)
Omegad[j-1,j]<-(-theta)
}
diag(Omegad) <- 1+theta^2
for (i in 1:20){

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad
}

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1%*%t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve

Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega
```

```
Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*% Vu %*% t(Z) %*% solve(V)
 %*% X %*% beta_noY #Falta por terminar
traza<- sum(diag(H))
mu<-X %*% beta_noY %*% Y+Z %*% Vu %*% t(Z) %*% solve(V) %*% (Y-X %*% beta_noY %*% Y)
Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)

N<-200
p<-dim(X)[2]
I<-diag(1,200)
P<-I-X %*% solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
rs<-solve(V) %*% P %*% Y
stau<-sum(diag((solve(V) %*% P) %*% ((solve(V) %*% P))))-2*t(rs) %*% solve(V) %*% rs
K<-traza-2*stau^(-1)*t(rs) %*% solve(V) %*% P %*% rs
cAIC<- -2*(Verosimilitud)+2*K

return(list(cAIC=cAIC,Verosi=Verosimilitud,K=K))
}
```

A.3. Funciones en R para el cálculo del cYMO

```
################################################
# cYMO
################################################

cYMO_AR<-function(fit,X,Y,sigma2edi,B=500){

#cargamos los datos del fit
Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
rho<-fit$SIGMA[3,1]

#generamos la matriz Omega
Omega<-matrix(0,nrow=200,ncol=200)

for (i in 1:20){
Omegadaux<-matrix(0,nrow=10,ncol=10)

for (j in 1:10){
Omegadaux[j,j:10]<-rho^(0:(10-j))
Omegad<-t(Omegadaux)+Omegadaux

}
```

```
diag(Omegad)<-rep(1,10)
Omegad<-Omegad*(1/(1-(rho^2)))

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad

}

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve
Vu1<-sigma1*diag(1,20)
Vu2<-sigma2*Omega
Vu<-matrix(0,220,220)
Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X)%*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY

mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
%*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)

######
#Boot.
######

Yb<-matrix(nrow=B,ncol=200)
estmub<-matrix(nrow=B,ncol=200)
residu<-matrix(nrow=B,ncol=200)
b<-1
bad<-0

while(b<=B){
eb<-numeric(200)
u1b<-rnorm(20,sd=sqrt(sigma1))
u2b<-numeric(200)

for (j in seq(1,191,by=10)){
u2b[j]<-(1/(1-rho^2)^0.5)*rnorm(1,sd=sqrt(sigma2))
}
```

```
for (j in seq(1,191,by=10)){
for (k in 1:9){
u2b[j+k]<-rho*u2b[(j+k-1)]+rnorm(1,sd=sqrt(sigma2))
}
}


eb<-rnorm(200,sd=sqrt(sigma2edi))
u2b<-u2b[1:200]
ub<-c(u1b,u2b)

mub<-X %*% beta_noY %*% Y+Z %*% ub

Yb[b,]<-mub+eb

emub<-try(eblup.saery(X,Yb[b,],D, md, sigma2edi,B=0,model="AR1"),TRUE)
fitb<-try(fit.saery(X,Yb[b,],D, md, sigma2edi,model="AR1"),TRUE)

res<-try(emub$resid,TRUE)
if(class(emub)=="try-error"){

b <- b
bad<-bad+1

if (bad>1000){
break()
}

}

else if(fitb$iteration==20){

b <- b
bad<-bad+1
}
else{
residu[b,]<-emub$resid
estmub[b,]<-emub$eblup
b<-b+1
}
}

meanmub<-apply(estmub,2,mean)
meanYb<-apply(Yb,2,mean)
diffmu<-t(t(estmub)-meanmub)
diffY<-t(t(Yb)-meanYb)

sumando<-numeric(B)
invV<-solve(V)
for (b in 1:B){
sumando[b]<-(1/(B-1))*(t(diffmu[b,]) %*% invV %*% diffY[b,])
}
```

```
print(bad/(B+bad))
GDF<-sum(sumando)
cYMO<- -2*Verosimilitud+GDF
return(list(cYMO=cYMO,Verosi=Verosimilitud,GDF=GDF,bad=bad))
}

########
#Indep
########
cYMO_Indep<-function(fit,X,Y,sigma2edi,B=500){

#cargamos los datos del fit
Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
rho<-fit$SIGMA[3,1]

Omega<-diag(1,200)
Z_1<-numeric(0)

for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)


V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vc<-sigma2*Omega+Ve

Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

#C?lculo de la verosimilitud
beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY


mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
%*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)
```

```
######
#Boot.
######

Yb<-matrix(nrow=B,ncol=200)
estmub<-matrix(nrow=B,ncol=200)
b<-1
bad<-0
while(b<=B){

eb<-numeric(200)


eb<-rnorm(200,sd=sqrt(sigma2edi))


u1b<-rnorm(20,sd=sqrt(sigma1))
u2b<-rnorm(200,sd=sqrt(sigma2))

ub<-c(u1b,u2b)
#media en la primera remuestra

mub<-X %*% beta_noY %*% Y+Z %*% ub

Yb[b,]<-mub+eb

fitb<-try(fit.saery(X,Yb[b,],D, md, sigma2edi,model="Indep"),TRUE)

emub<-try(eblup.saery(X,Yb[b,],D, md, sigma2edi,B=0,model="Indep"),TRUE)
res<-try(emub$resid,TRUE)


if(class(emub)=="try-error"){

b <- b
bad<-bad+1
if (bad>1000){
break()
}
}

else if(fitb$iteration==20){
b <- b
bad<-bad+1
}

else{
estmub[b,]<-emub$eblup
b<-b+1
}
}
```

```
meanmub<-apply(estmub,2,mean)
meanYb<-apply(Yb,2,mean)
diffmu<-t(t(estmub)-meanmub)
diffY<-t(t(Yb)-meanYb)

sumando<-numeric(B)
invV<-solve(V)
for (b in 1:B){
sumando[b]<-(1/(B-1))*(t(diffmu[b,]) %*% invV %*% diffY[b,])
}

GDF<-sum(sumando)
cYMO<- -2*Verosimilitud+GDF
return(list(cYMO=cYMO,Verosi=Verosimilitud,GDF=GDF,bad=bad))

}

#######
#MA1
#######
cYMO_MA<-function(fit,X,Y,sigma2edi,B=500){

Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
theta<-fit$SIGMA[3,1]

Omega<-matrix(0,200,200)
Omegad <- matrix(0,nrow=10,ncol=10)
for(j in 2:10){
Omegad[j,j-1]<-(-theta)
Omegad[j-1,j]<-(-theta)
}
diag(Omegad) <- 1+theta^2

for (i in 1:20){

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad
}

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
```

```
Vu1<-sigma1*diag(1,20)

Vu2<-sigma2*Omega

Vu<-matrix(0,220,220)

Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY

mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
 %*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(Ve))-
(1/2)*t((Y-mu)) %*% solve(Ve) %*% (Y-mu)

######
#Boot.
######
Yb<-matrix(nrow=B,ncol=200)
estmub<-matrix(nrow=B,ncol=200)
residu<-matrix(nrow=B,ncol=200)
b<-1
bad<-0

while(b<=B){
if (bad>1000){
break()
}
eb<-numeric(200)
u1b<-rnorm(20,sd=sqrt(sigma1))
u2b<-numeric(200)

whitenoise<-rnorm(201,sd=sqrt(sigma2))
u2b<-whitenoise[2:201]-theta*whitenoise[1:200]


eb<-rnorm(200,sd=sqrt(sigma2edi))

ub<-c(u1b,u2b)

mub<-X %*% beta_noY %*% Y+Z %*% ub

Yb[b,]<-mub+eb

emub<-try(eblup.saery(X,Yb[b,],D, md, sigma2edi,B=0,model="MA1"),TRUE)
fitb<-try(fit.saery(X,Yb[b,],D, md, sigma2edi,model="MA1"),TRUE)
sigma1b<-try(fitb$SIGMA[1,1])
sigma2b<-try(fitb$SIGMA[2,1])
```

```
thetab<-try(fitb$SIGMA[3,1])
res<-try(emub$resid)
#hist(eblup.saery(X,ydi,D, md, sigma2edi,model="MA1",B=0)$resid)

if(class(emub)=="try-error"){

b <- b
bad<-bad+1
if (bad>1000){
break()
}
}

else if(fitb$iteration==20){
b <- b
bad<-bad+1
}

else{
residu[b,]<-emub$resid
estmub[b,]<-emub$eblup
b<-b+1
}
}

meanmub<-apply(estmub,2,mean)
meanYb<-apply(Yb,2,mean)
diffmu<-t(t(estmub)-meanmub)
diffY<-t(t(Yb)-meanYb)

sumando<-numeric(B)
invV<-solve(V)
for (b in 1:B){
sumando[b]<-(1/(B-1))*(t(diffmu[b,]) %*% invV %*% diffY[b,])
}
print(bad/(B+bad))

GDF<-sum(sumando)
cYMO<- -2*Verosimilitud+GDF
return(list(cYMO=cYMO,Verosi=Verosimilitud,GDF=GDF,bad=bad))

}
```

A.4. Funciones en R para el cálculo del xGAIC

```
###########
# xGAIC
###########

xGAIC_AR<-function(fit,X,Y,sigma2edi,B=500){
```

```
#cargamos los datos del fit
Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
rho<-fit$SIGMA[3,1]

#generamos la matriz Omega
Omega<-matrix(0,nrow=200,ncol=200)

for (i in 1:20){
Omegadaux<-matrix(0,nrow=10,ncol=10)

for (j in 1:10){
Omegadaux[j,j:10]<-rho^(0:(10-j))
Omegad<-t(Omegadaux)+Omegadaux

}
diag(Omegad)<-rep(1,10)
Omegad<-Omegad*(1/(1-(rho^2)))
Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad
}

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vu1<-sigma1*diag(1,20)
Vu2<-sigma2*Omega
Vu<-matrix(0,220,220)
Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z
%*% Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY

mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
 %*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(V))-
(1/2)*t((Y-mu)) %*% solve(V) %*% (Y-mu)

######
#Boot.
######
```

```
Yb<-matrix(nrow=B,ncol=200)
estmub<-matrix(nrow=B,ncol=200)
residu<-matrix(nrow=B,ncol=200)
b<-1
bad<-0

while(b<=B){

eb<-numeric(200)
u1b<-rnorm(20,sd=sqrt(sigma1))
u2b<-numeric(200)

for (j in seq(1,191,by=10)){
u2b[j]<-(1/(1-rho^2)^0.5)*rnorm(1,sd=sqrt(sigma2))
}

for (j in seq(1,191,by=10)){
for (k in 1:9){
u2b[j+k]<-rho*u2b[(j+k-1)]+rnorm(1,sd=sqrt(sigma2))
}
}


eb<-rnorm(200,sd=sqrt(sigma2edi))
u2b<-u2b[1:200]
ub<-c(u1b,u2b)
#media en la primera remuestra

mub<-X %*% beta_noY %*% Y+Z %*% ub

Yb[b,]<-mub+eb

emub<-try(eblup.saery(X,Yb[b,],D, md, sigma2edi,B=0,model="AR1"),TRUE)
res<-try(emub$resid,TRUE)
fitb<-try(fit.saery(X,Yb[b,],D, md, sigma2edi,model="AR1"),TRUE)

if(class(emub)=="try-error"){
b <- b
bad<-bad+1
#print(b)
if (bad>1000){
break()
}
}

else if(fitb$iteration==20){
b <- b
bad<-bad+1
}
else{
residu[b,]<-emub$resid
estmub[b,]<-emub$eblup
```

```
b<-b+1
}
}

meanmub<-apply(estmub,2,mean)
meanYb<-apply(Yb,2,mean)
diffmu<-t(t(estmub)-meanmub)
diffY<-t(t(Yb)-meanYb)

sumando<-numeric(B)
invV<-solve(V)
for (b in 1:B){
sumando[b]<-(1/(B-1))*(t(diffmu[b,]) %*% invV %*% diffY[b,])
}
print(bad/(B+bad))
GDF<-sum(sumando)
xGAIC<- -2*Verosimilitud+GDF
return(list(xGAIC=xGAIC,Verosi=Verosimilitud,GDF=GDF,bad=bad))
}

##########
# INDEP
##########

xGAIC_Indep<-function(fit,X,Y,sigma2edi,B=500){

Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]

Omega<-diag(1,200)

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)


V<-sigma1*Z_1%*%t(Z_1)+sigma2*Omega+Ve
Vu1<-sigma1*diag(1,20)
Vu2<-sigma2*Omega
Vu<-matrix(0,220,220)
Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X)%*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY
```

```
mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
 %*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(V))-
(1/2)*t((Y-mu))%*%solve(V)%*%(Y-mu)

######
#Boot.
######

Yb<-matrix(nrow=B,ncol=200)
estmub<-matrix(nrow=B,ncol=200)
b<-1
bad<-0
while(b<=B){

eb<-numeric(200)

eb<-rnorm(200,sd=sqrt(sigma2edi))

u1b<-rnorm(20,sd=sqrt(sigma1))
u2b<-rnorm(200,sd=sqrt(sigma2))

ub<-c(u1b,u2b)


mub<-X %*% beta_noY %*% Y+Z %*% ub

Yb[b,]<-mub+eb

emub<-try(eblup.saery(X,Yb[b,],D, md, sigma2edi,B=0,model="Indep"),TRUE)
fitb<-try(fit.saery(X,Yb[b,],D, md, sigma2edi,model="Indep"),TRUE)

res<-try(emub$resid,TRUE)

if(class(emub)=="try-error"){
b <- b
bad<-bad+1
if (bad>1000){
break()
}
}

else if(fitb$iteration==20){
b <- b
bad<-bad+1
}

else{
estmub[b,]<-emub$eblup
```

```
b<-b+1
}
}

meanmub<-apply(estmub,2,mean)
meanYb<-apply(Yb,2,mean)
diffmu<-t(t(estmub)-meanmub)
diffY<-t(t(Yb)-meanYb)

sumando<-numeric(B)
invV<-solve(V)
for (b in 1:B){
sumando[b]<-(1/(B-1))*(t(diffmu[b,]) %*% invV %*% diffY[b,])
}

GDF<-sum(sumando)

xGAIC<- -2*Verosimilitud+GDF

return(list(xGAIC=xGAIC,Verosi=Verosimilitud,GDF=GDF,bad=bad))
}


#######
#MA1
#######

xGAIC_MA<-function(fit,X,Y,sigma2edi,B=500){

Ve<-diag(sigma2edi)
Y<-as.matrix(Y)
sigma1<-fit$SIGMA[1,1]
sigma2<-fit$SIGMA[2,1]
theta<-fit$SIGMA[3,1]

Omega<-matrix(0,200,200)
Omegad <- matrix(0,nrow=10,ncol=10)
for(j in 2:10){
Omegad[j,j-1]<-(-theta)
Omegad[j-1,j]<-(-theta)
}
diag(Omegad) <- 1+theta^2

for (i in 1:20){

Omega[((i-1)*10+1):(i*10),((i-1)*10+1):(i*10)]<-Omegad
}

Z_1<-numeric(0)
for (i in 1:20){
suvdr<-c(rep(0,(i-1)*10),rep(1,10),rep(0,(200-i*10)))
Z_1<-cbind(Z_1,suvdr)
```

```
}
Z_2<-diag(1,200)
Z<-cbind(Z_1,Z_2)

V<-sigma1*Z_1 %*% t(Z_1)+sigma2*Omega+Ve
Vu1<-sigma1*diag(1,20)
Vu2<-sigma2*Omega
Vu<-matrix(0,220,220)
Vu[1:20,1:20]<-Vu1
Vu[21:220,21:220]<-Vu2

beta_noY<-solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V)
H<- X %*% beta_noY+Z %*% Vu %*% t(Z) %*% solve(V)-Z %*%
 Vu %*% t(Z) %*% solve(V) %*% X %*% beta_noY

mu<-X %*% beta_noY %*% Y + Z %*% Vu %*% t(Z) %*% solve(V)
%*% (Y-X %*% beta_noY %*% Y)

Verosimilitud<-(-1/2)*200*log(2*pi)-(1/2)*log(det(V))-
(1/2)*t((Y-mu)) %*% solve(V) %*% (Y-mu)

################################################

Yb<-matrix(nrow=B,ncol=200)
estmub<-matrix(nrow=B,ncol=200)
residu<-matrix(nrow=B,ncol=200)
b<-1
bad<-0

while(b<=B){
if (bad>1000){
break()
}

eb<-numeric(200)
u1b<-rnorm(20,sd=sqrt(sigma1))

u2b<-numeric(200)

whitenoise<-rnorm(201,sd=sqrt(sigma2))

u2b<-whitenoise[2:201]-theta*whitenoise[1:200]

eb<-rnorm(200,sd=sqrt(sigma2edi))

ub<-c(u1b,u2b)
#media en la primera remuestra

mub<-X %*% beta_noY %*% Y+Z %*% ub

Yb[b,]<-mub+eb
```

```
emub<-try(eblup.saery(X,Yb[b,],D, md, sigma2edi,B=0,model="MA1"),TRUE)
fitb<-try(fit.saery(X,Yb[b,],D, md, sigma2edi,model="MA1"),TRUE)
sigma1b<-try(fitb$SIGMA[1,1])
sigma2b<-try(fitb$SIGMA[2,1])
thetab<-try(fitb$SIGMA[3,1])
res<-try(emub$resid)

if(class(emub)=="try-error"){

b <- b
bad<-bad+1
}

else if(fitb$iteration==20){
b <- b
bad<-bad+1
}

else{
residu[b,]<-emub$resid
estmub[b,]<-emub$eblup
b<-b+1
}
}

meanmub<-apply(estmub,2,mean)
meanYb<-apply(Yb,2,mean)
diffmu<-t(t(estmub)-meanmub)
diffY<-t(t(Yb)-meanYb)

sumando<-numeric(B)
invV<-solve(V)
for (b in 1:B){
sumando[b]<-(1/(B-1))*(t(diffmu[b,]) %*% invV %*% diffY[b,])
}
print(bad/(B+bad))

GDF<-sum(sumando)
xGAIC<- -2*Verosimilitud+GDF
return(list(xGAIC=xGAIC,Verosi=Verosimilitud,GDF=GDF,bad=bad))

}
```

Apéndice B

Tablas AIC

En este apéndice mostraremos en diversas tablas (una para cada tipo de ingreso) los modelos probados para la realización del Capítulo 5, mostrándose junto el resultado para los distintos criterios de seleccíon, las variables de cada uno de los modelos y el tipo de efecto temporal de los mismos.

ITOT						AIC			
EFECTO T.	CONS	PFBLIM	REND	PESUP	PM65	cAICVyB	cAICH	xGAIC	cYMO
AR(1)	✓	✓	✓			**632.83**	**634.00**	653.20	572.90
AR(1)	✓		✓			651.71	652.28	694.50	577.32
AR(1)	✓	✓				648.19	651.38	798.94	**553.57**
Indep	✓	✓	✓	✓	✓	641.09	642.15	**648.79**	576.41
Indep	✓		✓	✓	✓	666.70	667.47	692.29	577.50
Indep	✓			✓	✓	686.92	687.59	731.31	579.50
Indep	✓				✓	684.50	685.24	734.54	579.43
Indep	✓			✓		686.31	686.94	748.59	577.54
Indep	✓		✓		✓	666.54	667.34	693.27	576.28
Indep	✓		✓	✓		666.25	667.03	693.24	575.15
Indep	✓	✓		✓	✓	676.60	677.42	696.56	581.73
Indep	✓	✓			✓	673.32	674.22	700.31	581.17
Indep	✓	✓				674.23	675.08	720.00	579.25
Indep	✓		✓			665.84	666.66	695.40	575.19
Indep	✓	✓		✓		674.93	675.69	720.33	577.50
Indep	✓	✓	✓		✓	639.95	641.05	649.80	576.56
Indep	✓	✓	✓			638.87.63	639.97	650.57	575.12
Indep	✓	✓	✓	✓		640.04	641.12	650.39	575.86

Tabla B.1: Tabla con los modelos para los modelos considerados para la variable ITOT en el capítulo 5 y los resultados de los criterios de selección para los mismos.

IAJ					AIC			
EFECTO T.	CONS	PM65	REND	PESUP	cAICVyB	cAICH	xGAIC	cYMO
AR(1)		✓	✓	✓	804.22	804.96	813.21	735.51
AR(1)			✓	✓	809.80	810.55	842.80	732.16
AR(1)				✓	991.88	991.09	NA	NA
AR(1)			✓		823.22	823.95	878.70	733.46
AR(1)		✓		✓	999.50	998.98	NA	NA
AR(1)		✓			856.74	856.84	NA	NA
AR(1)		✓	✓		805.74	806.562	811.31	**730.87**
Indep	✓	✓	✓	✓	794.79	795.71	790.76	734.30
Indep	✓		✓	✓	795.01	796.00	**789.35**	734.14
Indep	✓			✓	820.96	821.93	861.44	743.27
Indep	✓		✓		793.31	794.29	791.37	734.32
Indep	✓	✓		✓	817.84	818.93	839.41	745.75
Indep	✓	✓			815.18	816.38	841.06	746.73
Indep	✓	✓	✓		792.89	793.81	792.34	735.31
MA(1)	✓	✓	✓	✓	793.99	794.95	799.35	745.17
MA(1)	✓		✓	✓	793.98	795.00	796.94	742.11
MA(1)	✓			✓	891.05	884.111	NA	NA
MA(1)	✓		✓		792.87	793.86	796.72	739.45
MA(1)	✓	✓		✓	835.19	839.00	965.69	772.38
MA(1)	✓	✓			898.98	888.05	1150.71	761.34
MA(1)	✓	✓	✓		**792.55**	**793.48**	800.94	740.99

Tabla B.2: Tabla con los modelos para los modelos considerados para la variable IAJ en el capítulo 5 y los resultados de los criterios de selección para los mismos. Los valores NA indican que el criterio no pudo obtener el resultado.

IPROP					AIC			
EFECTO T.	CONS	PM65	REND	PFTIPO3	cAICVyB	cAICH	xGAIC	cYMO
AR(1)	✓	✓	✓	✓	**890.26**	**891.00**	891.65	806.99
AR(1)	✓		✓	✓	891.03	891.76	894.59	804.81
AR(1)	✓			✓	895.29	896.03	909.11	807.28
AR(1)	✓		✓		893.51	894.26	900.25	807.57
AR(1)	✓	✓		✓	895.93	896.65	911.09	807.38
AR(1)	✓	✓			899.62	**9**00.54	922.72	**800.08**
AR(1)	✓	✓	✓		894.07	894.80	900.94	807.85
Indep	✓	✓	✓	✓	**893.06**	893.95	**890.32**	808.27
Indep	✓		✓	✓	895.19	896.06	893.63	807.37
Indep	✓			✓	904.85	905.73	913.40	810.38
Indep	✓		✓		898.15	899.018	899.86	805.92
Indep	✓	✓		✓	905.60	906.45	913.32	810.24
Indep	✓	✓			911.69	912.48	925.67	809.39
Indep	✓	✓	✓		898.47	899.32	901.54	807.14
MA(1)	✓	✓	✓	✓	891.41	892.26	895.38	811.94
MA(1)	✓		✓	✓	892.42	893.25	902.57	817.27
MA(1)	✓			✓	908.73	910.13	1044.51	858.80
MA(1)	✓		✓		898.21	899.07	967.84	842.40
MA(1)	✓	✓		✓	909.83.01	911.33	1048.01	856.78
MA(1)	✓	✓			929.20	939.04	1106.44	835.85
MA(1)	✓	✓	✓		898.69	899.54	965.96	844.49

Tabla B.3: Tabla con los modelos para los modelos considerados para la variable IPROP en el capítulo 5 y los resultados de los criterios de selección para los mismos.

ICON					AIC			
MODELO	CONS	PENMEAN	PESUP	PESEC	cAICVyB	cAICH	xGAIC	cYMO
AR(1)		✓	✓	✓	560.80	**561.49**	596.64	496.10
AR(1)			✓	✓	565.79	566.74	NA	NA
AR(1)				✓	**560.52**	561.81	NA	NA
AR(1)			✓		631.68	631.69	NA	NA
AR(1)		✓		✓	562.27	563.00	594.75	497.20
AR(1)		✓			566.04	566.94	598.33	496.13
AR(1)		✓	✓		565.62	566.51	597.96	**495.63**
Indep	✓	✓	✓	✓	571.44	572.42	589.39	496.26
Indep	✓		✓	✓	652.18	652.44	760.89	503.28
Indep	✓			✓	668.59	668.86	815.14	507.70
Indep	✓		✓		659.88	660.18	782.31	505.72
Indep	✓	✓		✓	574.25	575.28	588.79	497.90
Indep	✓	✓			575.09	576.16	**588.76**	499.03
Indep	✓	✓	✓		573.48	574.51	587.77	497.81
MA(1)		✓	✓	✓	562.71	563.55	600.61	498.25
MA(1)			✓	✓	646.11	646.47	787.59	509.65
MA(1)				✓	704.23	696.75	NA	NA
MA(1)			✓		640.81	640.91	868.01	530.14
MA(1)		✓		✓	709.02	704.90	NA	NA
MA(1)		✓			714.95	710.89	NA	NA
MA(1)		✓	✓		565.95	566.89	605.25	500.66

Tabla B.4: Tabla con los modelos para los modelos considerados para la variable ICON en el capítulo 5 y los resultados de los criterios de selección para los mismos. Los valores NA indican que el criterio no pudo obtener el resultado.

Bibliografía

[1] Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. *In International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267?81. Budapest: Akademia Kiado.

[2] Battese,G. E., Harter, R. M., and Fuller, W. A. (1998). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.

[3] Esteban, M.D., Morales, D. y Perez, A. (2014). saery: Small Area Estimation for Rao and Yu Model. R package version 1.0. URL: https://CRAN.R-project.org/package=saery.

[4] Esteban, M.D., Morales, D., Perez, A. y Santamaría, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*. 56 (10), pp. 2840-2855.

[5] Fay, R. and Herriot, R. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*. 74:269-277

[6] Han B. (2013) Conditional Akaike information in the Fay-Herriot model. *Statistical Methodology*, 11, 53-67.

[7] Herrador, M., Morales, D., Esteban, M. D., Sánchez, A., Santamaría, L., Marhuenda, Y., Pérez, A. y Molina, I. (2009). Estimadores de áreas pequeñas basados en modelos para la Encuesta de Población Activa. *Estadística Española*, vol. 51, 170, 133-172.

[8] Jiang J., Lahiri P. (2006) Mixed Model Prediction and Small Area Estimation. *Test* 15, 1-96

[9] Liang, H., Wu H. and Zou G. (2008) Miscellanea. A note on conditional AIC for linear mixed-effects models. *Biometrika* , 95, 3, 773-778.

[10] Lombardía, M., López-Vizcaíno, E., Rueda, C. (2017). Mixed generalized Akaike information criterion for small area models. *Journal of the Royal Statistical Society.*

[11] Marhuenda, Y., Molina, I. y Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

[12] Molina I, Marhuenda Y (2015). sae: An R Package for Small Area Estimation. *The R Journal*, 7(1), 81-98. URL: https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf

[13] Patterson, H. D. y Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

[14] Pfeffermann D. (2013) New Important Developments in Small Area Estimation. *Statistic Science* Vol. 28, 1 40-68.

[15] Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

[16] Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.

[17] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[18] Rao, J. N. K. (2003) Small Area Stimation.

[19] Rao, J. N. K. y Yu, M. (1994). Small area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics*, vol. 22, 4, 511-528.

[20] Thompson, J. W. A. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 33, 273-289.

[21] Vaida F. and Blanchard S. (2005) Conditional Akaike information for mixed-effects models. *Biometrika*, 92, 2, 351-370.

[22] You, C., Muller, S. and Ormerod, J. (2016) On generalized degrees of freedom with application in linear mixed models selection. *Statist. Comput.*, 26, 199?210.