Trabajo Fin de Máster

Selección de modelos en modelos lineales mixtos. Aplicación a datos económicos de la Comunidad de Galicia

Belén Vicente Mata

Máster en Técnicas Estadísticas

Curso 2015-2016

Propuesta de Trabajo Fin de Máster

Título en galego: Selección de modelos en modelos lineais mixtos. Aplicación a datos económicos da Comunidade de Galicia.
Título en español: Selección de modelos en modelos lineales mixtos. Aplicación a datos económicos de la Comunidad de Galicia
English title: Selection of models in mixed linear models. Application to economic data from the Community of Galicia.
Modalidad: Modalidad B
Autora: Belén Vicente Mata, Universidade da Coruña
Directora: María José Lombardía Cortiña, Universidade da Coruña.
Tutores: Esther López Vizcaíno, IGE; Javier Barriuso Noya, 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, profesora de la Universidade da Coruña; doña Esther López Vizcaíno responsable del Servicio de Difusión e Información del IGE; y don Javier Barriuso Noya, Jefe de Servicio de Estadísticas Sociales del IGE, informan que el Trabajo Fin de Máster titulado

Selección de modelos en modelos lineales mixtos. Aplicación a datos económicos de la Comunidad de Galicia

fue realizado bajo su dirección por doña Belén Vicente Mata para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 28 de junio de 2016.

La directora:

Doña María José Lombardía Cortiña

La tutora:

Doña Esther López Vizcaíno

El tutor:

Don Javier Barriuso Noya

La autora:

Doña Belén Vicente Mata

Agradecimientos

A mi directora del TFM, María José Lombarbardía Cortiña por su confianza en mi para realizar este proyecto, su paciencia, apoyo e interés. Gracias por haberme guiado en todo momento y estar siempre a disposición para cualquier duda.

Al IGE por permitirme realizar las prácticas de empresa con ellos. En especial a mi tutora, Esther López Vizcaíno por su ayuda, apoyo y disponibilidad. Gracias por haber contribuido tanto en mi formación durante los cinco meses de prácticas.

A mi familia y amigos.

Índice general

Resumen

Resumen en español

El objetivo de este trabajo es el estudio de la selección de modelos mixtos en áreas pequeñas mediante algún criterio de selección de variables. En concreto, se utilizará el criterio del AIC para modelos mixtos que nos permitirá saber que covariables se introducen finalmente en un modelo. Hablaremos de un AIC marginal y un AIC condicional según cual sea el objetivo del investigador.

Tras describir la metodología, se aplicará a unos datos reales ofrecidos por el Instituto Galego de Estadística (IGE). Trataremos de estimar el ingreso medio total en el hogar para el año 2013 en unas áreas pequeñas determinadas por el IGE. Los resultados obtenidos para el año 2013 serán aplicados para conocer los ingresos desde el año 2007 al 2012.

Estas nuevas técnicas de estimación, estimación en áreas pequeñas, nos permitirá obtener unos estimadores más precisos que los estimadores calculados directamente de la muestra, los estimadores basados en el modelo.

English abstract

The aim of this proyect is study of selection of mixed models in small areas through a selection criterion variables. Specifically, the criterion of AIC for mixed models that allows us to know what covariates are finally introduced in to a model. We will talk of a marginal AIC and conditional AIC whichever the goal of the researcher.

After describing the methodology, applied to actual data provided by the IGE. We will try to estimated the average total household income for 2013 and in some areas determinated by the IGE. The result for 2013 will be applied to find out the revenue from 2007 to 2012.

This new estimation techniques like small areas estimation allow us to obtain more accurate estimates than the ones calculated directly from the sample; these new estimated are called estimated based on the modelo. The sample used is Living Conditions

Survey.

Introducción

El tema principal de este trabajo será la estimación en áreas pequeñas por lo que antes de comenzar con la introducción definiremos lo que se conoce como "áreas" o "áreas pequeñas". Comúnmente se utiliza la expresión de " áreas pequeñas "para referirse a áreas geográficas pequeñas, como municipios, comarcas o distritos sanitarios, o a pequeñas subpoblaciones, como desempleados, juventud, minorías étnicas o minusválidos, donde la muestra de ese área no es suficientemente grande para obtener estimadores directos de los parámetros con cierta precisión. Hace referencia a una división de la población para la variable de interés.

Cada año, desde 1999, el Instituto Galego de Estadística (IGE) realiza una encuesta dirigida a los hogares gallegos con el objetivo de obtener información sobre las características socioeconómicas de dichos hogares y analizar las diferencias existentes entre las veinte áreas territoriales de Galicia. Esta actividad estadística se denomina "Encuesta de Condiciones de Vida"(ECV) y consta de dos módulos: uno primero de carácter general, igual cada año; y otro segundo módulo, que variará cada año pero se puede repetir aproximadamente cada cinco años. El primer módulo fue diseñado para recoger información sobre variables básicas como tamaño del hogar; profesión, edad, sexo, nivel de estudio y parentesco de los integrantes del hogar; y principalmente los ingresos del hogar. Con el segundo módulo se han ido estudiando aspectos más particulares como equipamientos del hogar, gastos familiares, nuevas tecnologías, uso y conocimiento del gallego, condiciones laborales y estado de salud de los integrantes del hogar o el cuidado de los menores de 12 años y personas dependientes.

La división por áreas que utiliza la ECV es una división que se utiliza en muchas encuestas del IGE. El origen de esta división territorial (que es una agrupación de comarcas y que veremos más detalladamente en el Capítulo 1) es por cuestiones económicas. Si en el diseño de la encuesta nos planteamos como objetivo ofrecer información a nivel de comarcas el tamaño de la muestra se incrementa mucho y por tanto el coste económico se hace inasumible, por esta razón se decidió definir unas áreas que fuesen agrupaciones de comarcas para las cuales se ofreciese información y que el coste económico no fuese elevado. Además aumentar el tamaño de muestra supondría mayores errores extramuestrales y un aumento en el tiempo de realización de la encuesta que ya por sí tiene una

duración aproximada de un trimestre.

El problema que nos encontramos después de realizar esta agrupación de comarcas es que la información que se ofrece por áreas a lo largo de los años no es suave y los estimadores obtenidos directamente de la ECV son poco precisos, **los estimadores indirectos**; de ahí la idea de utilizar técnicas de **estimación en áreas pequeñas** (SAE) que suavicen esta información obteniendo unos estimadores más precisos, los **estimadores indirectos**.

En la actualidad está siendo de gran interés la estimación en áreas pequeñas, principalmente de datos socio-económicos, donde los datos de los censos o registros administrativos no son suficientes, o en periodos intercensales. En esos casos se suelen utilizar valores de interés en otras áreas similares, valores anteriores de las mismas áreas o valores de otras variables relacionadas con la variable de interés. Los países que principalmente usan esta metodología son Australia, Canadá y Estados Unidos donde existen muchas poblaciones con tamaños de muestra muy pequeños.

Los estimadores indirectos obtenidos mediante métodos SAE que permiten introducir información auxiliar en el proceso de estimación, se denominan **estimadores basados en el modelo**; dichos modelos se dividen en dos grupos: **modelos de efectos fijos** si la variabilidad entre los dominios en la variable respuesta puede ser explicada completamente por las variables explicativas o **modelos de efectos aleatorios**, modelos mixtos (MM), si la variabilidad de cada área no puede ser completamente explicada por las variables explicativas surgiendo la necesidad de introducir efectos aleatorios al modelo.

En la actualidad los MM son cada vez más utilizados ya que son particularmente adecuados para estimaciones de áreas pequeñas por ser flexibles para combinar con eficacia las distintas fuentes de información y modelizar adecuadamente las distintas fuentes de error. Además suelen incorporar efectos aleatorios asociados a las áreas para explicar la variabilidad de los datos que no es recogida por parte de los efectos fijos del modelo. Pueden combinar efectos fijos y efectos mixtos de manera simultánea, esta es una de las principales características del éxito de estos modelos.

En el contexto de áreas pequeñas, los MM se dividen en dos grandes grupos en función del tipo de dato del que disponemos: **modelos a nivel de área**, datos agregados por áreas, y los **modelos a nivel de individuo**, datos disponibles a nivel de unidad.

En este trabajo con la ayuda de los modelos lineales mixtos (MML) trataremos de estimar datos socio-económicos de cada una de las veinte áreas de la Comunidad de Galicia que se detallarán en el Capítulo 1, en particular los ingresos medios por hogar en cada área.

Para ello, disponemos de los datos de la encuesta del 2014. Cada año a los encuestados en la ECV se les pregunta por los ingresos que tuvieron el año anterior. Por tanto, los datos de ingresos corresponden con el 2013.

Los datos que tenemos se encuentran disponibles a nivel de hogar, es decir, a nivel unidad; por lo que, en el Capítulo 1, además de explicar detalladamente los ingresos que estamos interesados en estudiar, explicaremos el procedimiento utilizado para agregar los ingresos en áreas. En ese primer capítulo también serán detalladas las áreas de interés y el muestreo utilizado para obtener los datos de la ECV.

En el capítulo 2 explicaremos la metodología de estimación en áreas pequeñas. En concreto la selección de modelos mediante algún criterio de selección de variables en áreas pequeñas, el AIC para modelos mixtos.

En el Capítulo 3 trataremos de aplicar dicha metodología a los datos reales de los ingresos. Como se tiene un tamaño de muestra pequeño, para estimar dichos ingresos será necesario utilizar ciertas variables explicativas de la ECV y de otras fuentes de información relacionadas con los ingresos para posteriormente poder elegir el mejor modelo mediante el AIC para MM.

Posteriormente, en el Capítulo 4, utilizaremos los resultados obtenidos con los datos del año 2013 para estimar los datos desde el 2007 al 2012; pudiendo así ver la evolución. Compararemos las estimaciones directas, las basadas en el modelo y las sintéticas.

En un último capítulo explicaremos las conclusiones obtenidas de los Capítulos 3 y 4.

Capítulo 1

Datos

Queremos estudiar la selección de modelos en áreas pequeñas mediante el AIC para estimar **el ingreso medio mensual por hogar** en cada una de las **veinte área de la Comunidad de Galicia**. Para ello utilizaremos los modelos lineales mixtos, en concreto los modelos a nivel de área ya que aunque dispongamos de datos a nivel de individuo necesitaremos de otras variables explicativas procedentes de otras fuentes de información diferentes a la ECV que solo se encuentran disponibles a nivel de área. Por parte del IGE no solo es de interés conocer las variables que serían capaces de explicar el ingreso medio mensual total, sino también los cuatro ingresos en los que se divide dicho ingreso. Por ello, también estudiaremos **el ingreso medio mensual procedente de ingresos por cuenta ajena**, **el ingreso medio mensual procedente de ingresos por cuenta propia**, **el ingreso medio mensual procedente de ingresos por prestaciones contributivas** y **el ingreso medio mensual procedente de ingresos por prestaciones no contributivas**. Estos cinco tipos de ingresos medios se estudiarán por hogar en las veinte áreas en las que se divide la Comunidad de Galicia.

En la primera sección de este capitulo explicaremos cuales son las veinte áreas que utiliza el IGE para dividir la Comunidad Autónoma de Galicia. Posteriormente, en la Sección 2.2, haremos una explicación detallada de las variables de interés (variables respuesta), del tipo de muestreo utilizado para la ECV, como se han obtenido las estimaciones directas y la medida de error muestral para esas estimaciones. Por último veremos los tamaños de muestra utilizados para calcular las estimaciones directas de las cinco variables de interés.

1.1. Áreas de estudio

Antes de explicar detalladamente las variables de interés y el muestreo utilizado para obtener sus estimaciones directas, debemos conocer cuales son las áreas en las que se ha dividido la Comunidad de Galicia para realizar el muestreo. Además estas veinte áreas también serán de interés ya que son las áreas en las que queremos conocer cada

variable de interés.

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 de A Costa da Morte: comprende las comarcas de Bergantiños, Fisterra, Muros, Terra de Soneiras y Xallas
4. A Barbanza - Noia: comprende as comarcas de A Barbanza y Noia.
5. Área de A Coruña: comprende las comarcas de A Coruña y Betanzos.
6. Área de Santiago: comprende las comarcas de A Barcala, O Sar y Santiago.

Provincia de Lugo

1. Lugo sur: comprende las comarcas de Chantada, Quiroga y Terra de Lemos.
2. Lugo oriental: comprende las comarcas de A Fonsagrada, Os Ancares y Sarria.
3. Lugo central: comprende las comarcas de Ulloa, Lugo, Meira y A Terra Chá.
4. A Mariña: comprende las comarcas de A 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 y Maceda, Terra de Caldelas, Terra de Trives y Valdeorras.
3. Ourense sur: comprende las comarcas de A Limia, Baiza Limia, Terra deCelanova, Verín y Viana.
4. Área de Ourense: comprende la comarca de Ourense.

Provincia de Pontevedra

1. Pontevedra nororiental: comprende las comarcas de Dez y Tabeiros.Terra de Montes.
2. Pontevedra sur: comprende las comarcas de A Paradanta, O Baixo Miño y Condado.
3. Caldas - O Salnés: comprende las comarcas de Cladas y O Salnés.
4. O Morrazo: compre 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 podemos ver gráficamente donde se encuentra cada área en particular.

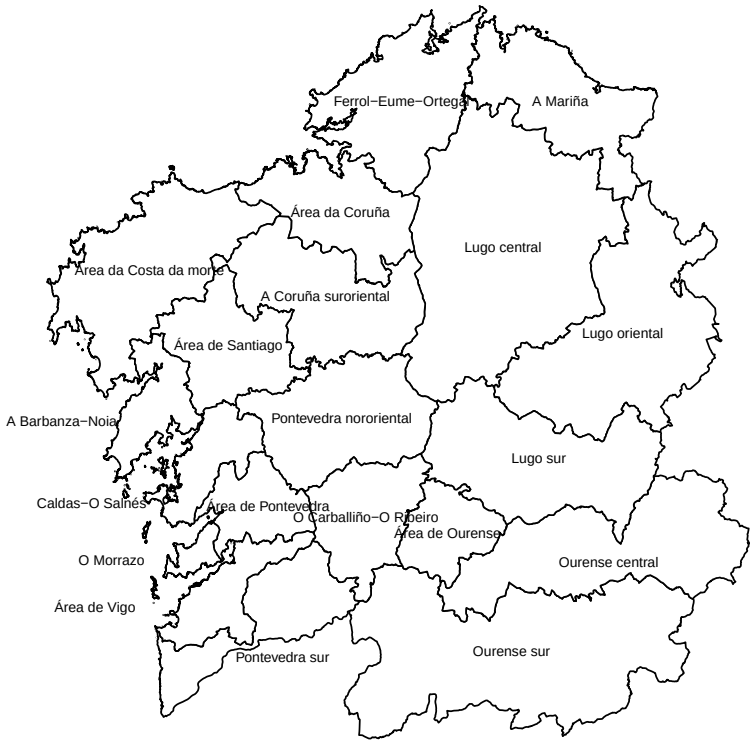

Figura 1.1: Mapa de la Comunidad de Galicia dividida en las 20 áreas.

1.2. Descripción de los datos

En esta Sección definiremos detalladamente cuales son las variables de estudio, pero antes debemos definir que se conoce como hogar.

Hogar: persona o conjunto de personas que ocupan en común una vivienda princi-

pal o parte de ella, y que consumen y/o comparten alimentos o bienes con cargo a un mismo presupuesto.

Las variables de las que estamos interesados en conocer su estimación están disponibles a nivel de hogar, nivel de unidad; pertenecen a los datos del 2013 y son las siguientes:

- **IMTOT**: media mensual de los ingresos netos monetarios de todos los miembros del hogar en el año anterior al de la encuesta. Tenemos un tamaño de muestra de 9216 respecto al cerca del millón de hogares que había en el año 2013.
- **IM_AJENA**: ingreso medio mensual percibido por trabajo por cuenta ajena. Tenemos un tamaño de muestra de 4540.
- **IM_PROPIA**: ingreso medio mensual percibido por trabajo por cuenta propia. Tenemos un tamaño de muestra de 1715.
- **IM_CONTRIB**: ingreso medio mensual percibido por pensiones ya sean de España o del extranjero. Tenemos un tamaño de muestra de 5149.
- **IM_NO_CONTRIB**: ingreso medio mensual percibido por prestaciones o subsidios por desempleo, rentas, becas u otras fuentes de ingreso. Tenemos un tamaño de muestra de 3127.

Todas estas variables se han obtenido de la ECV. El muestreo utilizado para elegir la muestra variará en función del número de habitantes que tenga el municipio al que pertenece el hogar y el marco que se emplea es el Padrón de habitantes. Por ello, cada una de las veinte áreas en las que se divide Galicia están a su vez divididas en estratos de acuerdo a la siguiente clasificación:

- **Estrato 0:** municipios autorrepresentados, con más de 50000 habitantes (A Coruña, Ferrol, Santiago de Compostela, Pontevedra, Vigo, Ourense y Lugo).
- **Estrato 1:** municipios de más de 20000 habitantes.
- **Estrato :** municipios de más de 20000 habitantes.
- **Estrato 2:** municipios de 15000 a 20000 habitantes.
- **Estrato 3:** municipios de más 10000 a 15000 habitantes.
- **Estrato 4:** municipios de 5000 a 10000 habitantes.
- **Estrato 5:** minucipios de menos de 5000 habitantes.

En los municipios corespondientes al estrato 0 el muestreo es unietápico: se ordenan los hogares según características sociodemográficas y, a continuación, se escoge la muestra mediante muestreo sistemático con arranque aleatorio.

Para el resto de los estratos de las distintas áreas el muestreo es bietápico: en primer lugar se ordenan las secciones censales de ese estrato según características sociodemográficas y después se escoge la muestra de secciones mediante muestreo sistemático con arranque aleatorio. En una segunda etapa, para la muestra de secciones ya escogidas, se ordenan los hogares de esas secciones mediante esas características sociodemográficas y , a continuación, se escoge la muestra de hogares mediante muestreo sistemático con arranque aleatorio.

Para los datos del año 2013 se utilizaron 512 secciones. En A Coruña 180, Lugo 90, Ourense 91, Pontevedra 151. En cada sección se entrevistaron a 18 viviendas, con lo que resulta un total de 9216 viviendas.

Una vez obtenidos los resultados a nivel de hogar, se quiere calcular cada ingreso a nivel de área obteniendo así las estimaciones directas para cada ingreso, las denotaremos por y_{id} con $i = 1, ..., 5$ los cinco tipos de ingresos y $d = 1, ..., 20$ área al que pertenece ese tipo de ingreso.

Cada hogar j no tiene el mismo peso para todas las áreas, dependerá del tamaño del hogar, de la edad y sexo de las personas de cada hogar y área, y de la distribución poblacional utilizada procedente de proyecciones de población especialmente elaboradas para la ECV. Lo denotaremos por w_{jd}. De modo que las estimaciones directas para cada área, y_{id}, serán estimaciones ponderadas dadas por la siguiente expresión:

$$y_{id} = \frac{\sum_{j=1}^{nd} w_{jd}\, y_{ijd}}{\sum_{j=1}^{n_d} w_{jd}} = \frac{\sum_{j=1}^{nd} w_{jd}\, y_{ijd}}{n_d} \qquad d = 1, .., 20 \qquad i = 1, ..., 5. \tag{1.1}$$

con j cada hogar, w_{jd} el peso para cada hogar en el área d, y_{ijd} valor de cada ingreso i obtenido de la ECV para cada hogar j en el área d y n_d número de hogares en cada área.

Obtendremos las siguientes estimaciones y_{1d}, y_{2d}, y_{3d}, y_{4d} e y_{5d} para cada área $d = 1, ..., 20$. Recibirán respectivamente el nombre de **IMTOT**, **IM_AJENA**, **IM_PROPIA**, **IM_CONTRIB** e **IM_NO_CONTRIB**.

Una vez obtenidas las estimaciones directas para cada área (y_{1d}, y_{2d}, y_{3d}, y_{4d} e y_{5d}), una forma de medir su precisión será con el error de muestreo. Cuánto más pequeño sea éste más precisas serán las estimaciones. Para estimar estos errores de muestreo para cada uno de los cinco ingresos se utiliza un método de remuestreo Jackknife. El estimador Jackknife de cada ingreso i para cada área d, viene dado por la siguiente expresión:

$$\hat{V}(y_{id}) = \sum_{h=1}^{L_d} \frac{S_h^{-1}}{S_h} \Big(\sum_{s=1}^{S_h} (y_J^s - \hat{y}_{Jh})^2 \Big) \qquad \forall i = 1, ..., 5 \qquad d = 1, ..., 20.$$

siendo L_d el número de estratos de área d, S_h el número de secciones muestrales en el estrato h, y_J^s el estimador obtenido después de suprimir de la muestra la sección s, $\hat{y}_{Jh}$ la media de los estimadores y_J^s correspondiente al estrato h.

Una aproximación del error de muestreo es el coeficiente de variación en porcentaje:

$$CV = \frac{\sqrt{\hat{V}(y_{id})}}{\hat{y}_{id}} \cdot 100.$$

Cuanto menor sea este coeficiente de variación más precisas serán las estimaciones directas.
En la parte práctica este coeficiente de variación será denotado por CV_1.

1.3. Tamaños de muestra de las áreas

Por último veremos el tamaño de muestra que se tiene para calcular las estimaciones directas de cada una de las cinco variables de interés en las veinte áreas. Para ello haremos una tabla con las cinco variables de interés en las que aparecerá el mínimo, máximo, primer y tercer cuartil, media y mediana de los tamaños de muestra para las veinte áreas de estudio.

	Mínimo	Máximo	Primer cuartil	Tercer cuartil	Media	Mediana
IMTOT	284	1059	287,8	597,8	459,4	341
IM_AJENA	98	621	123	295,2	227	153
IM_PROPIA	37	186	55,75	98,5	85,75	80
IM_CONTRIB	152	496	189,8	281,2	257,4	207
IM_NO_CONTRIB	57	383	99	192,2	156,4	122

Cuadro 1.1: Descriptiva de los tamaños de muestra para las 20 áreas en cada tipo de ingreso.

Como se observa en el Cuadro 1.1. tenemos que el mínimo tamaño de muestra es 37. Además hay mucha diferencia entre el mínimo y el máximo para las cinco variables. Las variables donde tenemos menos tamaño de muestra son IM_AJENA, IM_PROPIA e IM_NO_CONTRIB. Concluimos que, en general, tenemos tamaños de muestra pequeños para muchas áreas ya que algunas variables tienen el mínimo y el primer cuartil con valores muy bajos lo que podría hacer que las estimaciones directas fuesen poco precisas en esas áreas. Por lo que, en los siguientes capítulos se hará uso de las estimaciones en áreas pequeñas para obtener unos estimadores más precisos que los directos,se calcularán los estimadores basados en el modelo para las cinco variables de interés.

Capítulo 2

Metodología

En este segundo capítulo haremos primero una pequeña introducción para diferenciar los dos modelos en los que se dividen los MM cuando hablamos de estimación en áreas pequeñas: modelos a nivel de área y modelos a nivel de individuo. Después construiremos el modelo de Fay-Herriot, modelo a nivel de área en el contexto de áreas pequeñas; además, estimaremos los parámetros de dicho modelo para poder, finalmente, estimar la verdadera media de los datos en cada área, objetivo del trabajo.

Por último, estudiaremos la selección de modelos mediante el AIC que nos permitirá saber qué covariables necesitamos en el modelo y cuáles podrán ser no significativas en el modelo completo. A la hora de hacer inferencia los modelos pueden ser tratados de dos formas distintas según cuál sea el objetivo del investigador: un enfoque condicional y un enfoque marginal. Para la construcción del AIC, en un enfoque marginal usaremos la verosimilitud marginal y en el enfoque condicional usaremos la verosimilitud condicional.

Toda la metodología que explicaremos en este capítulo será en los siguientes capítulos aplicada de forma práctica a unos datos reales. En concreto, se tratará de estudiar el ingreso medio mensual por hogar en las veinte áreas territoriales gallegas.

2.1. Modelos a nivel del area vs. modelos a nivel de individuo.

En esta primera sección daremos una visión general de los dos tipos de modelos más utilizados en estimación de áreas pequeñas y luego en la Sección 2.2. se explicará en particular el modelo con el que trabajaremos en el Capítulo 3 con los datos reales.

Tomando como referencia el artículo de *Pfeffermann (2013)*, dividiremos los modelos lineales mixtos en dos grandes grupos:

1. **Modelos a nivel de individuo**: son aquellos en los que se dispone de información

auxiliar sobre las unidades individuales de la población. Estos modelos se pueden expresar como:

$$y_{dj} = X_{dj}\boldsymbol{\beta} + z_{dj}u_d + \epsilon_{dj} \qquad \forall d = 1, ..., D, j = 1, ..., n_d$$

donde d denota el área, j el individuo en el área, y_{dj} el valor de la variable respuesta de cada uno de los j individuos en el área d, X_{dj} matriz de diseño formada por información auxiliar conocida, $\boldsymbol{\beta}$ parámetro de regresión fijo, z_{dj} peso que tiene el individuo j en el área d, u_d efecto aleatorio en el área d con distribución $N(0, \sigma_u^2)$, σ_u^2 desconocida y ϵ_{dj} vector aleatorio de errores desconocidos los cuales siguen una distribución $N(0, \sigma_{\epsilon d}^2)$, $\sigma_{\epsilon d}^2$ también desconocida. En este modelo los parámetros a estimar serán $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2, \sigma_{\epsilon d}^2)$.

Tomando como referencia el artículo de *Jiang y Lahiri* (2006), los modelos a nivel de individuo fueron primeramente usados por *Battese et al.* (1988) para estimar las áreas plantadas con maíz y soja en 12 países (áreas pequeñas con tamaños de muestra entre 1 y 6) de Iowa Norte-Central, EEUU. Combinaron la información de un conjunto de encuestas y datos de satélite que proporcionan información de píxeles de terreno plantados de maíz y soja en cada país. El pequeño tamaño de las muestras hacía que la estimación de los parámetros por los métodos usuales de muestreo fuese poco precisa; además asumiendo una constante del modelo fija para todos los países no se tendría en cuenta la variabilidad adicional de cada país. Por lo que el modelo usado para conocer el número de áreas plantadas por maíz y soja en Iowa fue el siguiente:

$$y_{ij} = \beta_{0i} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \epsilon_{ij}$$

donde i denota el país ($i = 1, ..,6$), j es el tamaño de muestra en cada país, y_{ij} es el número de hectáreas de maíz y soja en el j-ésimo segmento del i-esimo país según la encuesta, $X_{1ij}(X_{2ij})$ número de píxeles clasificados como maíz (soja) en el j-ésimo segmento del i-ésimo paíz, β_{0i} constante del modelo en cada país específico que son asumidos i.i.d con media β_0 y varianza σ_u^2 y, por último, ϵ_{ij} errores i.i.d con media cero y varianza $\sigma_{\epsilon d}^2$.

Otro caso importante donde ha sido utilizado este tipo de modelos es en el análisis de datos longitudinales. Para más información puede consultarse el artículo de *Diggle et al.* (1996).

2. **Modelos a nivel de área**: son aquellos en los que se dispone de información auxiliar sólo a nivel de área, tenemos datos agregados, principalmente por motivos de privacidad.

   Para estudiar este tipo de modelos usaremos **el modelo de Fay -Herriot** diseñado por *Fay y Herriot* (1979) para estimar la renta per-cápita para poblaciones

con tamaños de muestra inferiores a 1000 en los condados de Estados Unidos. El modelo usado fue el siguiente:

$$y_d = X_d\boldsymbol{\beta} + z_d u_d + \epsilon_d \qquad d = 1, ..., D$$

donde d denota el condado donde se realiza el estudio y D el número total de condados, y_d es el estimador de interés (por ejemplo la renta media en cada área), X_d matriz de diseño, $\boldsymbol{\beta}$ vector de efectos fijos del modelo, z_d peso que tiene el área d respecto al resto de áreas, u_d efecto aleatorio para cada área d siendo independientes e idénticamente distribuidos $N(0, \sigma_u^2)$, σ_u^2 desconocida y ϵ_d error muestral en cada área d siendo independientes e identicamente distribuidos $N(0, \sigma_{\epsilon d}^2)$, $\sigma_{\epsilon d}^2$ asumida conocida; además u_d son independientes de ϵ_d. Por tanto, los parámetros a estimar en este modelo son $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2)$.

Puesto que este será nuestro modelo de partida, se tratará en detalle en la Sección 2.2

2.2. Modelo de Fay-Herriot

En esta sección se tratará en más detalle los modelos de Fay-Herriot también llamados modelos a nivel de área ya que nuestros datos de trabajo, introducidos en el Capítulo 1, se encuentran disponibles de forma agregada por áreas. Por tanto el modelo de Fay-Herriot es nuestro modelo de trabajo.

La construcción del modelo introducido por *Fay y Herriot* (1979) sigue dos etapas:

1. Sea una población finita dividida en D áreas, con n_d tamaño de muestra en cada área, μ_d el parámetro de interés en cada área y y_d ($y_d = \bar{y}_d$) el estimador directo de μ_d ($d = 1, ...D$). Se quiere estimar μ_d por y_d, cuando se estima muestralmente una característica se comete un error, el error muestral; por tanto, podemos definir el estimador directo de μ_d como:

   $$y_d = \mu_d + \epsilon_d \qquad d = 1, ..., D$$

   con ϵ_d error muestral en cada área con distribución $N(0, \sigma_{\epsilon d}^2)$, $\sigma_{\epsilon d}^2$ asumida conocida. Por tanto:

   $$y_d|\mu_d \sim N(0, \sigma_{\epsilon d}^2).$$

2. Se supone que la verdadera media en cada área, μ_d, se puede expresar como una combinación lineal de los elementos fijos de $\boldsymbol{\beta}$ y de los elementos aleatorios de u_d:

   $$\mu_d = X_d\boldsymbol{\beta} + z_d u_d \qquad d = 1, ..., D$$

   con X_d matriz de diseño, $\boldsymbol{\beta}$ vector de parámetros de regresión fijos, z_d peso que tiene el área d y u_d efecto aleatorio en el área d el cual sigue una distribución $N(0, \sigma_u^2)$, σ_u^2 desconocida. Por tanto,

   $$\mu_d \sim N(X_d\boldsymbol{\beta}, \sigma_u^2).$$

Uniendo las dos fases del proceso, podemos expresar el modelo de Fay-Herriot como:

$$y_d = X_d\boldsymbol{\beta} + z_d u_d + \epsilon_d \qquad d = 1, ..., D \tag{2.1}$$

Los parámetros a estimar en este modelo son $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2)$.

Observación 2.1: En la práctica $\sigma_{\epsilon d}^2$ es muy difícil que sea conocida. Por lo que, aunque para los posteriores desarrollos se suponga conocida debe ser estimada con la muestra, se toma la varianza muestral. La varianza del error puede ser constante a lo largo de todas las áreas, el caso más simple, o bien variar con el área que será nuestro caso. En este último caso el estimador de la varianza del error será:

$$\hat{\sigma}_{\epsilon d}^2 = \frac{1}{n_d}\sum_{i=1}^{n_d}(y_{id} - y_d)^2 \qquad \forall d = 1, ..., D$$

con y_{id} individuo i en el área d, y_d estimación directa de la media poblacional μ_d y n_d tamaño de muestra en cada área d.

Observación 2.2: De forma matricial el modelo se define como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZU} + \boldsymbol{\epsilon}$$

con $\mathbf{y} = (y_1, ..., y_D)'$ vector objetivo en cada área, variable respuesta; $\mathbf{X} = \mathbf{X}_{DxP}$ y $\mathbf{Z} = \mathbf{Z}_{DxD}$ matrices de diseño, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ vector de efectos fijos, p número de covariables, $\mathbf{U} = (u_1, ..., u_D)'$ vector de efectos aleatorios compuesto por el efecto aleatorio para cada área y sigue una distribución $N(\vec{0}, \Sigma_u)$ con $\Sigma_u = \sigma_u^2\mathbf{I}$ matriz de varianzas-covarianzas de los efectos aleatorios, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_D)$ vector de errores asociados a cada área con distribución $N(\vec{0}, \Sigma_\epsilon)$, $\Sigma_\epsilon = \sigma_{\epsilon d}^2\mathbf{I}$ matriz de varianzas-covarianzas del error muestral, $\mathbf{I}$ matriz identidad y $\boldsymbol{V} = \boldsymbol{V}(\nu) = \text{Var}\{\mathbf{y}\} = \text{Var}\{\mathbf{X}\boldsymbol{\beta} + \mathbf{ZU} + \boldsymbol{\epsilon}\} = \text{Var}\{\mathbf{ZU} + \boldsymbol{\epsilon}\} = \mathbf{Z}\Sigma_u\mathbf{Z}' + \Sigma_\epsilon$ varianza del modelo con $\nu = (\sigma_{\epsilon d}, \sigma_u)$ vector de las componentes de la varianza.

Para la estimación del modelo trabajaremos con el modelo definido de esta forma.

Observación 2.3: El modelo de Fay-Herriot definido matricialmente como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZU} + \boldsymbol{\epsilon}$$

en nuestro caso podría ser tratado como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U} + \boldsymbol{\epsilon}$$

ya que $\mathbf{Z}$ corresponde con la matriz identidad, lo que indicaría 1 si el elemento está en el área y 0 si no pertenece al área. Este es el modelo habitual de trabajo.

2.2.1. Estimación del modelo

Una vez introducido el modelo de Fay-Herriot, el siguiente paso es estimar los parámetros del modelo para poder estimar la verdadera media en cada área, objetivo de estos modelos, que denotamos por $\mu_d = X_d\beta + u_d$.

Tomando como referencia los artículo de *Prasad y Rao* (1990) y de *Gonzalez-Manteiga et al.* (2008) la estimación se podría llevar a cabo en dos fases:

1. Estimación de $\boldsymbol{\beta}$ y de los efectos aleatorios $\mathbf{U}$.
2. Estimación de la varianza de y, varianza de la variable respuesta, compuesta por la varianza de los efectos aleatorios y la varianza del error muestral. La varianza del error se asumía conocida.

Fase I

Lo primero que tenemos que hacer es estimar $\boldsymbol{\beta}$:

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

como vemos $\boldsymbol{\beta}$ depende de $\mathbf{V}$ que es desconocida. $\widetilde{\boldsymbol{\beta}}$ es el BLUE (mejor estimador lineal insesgado) estimado mediante mínimos cuadrados generalizados.

Después estimamos el vector de efectos aleatorios que claramente va a depender de su varianza la cual es desconocida; y además también de $\boldsymbol{\beta}$ de la que conocemos su estimador pero no su estimación. El BLUP (mejor predictor lineal insesgado) de $\boldsymbol{U}$ viene dado por (ver *Rao* (2003)):

$$\widetilde{\boldsymbol{U}} = E[\mathbf{U}|\mathbf{y}, \boldsymbol{\beta}] = \Sigma_u\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Fase II

Como vimos en la Fase I, $\boldsymbol{\beta}$ y $\mathbf{U}$ dependen de ciertos parámetros desconocidos, Σ_u, y, en consecuencia, $\mathbf{V}$; los cuales deben ser estimados mediante algunos de los métodos nombrados posteriormente en la siguiente subsección. Una vez estimados se introducen en $\widetilde{\boldsymbol{\beta}}$ y $\widetilde{\boldsymbol{U}}$ de modo que obtenemos los estimadores empíricos, EBLUE en el caso de $\boldsymbol{\beta}$ y EBLUP en el caso de $\mathbf{U}$:

$$\widetilde{\boldsymbol{\beta}}(\hat{\Sigma}_u, \hat{\boldsymbol{V}}) = \hat{\boldsymbol{\beta}}, \qquad \widetilde{\boldsymbol{U}} = (\hat{\Sigma}_u, \hat{\boldsymbol{V}}) = \hat{\boldsymbol{U}}.$$

Una vez conocidas $\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{U}}$, las reemplazamos y obtenemos que el EBLUP, BLUP empírico, de $\boldsymbol{\mu}$ es:

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{U}}.$$

Observación 2.4: Si la varianza de los efectos aleatorios fuese conocida, algo bastante improbable en la práctica, en la estimación del modelo solo tendríamos la Fase I y $\hat{\boldsymbol{\mu}}$ sería el BLUP de $\boldsymbol{\mu}$.

Estimación de las componentes de la varianza

Tomando como referencia el artículo de *Jiang y Lahiri* (2006) para estimar las componentes de la varianza se conocen, principalmente, tres métodos: ANOVA, REML (máxima verosimilitud restringida) y MLE (máxima verosimilitud).

El método ANOVA fue uno de los primeros métodos para la estimación de las componentes de la varianza pero se comprobó que no era muy eficiente; por lo que, generalmente se usan los otros dos métodos, que aunque fueron desarrollados bajo la hipótesis de normalidad, en la vida real afortunadamente esa hipótesis puede ser violada lo que significa que son métodos robustos ante la falta de normalidad. *Lange y Rian* (1989) mostraron ejemplos en los que se había usado MLE y REML para la estimación de las componentes de la varianza donde los efectos aleatorios no eran normales.

El procedimiento de MLE tiene su origen con *Fisher* (1925). Fue aplicado por primera vez al modelo mixto por *Hartley y Rao* (1967). Supuesto el modelo lineal mixto

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZU} + \boldsymbol{\epsilon}$$

con $\mathbf{E} = E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ y con verosimilitud

$$L = (2\pi)^{-2n}|\mathbf{V}|^{-1/2} \exp\Big\{ -\frac{1}{2}[(\mathbf{y}-\mathbf{E})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{E})]\Big\}.$$

El método consiste en maximizar L respecto a $\boldsymbol{\beta}$ y a $\mathbf{V}$.

En general no existen expresiones explícitas para los estimadores MLE por ser ecuaciones no lineales en los parámetros a estimar; de modo que, los estimadores se obtienen por métodos iterativos basados en derivadas. Destaca el "método-scoring", método de Newton-Raphson modificado.

Se observó que el estimador ML suele producir estimaciones sesgadas de la varianza porque no tiene en cuenta los grados de libertad, que se pierden al estimar la media. Para evitar este problema surgió la idea de los estimadores REML.

En el caso del procedimiento REML propuesto por *Patterson y Thompson* (1971) consiste en factorizar la verosimilitud completa en dos partes independientes, una de las cuales no contiene la media, asumiendo que por usar esta parte de la verosimilitud no se pierde información con respecto a usar la verosimilitud completa. La verosimilitud restringida se corresponde con la verosimilitud asociada a una combinación lineal de las observaciones, es decir, suponemos que:

$$L\mathbf{y} = L\mathbf{X}\boldsymbol{\beta} + L\mathbf{ZU} + L\boldsymbol{\epsilon}$$

cuya media es nula, es decir $E(L\mathbf{y}) = E(L\mathbf{X}\boldsymbol{\beta} + L\mathbf{ZU} + L\boldsymbol{\epsilon}) = L\mathbf{X}\boldsymbol{\beta} = 0$ y cumple las condiciones mencionadas anteriormente (ser un factor independiente del otro con el

que se reproduce la verosimilitud completa y no suponer pérdida de información con respecto a usar los datos originales).

Los estimadores REML son insesgados para la varianza en modelos mixtos donde se satisface la hipótesis de normalidad para los errores y efectos aleatorios, en la práctica.

Los desarrollos de las expresiones obtenidas con estos métodos para la estimación de las componentes de la varianza pueden consultarse en *Rao* (2003).

2.3. AIC

La selección del modelo óptimo entre múltiples candidatos representa un papel importante para interpretar adecuadamente los datos. Esta selección se puede realizar mediante el uso de algún criterio entre los que destacan los criterios de información de Akaike (*Akaike* (1974)), AIC, y el Bayesiano de Schwarz (*Schwarz* (1978)), BIC:

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

$$\text{BIC} = -2log(\hat{\theta}) + Klog(D)$$

con $\hat{\theta}$ máximo valor de función de verosimilitud para el modelo estimado, K número de parámetros a estimar en el modelo y D número de áreas. Ambos criterios son funciones del logaritmo de la verosimilitud pero se diferencian por el valor de penalización.

Estos criterios fueron desarrollados, originalmente, para los modelos lineales clásicos y luego adaptados a los modelos lineales mixtos. Este trabajo se centrará en el uso del AIC introducido por *Akaike* (1974) y extendido a los modelos mixtos por *Vaida y Blanchard* (2005).

Tomando como referencia el artículo de *Vaida y Blanchard* (2005), el AIC es una verosimilitud penalizada basada en modelos utilizando la diferencia de Kullback-Leibler la cual mide la diferencia entre la verdadera densidad de la distribución generada por los datos, f, y el modelo de aproximación para evaluar los datos, g_θ. Valores más bajos de esa distancia corresponden a una mejor aproximación de f por g_θ; por eso, cuánto menor sea el AIC de un modelo mejor será el modelo, siempre que se cumplan las hipótesis del modelo.

El AIC usual viene dado por:

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

siendo $\hat{\theta}$ máximo valor de función de verosimilitud para el modelo estimado y K número de parámetros a estimar en el modelo (parámetros fijos).

El AIC tradicional introducido por *Akaike* (1974) no tiene en cuenta los efectos aleatorios de los modelos mixtos. Por ello, *Vaida y Blanchard* (2005) proponen un nuevo AIC que sí tenga en cuenta esos efectos aleatorios. El AIC para modelos mixtos tiene distintas expresiones atendiendo a la log-verosimilitud considerada y al término de penalización considerado. En modelos mixtos se trabaja con dos versiones de verosimilitud, marginal y condicional. Por tanto hablaremos de un AIC marginal, mAIC, y un AIC condicional, cAIC. El término de penalización considerado en ambos será el mismo.

En la siguiente subsección se explicarán detalladamente cada enfoque, centrándonos en el enfoque condicional ya que estamos interesados en estudiar cada área en particular y no la población en general. Posteriormente se explicará el valor de la penalización.

2.3.1. Enfoque marginal vs. enfoque condicional

A la hora de hacer inferencia en modelos lineales mixtos el AIC puede adquirir una forma u otra en función de cuál sea el objetivo de interés. Esto se manifestará en la verosimilitud que usemos.

Por ejemplo, en un ensayo clínico para probar el efecto de un nuevo tratamiento frente al tratamiento estandar, las personas están inscritas dentro de cada hospital (área) d. El enfoque marginal se centra en el efecto global del tratamiento, es decir en β, en cambio el enfoque condicional se centra en conocer el efecto del tratamiento en cada hospital d, es decir en $\beta + u_d$.

1. **Enfoque marginal**: inferencia relativa a los parámetros de la población. Foco en los efectos fijos. Los efectos aleatorios no son de interés en sí mismos sino que son un dispositivo para el modelado de la correlación de la variable respuesta en cada área.

   Sea el modelo de Fay-Herriot:

   $$y_d = X_d\boldsymbol{\beta} + u_d + \epsilon_d$$

   con un enfoque marginal podemos decir que y_d sigue una distribución:

   $$y_d \sim N(X_d\boldsymbol{\beta}, \sigma_u^2 + \sigma_{\epsilon d}^2).$$

   Calculando su esperanza y varianza tenemos:

   $$\mu_d = E(y_d) = X_d\boldsymbol{\beta}$$

   $$v_d = Var(y_d) = Var(X_d\boldsymbol{\beta} + u_d + \epsilon_d) = Var(u_d) + Var(\epsilon_d) = \sigma_u^2 + \sigma_{\epsilon d}^2$$

   con $u_d \sim N(0, \sigma_u^2)$ y $\epsilon_d \sim N(0, \sigma_{\epsilon d}^2)$ independientes e idénticamente distribuidos.

Por tanto, la función de densidad necesaria para construir el AIC marginal, mAIC, viene dada por:

$$f_\theta(y) = \prod_{d=1}^{D} f_\theta(y_d) = \prod_{d=1}^{D} f(\mu_d, v_d) =$$

$$= \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi v_d}}\, exp\Big\{ -\frac{1}{2}[(y_d - X_d\boldsymbol{\beta})' v_d^{-1}(y_d - X_d\boldsymbol{\beta})]\Big\}.$$

Luego la verosimilitud marginal es:

$$log(f(\boldsymbol{\mu}, \boldsymbol{V})) = \prod_{d=1}^{D} log(f(\mu_d, v_d)) = \tag{2.2}$$

$$= -\frac{D}{2} log(2\pi) - \frac{1}{2}\sum_{d=1}^{D} log\{v_d\} - \frac{1}{2}\sum_{d=1}^{D}[(y_d - X_d\boldsymbol{\beta})'\, v_d^{-1}(y_d - X_d\boldsymbol{\beta})].$$

2. **Enfoque condicional**: inferencia sobre los parámetros específicos del área. Foco en los efectos aleatorios. Los efectos aleatorios sí que son de interés en sí mismos, de hecho son parámetros a estimar.

   En este caso el parámetro de interés sigue la distribución:

   $$y_d|u_d \sim N(X_d\boldsymbol{\beta} + u_d, \sigma^2_{\epsilon d}),$$

   pues calculando la esperanza y varianza se tiene:

   $$\mu_d = E[y_d|u_d] = X_d\boldsymbol{\beta} + u_d$$

   $$v_d = Var[y_d|u_d] = \sigma^2_{\epsilon d}.$$

   Para construir el cAIC usaremos la verosimilitud condicional, es decir, sustituir $X_d\boldsymbol{\beta}$ por $X_d\boldsymbol{\beta} + u_d$ y v_d también cambia por $\sigma^2_{\epsilon d}$ en la ecuación 2.2:

   $$log(f(\boldsymbol{\mu}, \boldsymbol{V})) = log(\prod_{d=1}^{D} f(\mu_d, v_d)) =$$

   $$= -\frac{D}{2} log(2\pi) - \frac{1}{2}\sum_{d=1}^{D} log\{v_d\} - \frac{1}{2}\sum_{d=1}^{D}[(y_d - X_d\boldsymbol{\beta} - u_d)'\, v_d^{-1}(y_d - X_d\boldsymbol{\beta} - u_d)] =$$

   $$= -\frac{D}{2} log(\sqrt{2\pi}) - \frac{1}{2}\sum_{d=1}^{D} log\{\sigma^2_{\epsilon d}\} - \frac{1}{2}\sum_{d=1}^{D}[(y_d - X_d\boldsymbol{\beta} - u_d)'\sigma^{-2}_{\epsilon d}(y_d - X_d\boldsymbol{\beta} - u_d)].$$

Trabajaremos siempre con el cAIC ya que, principalmente, el objetivo que tenemos es analizar cada área y no la población en general.

2.3.2. Término de penalización, K

El valor de la penalización, K, ha sido estudiado por múltiples autores y tiene un valor u otro según cuál sea el método de estimación de los parámetros del modelo y si la varianza del efecto aleatorio es conocida o no.

Los primeros en proponer un valor para K estableciendo la hipotesis de varianza del efecto aleatorio conocida para modelos mixtos fueron *Vaida y Blanchard* (2005). Tomaron como valor de K:

$$K = \rho = \text{traza}(H)$$

donde H es la "*matriz hat*" del vector de observaciones $\boldsymbol{y}$ en el vector evaluado $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}$ tal que $\hat{\mu}_d = Hy_d$. H es una matriz de dimensión DxD, simétrica, también denominada matriz de proyecciones ya que se puede considerar que la predicción $\hat{\boldsymbol{\mu}}$ es la proyección de $\boldsymbol{y}$; su estimación viene dada por:

$$\hat{H} = (1 - \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1})\Big(\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\Big) + \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}$$

ya que

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}} = \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} + \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) =$$
$$= \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} + \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}) =$$
$$= \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} + \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} - \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} =$$
$$= \Big[\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1} + \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1} - \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\Big]\boldsymbol{y} =$$
$$= \Big[(1 - \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1})\Big(\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\Big) + \hat{\boldsymbol{\Sigma}}_u\hat{\boldsymbol{V}}^{-1}\Big]\boldsymbol{y}$$

Como la hipotesis propuesta por *Vaida y Blanchard* (2005), Σ_u conocida, era muy restrictiva y bastante improbable en la práctica, *Liang et al.* (2008) proponen un K genérico cuando Σ_u es desconocida:

$$K = \sum_{d=1}^{D} \frac{\partial \hat{\mu}_d}{\partial y_d} = \text{traza}\Big(\frac{\partial \hat{\boldsymbol{\mu}}'}{\partial \boldsymbol{y}}\Big).$$

Como se observa, para modelos lineales mixtos ese valor de K va a depender del vector de observaciones $\boldsymbol{y}$. Para obtener ese valor del parámetro de penalización, K, mediante derivadas se puede calcular directamente o aproximando numéricamente por:

$$\frac{\partial \hat{\mu}_d}{\partial y_d} \approx \frac{\hat{\boldsymbol{\mu}}_d(h) - \boldsymbol{y}_d}{h} \qquad h \to 0.$$

Han (2013) propuso una aproximación para el K de *Liang et al.* (2008) en los casos

particulares de estimar los parámetros del modelo por ML y REML. El K utilizado es igual a:

$$K = \rho - C$$

donde $\rho = \text{traza}(H)$ y C es una cantidad debida a la estimación de $\boldsymbol{\Sigma}_u$. El valor de C variará si hemos estimado los parámetros por ML o por REML:

$$C = 2\,\boldsymbol{T}^{-1}\boldsymbol{r}_\epsilon\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{P}}^*\hat{\boldsymbol{\Sigma}}_\epsilon\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{P}}^*\boldsymbol{r}_\epsilon.$$

Si estimamos por ML:

$$\boldsymbol{T} = \text{traza}\Big((\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{P}}^*)^2\Big) - 2\boldsymbol{r}_\epsilon'\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{P}}^* r_\epsilon.$$

Si estimamos por REML:

$$\boldsymbol{T} = \text{traza}(\hat{\boldsymbol{V}}^{-2}) - 2\boldsymbol{r}_\epsilon'\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{P}}^*\boldsymbol{r}_\epsilon$$

siendo:

$$\boldsymbol{r}_\epsilon = \hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{P}}^*\boldsymbol{y}$$
$$\hat{\boldsymbol{P}}^* = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}$$
$$\hat{\boldsymbol{V}} = \hat{\boldsymbol{\Sigma}}_u + \hat{\boldsymbol{\Sigma}}_\epsilon.$$

En nuestros datos tenemos que $\boldsymbol{\Sigma}_u$ es desconocida, de modo que el K apropiado será el propuesto por *Han* (2013).

2.3.3. cAIC

El AIC empleado para modelos mixtos en los que estamos interesados en estudiar cada área en particular y no la población en general es el cAIC. Viene dado por la siguiente expresión,

$$cAIC = -2\,log(f(\boldsymbol{\mu}, \boldsymbol{V})) + 2K$$

es decir, se utiliza la verosimilitud condicional y un parámetro de penalización K. Fueron muchos los que trataron de estudiar cuál era ese parámetro de penalización, en concreto *Vaida y Blanchard* (2005) y *Han* (2013).

cAIC de *Vaida y Blanchard* (2005)

El cAIC propuesto por *Vaida y Blanchard* (2005) se utiliza para modelos en los que conocemos la varianza del efecto aleatorio, algo bastante improbable en la práctica. Este cAIC viene dado por la siguiente expresión.

$$cAIC1 = -2\,log(f(\boldsymbol{\mu}, \boldsymbol{V})) + 2K =$$
$$= -2\Big(-\frac{D}{2}log(\sqrt{2\pi}) - \frac{1}{2}\sum_{d=1}^{D} log\{\sigma^2_{\epsilon d}\} - \frac{1}{2}\sum_{d=1}^{D}[(y_d - X_d\boldsymbol{\beta} - u_d)'\sigma^{-2}_{\epsilon d}(y_d - X_d\boldsymbol{\beta} - u_d)]\Big) + 2\rho =$$

$$= D\,log(\sqrt{2\pi}) + \sum_{d=1}^{D} log\{\sigma^2_{\epsilon d}\} + \sum_{d=1}^{D}[(y_d - X_d\boldsymbol{\beta} - u_d)'\sigma^{-2}_{\epsilon d}(y_d - X_d\boldsymbol{\beta} - u_d)] + 2\rho$$

con ρ traza de la "matriz hat".

En el caso de no tener efecto aleatorio, ρ coincide con p donde p son el número de variables explicativas introducidas en el modelo.

cAIC *Han* (2013)

Debido a la improbabilidad de conocer la varianza del efecto aleatorio en la práctica, *Han* propuso un cAIC en el caso de estimar los parámetros del modelo por ML y REML y de no conocer la varianza del efecto aleatorio. Dicho cAIC tiene la misma verosomilitud que el cAIC de *Vaida y Blanchard* (2005) pero varía en el parámetro de penalización K. Este nuevo parámetro es $\rho - C$ con ρ traza de la "matriz hat" y C una cantidad debida a la estimación de $\boldsymbol{\Sigma}_u$.

Como se ve, este nuevo cAIC permite tener en cuenta la estimación de la varianza del efecto aleatorio. El término de penalización coincide con los grados de libertad que nos permitirán conocer la complejidad del modelo. Por tanto, este cAIC viene dado por la siguiente expresión:

$$cAIC2 = -2\,log(f(\boldsymbol{\mu}, \boldsymbol{V})) + 2K =$$

$$= D\,log(\sqrt{2\pi}) + \sum_{d=1}^{D} log\{\sigma^2_{\epsilon d}\} + \sum_{d=1}^{D}[(y_d - X_d\boldsymbol{\beta} - u_d)'\sigma^{-2}_{\epsilon d}(y_d - X_d\boldsymbol{\beta} - u_d)]\Big) + 2(\rho - C).$$

Como se vio, ambos cAIC tienen la misma verosimilitud pero varían en el parámetro de penalización por lo que la complejidad del modelo no será la misma y dependerá del cAIC usado. En el cAIC de *Han* (2013), C siempre es una cantidad negativa, de modo que el valor de $\rho - C$ será mayor que ρ. Por ello si no conocemos la varianza del efecto aleatorio el modelo será más complejo; de hecho cuánto mayor sea la varianza del efecto aleatorio, mayor será K.

Además si la varianza del efecto aleatorio es cero, C también será cero y, al igual que ocurría con el cAIC de *Vaida y Blanchard* (2005), ρ será igual a p. Por tanto, en el caso de no tener efecto aleatorio cAIC1 y cAIC2 coincidirán, la verosimilitud condicional siempre es la misma y K será igual a p.

En nuestros datos vamos a tener varianza del efecto aleatorio desconocida por lo que para elegir el mejor modelo se utilizará el cAIC2. Aún así el cAIC1 será calculado a

efectos de comparación.

Burnham y Anderson (2002) establecieron el criterio de una diferencia de a lo sumo 2 unidades en el AIC no es confiable para la clasificación de dos modelos, mientras que una diferencia de 10 es a favor del modelo con el cAIC más pequeño. *Vaida y Blanchard* (2005) utilizaron este criterio para modelos mixtos.

2.3.4. El AIC para modelos mixtos en R

Al igual que en la literatura se han estudiado las estimaciones en áreas pequeñas mediante modelos mixtos, en el programa R existen una serie de funciones para dichas estimaciones. En este trabajo se hará uso de las funciones *eblupFH* y *mseFH*, funciones que utilizan los modelos de Fay-Herriot y que pertenecen a la librería *sae*. Aunque nosotros trabajaremos con modelos a nivel de área, esta librería también se puede utilizar para modelos a nivel de individuo. Estas funciones nos permitirán conocer las estimaciones basadas en el modelo, las estimaciones de los parámetros, los errores cuadráticos medios de las estimaciones y el AIC.

El AIC que utiliza esta librería toma como verosilimilitud la marginal y como parámetro de penalización $K = p + 1$, donde p son el número de variables explicativas y 1 es una unidad de estimar $\mathbf{\Sigma}_u$.

Esta librería además de no tener en cuenta el efecto aleatorio cuando estamos interesados en estudiar cada área en particular, ya que considera la verosimilitud marginal, no considera bien el parámetro de penalización. Además, en el caso de no tener efecto aleatorio y tratar de estimar un modelo mediante el paquete *sae* sigue considerando como $K = p + 1$. *Vaida y Blanchard* (2005) y *Han* (2013) establecieron que si no existe efecto aleatorio K es igual a p. De modo que para calcular el AIC en modelos mixtos para datos reales, Capítulo 3, éste deberá ser programado.

Aunque este AIC no es correcto, para nuestros datos (Capítulo 3) también será calculados a efectos de comapración; lo denotaremos por mAIC.

2.3.5. Otras aportaciones para la selección de modelos lineales mixtos

Aunque el cAIC sea el criterio más utilizado y conocido para la selección de modelos en estimación en áreas pequeñas, tomando como referencia el artículo de *Pfeffermann* (2013) existieron otros autores que propusieron otros métodos para dicha selección:

- *Pan y Lim* (2005) propusieron un test de bondad de ajuste basados en sumas acumuladas de residuos para modelos lineales mixtos generalizados.

- *Jiang et al.* (2008) propone una clase de estrategias para la selección de modelos llamada *fence methods* aplicada tanto para modelos a nivel de área como de individuos.

Aunque estos métodos fueron introducidos bajo una suposición frecuentista, también se han utilizado métodos bayesianos (ver *Pfeffermann* (2013).)

Capítulo 3

Aplicación a datos reales. Estudio de los ingresos medios mensuales en el hogar para el año 2013

Como se vio en el Capítulo 1, teníamos tamaños de muestra muy pequeños sobre todo para IM_AJENA, IM_PROPIA e IM_NO_CONTRIB. Por ello era necesario introducir unas variables explicativas que nos ayuden a obtener buenas estimaciones de IMTOT, IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB mediante metodos SAE.

En la primera sección de este capítulo se detallarán las variables explicativas con las que comenzaremos el estudio. En las siguientes secciones se estudiarán IMTOT, IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB por separado con el siguiente procedimiento: normalidad de las variables respuesta (Subsección 1), correlación de cada variable respuesta con las variables explicativas introducidas en la Sección 3.1. (Subsección 2) y, por último en las Subsecciones 3, selección del mejor modelo mediante el AIC para MM.

El motivo de estudiar la normalidad para las cinco variables respuesta es porque el modelo que se utiliza para el estudio de estas cinco variables es el modelo de Fay-Herriot y que, como vimos en el Capítulo 2, requiere de esta hipótesis (errores y efectos aleatorios propuestos normales). Además también los métodos de REML y ML la necesitan. Para ello, realizaremos los histogramas y el correspondiente gráfico qqPlot de las cinco variables respuesta junto con el p-valor obtenido en el contraste de normalidad Shapiro-Wilk. Utilizaremos el test de Shapiro-Wilk ya que disponemos de pocos datos, veinte datos, uno por área.

En siguiente paso para conocer cada ingreso es estudiar sus correlaciones respecto a las variables explicativas detalladas en la Sección 3.1. Aquellas variables explicativas

más relacionadas con las variables respuesta serán las elegidas para introducir en el modelo saturado y con el que comenzaremos el estudio del AIC para MM en cada subsección 3. Como se verá no todas las variables respuesta están relacionadas con las mismas variables explicativas, por lo que para cada variable respuesta no se partirá de las mismas variables explicativas.

La variable IMTOT es la combinación de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB . Por lo que, en una última sección, se compararán las estimaciones basadas en el modelo de IMTOT con las estimaciones basadas en el modelo obtenidas de combinar las estimaciones basadas en el modelo de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB.

3.1. Variables explicativas

Para el estudio de IMTOT, IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB se utilizarán las siguientes variables auxiliares procedentes de la ECV, la"Agencia Estatal de la Administración Tributaria" (AEAT) o la "Seguridad Social" (SS).

1. **EDAD**: Porcentaje de personas en el hogar según edad. Categorías:
   - *a*) **NPER_18**: Porcentaje de personas menores de 18 años.
   - *b*) **NPER_18A64**: Porcentaje de personas entre 18 y 64 años.
   - *c*) **NPER_65**: Porcentaje de personas mayores de 65 años.

2. **NPER**: número medio de personas en el hogar.

3. **UC**: las unidades de consumo es una expresión que se utiliza para calcular el ingreso equivalente del hogar, permitiendo comparar los ingresos por hogar con distinto número de personas. Estará relacionado con NPER y EDAD. Se definirá utilizando la escala de equivalencia de la OCDE (Organización para la Cooperación y el Desarrollo Económico) modificada, teniendo la siguiente expresión:

   $$[1 + 0,5 \cdot (a - 1) + 0,3 \cdot b]$$

   Siendo:
   "a" el número de personas de 14 o más años del hogar,
   "b" el número de personas menores de 14 años del hogar.

4. **ESTUDIOS**: Porcentaje de personas en el hogar según estudio. Categorías:

   a) **NPER_PRIM**: Porcentaje de personas con estudios primarios. Se consideran personas con estudios primarios aquellas que no saben ni leer ni escribir y que son menores de 16 años o personas con la primera etapa de educación secundaria superada o EGB completa.

   b) **NPER_SEC**: Porcentaje personas con estudios secundarios. Dentro de las personas con estudios secundarios se incluyen las personas que tienen una titulación de bachillerato, COU o preuniversitarios, las personas con una titulación de formación profesional, artes plásticas y diseño y deportivas de grado medio y similares y personas con certificado de profesionalidad de nivel 3 o título propio universitario que precisa del título de bachiller de duración igual o superior a un semestre o inferior a dos años.

   c) **NPER_SUP**: Porcentaje personas con estudios superiores. Se incluye a las personas con una titulación de formación profesional, artes plásticas y diseño y deportivas de grado superior, las personas con un título propio universitario que precisa del título de bachiller de 2 o más años de duración, diploma, licenciatura, máster o especialidad en ciencias de la salud por el sistema de residencia y las personas con un doctorado universitario.

5. **NACIONALIDAD**: Porcentaje de personas en el hogar según la nacionalidad. Categorías:

   a) **NPER_ESP**: Porcentaje de personas con nacionalidad española.

   b) **NPER_NO_ESP**: Porcentaje de personas residentes en España con nacionalidad no española.

OBSERVACIÓN 3.1: Será necesario definir que es el ingreso equivalente por hogar para después poder definir qué son los hogares bajo el umbral de pobreza.

El ingreso equivalente por hogar viene dado por la siguiente expresión:

$$\frac{\text{IMTOT}}{\text{UC}}.$$

Este concepto corrige el efecto del número de personas en el volumen de ingresos del hogar, haciendo comparables los ingresos del hogar con distintos números de miembros.

6. **HOG_UMBRAL_POBREZA**: Porcentaje de hogares según el umbral de pobreza. Categorías:

   a) **HOG_BAJO_UMBRAL**: Porcentaje de hogares cuyos ingresos equivalentes están por debajo del 60 % de la mediana de los ingresos equivalentes de todas las personas gallegas.

   b) **HOG_SOBRE_UMBRAL**: Porcentaje de hogares cuyos ingresos equivalentes están por encima del 60 % de la mediana de los ingresos equivalentes de todas las personas gallegas.

7. **TIPOLOGÍA_HOGAR**: Porcentaje de hogares según la tipología del hogar. Categorías:

   a) **HOG_TIP1**: Porcentaje de hogares unipersonales, compuestos por una sola persona.

   b) **HOG_TIP2**: Porcentaje de hogares sin núcleo (más de una persona).

   c) **HOG_TIP3**: Porcentaje de hogares de parejas con hijos.

   d) **HOG_TIP4**: Porcentaje de hogares de parejas sin hijos.

   e) **HOG_TIP5**: Porcentaje de hogares monoparentales ya sea masculino o femenino.

   f) **HOG_TIP6**: Porcentaje de hogares compuestos por un núcleo y alguien más.

   g) **HOG_TIP7**: Porcentaje de hogares con varios núcleos.

8. **RENDIMIENTO**: Rendimiento medio anual declarado en el impuesto de la renta de las personas físicas (IRPF).

9. **PENSIONES**: importe medio mensual declarado de las pensiones contributivas de la Seguridad Social: incapacidad permanente, jubilación y muerte y supervivencia.

En el Cuadro 3.1. podemos ver un resumen de las variables explicadas anteriormente.

Comentar que la suma de los % de personas de cada grupo de EDAD (NPER_18, NPER_65, NPER_18A64) suman el 100 %, por tanto cada una de las variables de edad es combinación lineal de las otras dos. Ocurre lo mismo en el caso de las variables de los grupos: ESTUDIOS, NACIONALIDAD, HOG_UMBRAL_POBREZA y TIPOLOGÍA_DEL_HOGAR. Para solucionar esto se tomará dentro de cada grupo de variables una categoría de referencia.

VARIABLE	EXPLICACIÓN	CATEGORÍAS	FUENTE
EDAD	Porcentaje de personas en el hogar atendiendo a la edad: menores de 18 años (NPER_18), personas de 18 a 64 años (NPER_18A64) y mayores de 65 años (NPER_65)	NPER_18, NPER_18A64, NPER_65	ECV
NPER	Número medio de personas por hogar.		ECV
UC	Unidades de consumo. Tanto NPER como UC pertenecen al mismo grupo de variables, al número de personas en el hogar.		ECV
ESTUDIOS	Porcentaje de personas en el hogar atendiendo al tipo de estudios : personas con estudios primarios (NPER_PRIM), personas con estudios secundarios (NPER_SEC) y personas con estudios superiores (NPER_SUP).	NPER_PRIM, NPER_SEC, NPER_SUP	ECV
NACIONALIDAD	Porcentaje de personas en el hogar atendiendo a la nacionalidad: españoles (NPER_ESP) y extranjeros (NPER_NO_ESP).	NPER_ESP, NPER_NO_ESP	ECV
HOG_UMBRAL_ POBREZA	Porcentaje de hogares atendiendo al umbral de pobreza: hogares bajo el umbral de pobreza (HOG_BAJO_UMBRAL) y hogares sobre el umbral de pobreza (HOG_SOBRE_UMBRAL).	HOG_BAJO_ UMBRAL, HOG_SOBRE_ UMBRAL	ECV
TIPOLOGIA_ HOGAR	Porcentaje de hogares atendiendo a la tipología del hogar: hogares unipersonales (HOG_TIP1), hogares sin núcleo (HOG_TIP2), hogares de parejas con hijos (HOG_TIP3), hogares de parejas sin hijos (HOG_TIP4), hogares monoparentales (HOG_TIP5), hogares con un núcleo y alguien más (HOG_TIP6) y hogares con varios núcleos (HOG_TIP7).	HOG_TIP1. HOG_TIP2, HOG_TIP3, HOG_TIP4, HOG_TIP5, HOG_TIP6, HOG_TIP7	ECV
RENDIMIENTO	Rendimiento medio anual por persona.		AEAT
PENSIONES	Pensión media mensual por persona.		SS

Cuadro 3.1: Resumen de las variables explicativas con las que comenzamos en estudio.

En este trabajo las categorías de las variables serán tomadas como si fueran variables independientes a la hora de meterlas en el modelo. Así se suele hacer en el IGE. Además, al ser nuestra información agregada cada una de estas categorías de cada variable es un porcentaje o cantidad, no es una variable indicadora. Por lo tanto, partiremos inicialmente de veintiuna variables explicativas: NPER_18, NER_18A64, NPER_65, NPER, UC, NPER_PRIM, NPER_SEC, NPER_SUP, NPER_ESP, NPER_NO_ESP, HOG_BAJO_UMBRAL, HOG_SOBRE_UMBRAL, HOG_TIP1. HOG_TIP2, HOG_TIP3, HOG_TIP4, HOG_TIP5, HOG_TIP6, HOG_TIP7, RENDIMIENTO y PENSIONES.

3.2. IMTOT

En esta Sección trataremos de explicar el ingreso medio total en el hogar, IMTOT, con la ayuda de las variables explicativas, citadas en la Sección 3.1, mediante métodos SAE. Se estimará en las 20 áreas en las que se divide la Comunidad de Galicia, por lo tanto D=20.

Se quiere conocer el verdadero valor de IMTOT, μ_{1d}:

$$\mu_{1d} = X_d\boldsymbol{\beta} + u_d.$$

Partimos de las veinte estimaciones directas de IMTOT, y_{1d}, una para cada área:

$$y_{1d} = X_d\boldsymbol{\beta} + u_d + \epsilon_d \qquad \forall d = 1, ..., D$$

donde $\epsilon_d \sim N(0, \sigma^2_{\epsilon d})$ con $\sigma^2_{\epsilon d}$ conocida y $N(0, \sigma^2_u)$ con σ^2_u desconocida. Se tomará como variables respuesta y_{1d}, el estimador directo de IMTOT calculado de acuerdo a la expresión (1.1) del Capítulo 1.

Mediante los métodos SAE podremos obtener un buen estimador de μ_{1d}, $\hat{\mu}_{1d}$, definido como:

$$\hat{\mu}_{1d} = X_d\hat{\boldsymbol{\beta}} + \hat{u}_d$$

Para poder medir los errores de estimación de $\hat{\mu}_{1d}$ se utilizará una aproximación del error cuadrático medio:

$$\mathrm{MSE}(\hat{\mu}_{1d}) = \mathrm{E}[(\hat{\mu}_{1d} - \mu_{1d})^2] = g_1 + g_2 + g_3 \qquad \forall d = 1, ..., D$$

calculado por *Prasad y Rao* (1990). En las páginas 165-167 se puede ver el cálculo detallado de las funciones g_i, $i = 1, 2, 3$.

Se medirá el error con el coeficiente de variación, en porcentaje:

$$\mathrm{CV}_2 = \frac{\sqrt{\mathrm{MSE}(\hat{\mu}_{1d})}}{\hat{\mu}_{1d}} \cdot 100.$$

Cuanto menor sean CV_2 $\forall d = 1, ..., D$ más precisas serán las estimaciones.

Por tanto, para obtener μ_{1d} se realizará lo siguiente: estudio de la normalidad de IMTOT, necesaria en los modelos de Fay-Herriot; correlación de IMTOT con las variables explicativas citadas en la Sección 3..; y por último, selección del mejor modelo mediante el AIC para modelos mixtos partiendo de un modelo saturado con aquellas variables explicativas más relacionadas con IMTOT.

3.2.1. Normalidad

Para comprobar la normalidad de IMTOT realizaremos el histograma y el correspondiente gráfico qqPlot junto con el p-valor obtenido en el contraste de normalidad Shapiro-Wilk.

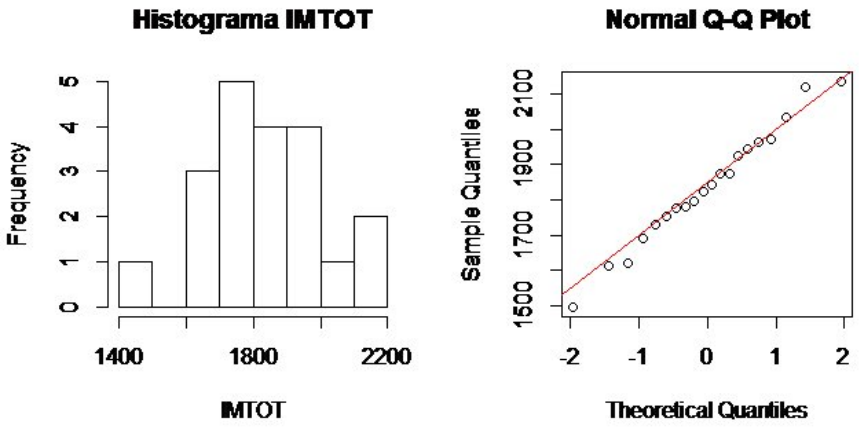

Figura 3.1: Histograma y gráfico qq-Plot del IMTOT.

Como se observa en la Figura 3.1, IMTOT sigue una distribución normal; además en el contraste de Shapiro-Wilk hemos obtenido un p-valor de 0.99.

3.2.2. Correlación

El segundo paso para conocer IMTOT en las veinte áreas gallegas es ver aquellas variables con las que más está relacionado. Al partir de 21 variables explicativas, este es un primer criterio para elegir las siete u ocho variables que posiblemente se introducirán en el modelo final.

Como veremos en las Figuras 3.2, 3.3, 3.3 y 3.5, el IMTOT está principalmente relacionado con: NPER_65, NPER_18 y NPER_18A64; NPER_PRIM y NPER_SUP; HOG_BAJO_UMBRAL Y HOG_SOBRE_UMBRAL; RENDIMIENTO y PENSIONES.

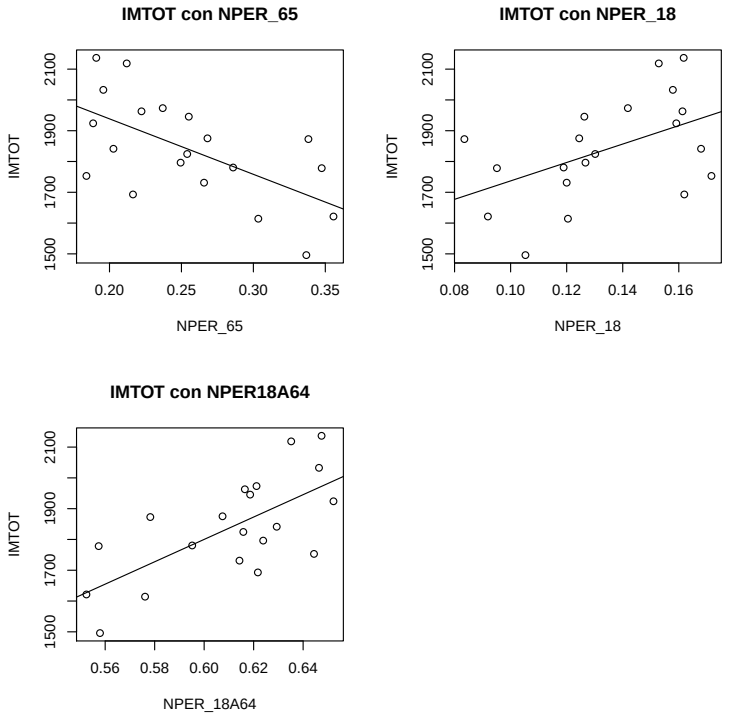

Figura 3.2: Gráfico de dispersión de IMTOT con EDAD: NPER_18, NPER_18A64 Y NPER_65.

Aunque las tres categorías del grupo EDAD están muy relacionadas con IMTOT, como cada una de ellas es combinación lineal de las otras dos, para el modelo nos quedamos con NPER_65 y NPER_18A64, pues de las tres variables son las dos más relacionadas con IMTOT. Esto sería equivalente a tomar como categoría de referencia NPER_18.

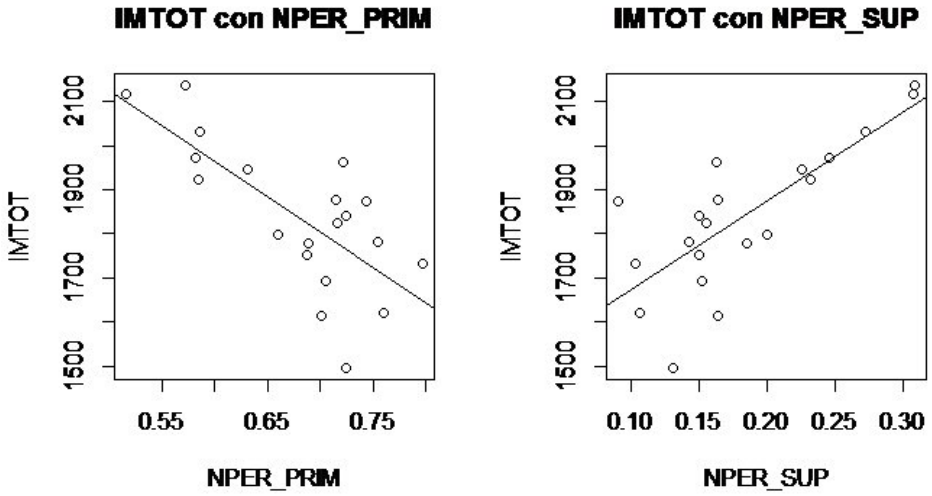

Figura 3.3: Gráficos de dispersión de IMTOT con NPER_PRIM y NPER_SUP.

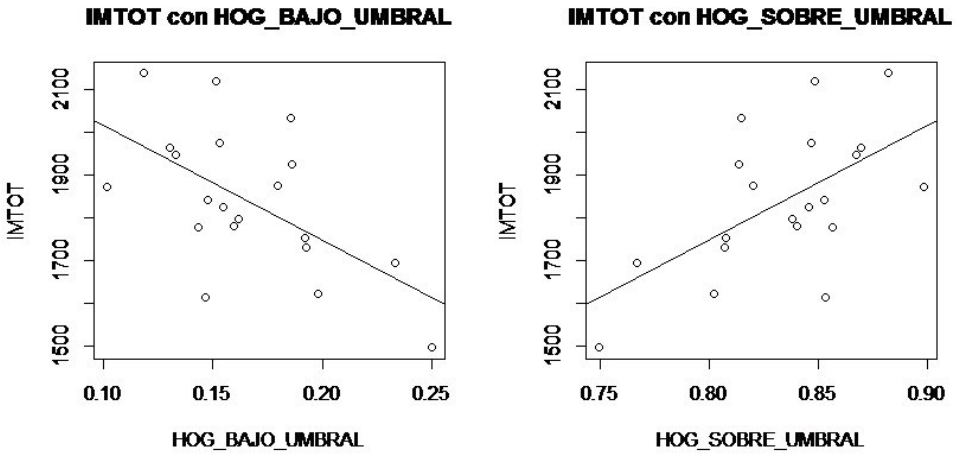

Figura 3.4: Gráficos de dispersión de IMTOT con HOG_BAJO_UMBRAL y HOG_SOBRE_UMBRAL.

HOG_BAJO_UMBRAL es combinación lineal de HOG_SOBRE_UMBRAL y viceversa; por tanto, para el modelo solo usaremos una de ellas, por ejemplo, HOG_BAJO_UMBRAL tomando como categoría de referencia HOG_SOBRE_UMBRAL.

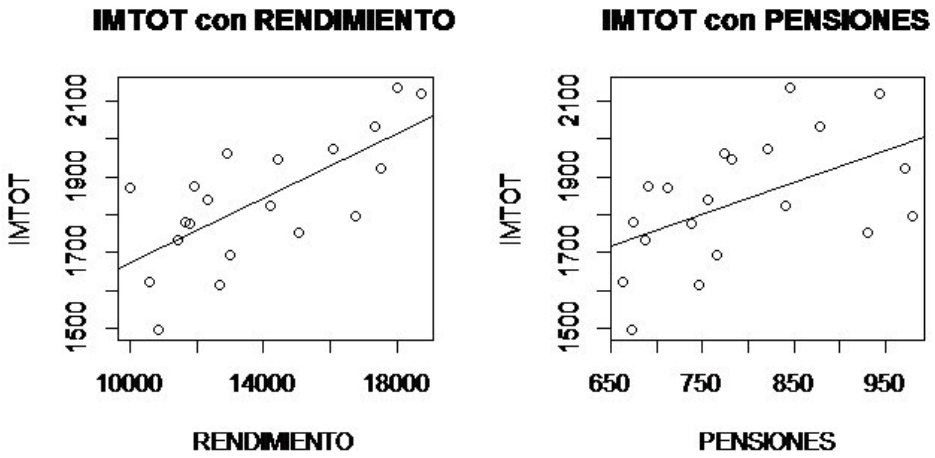

Figura 3.5: Gráficos de dispersión de IMTOT con RENDIMIENTO y PENSIONES.

En el Cuadro 3.2 se puede ver las correlaciones de IMTOT con las variables que introduciremos en el modelo en la Subsección 3.2.3. Además hemos introducido las correlaciones de las variables explicativas entre sí. Las variables elegidas para introducir en el modelo son aquellas variables más correlacionadas con IMTOT, se elegirá como criterio tomar aquellas variables que tengan un coeficiente de correlación mayor al 50 % con IMTOT.

	IMTOT	NPER_18A64	NPER_65	NPER_PRIM	NPER_SUP	HOG_BAJO_UMBRAL	RENDIMIENTO	PENSIONES
NPER_18A64	**0,68**	X	-0,97	-0,58	0,60	-0,10	0,78	0,74
NPER_65	**-0,61**	-0,97	X	0,56	-0,58	0,02	-0,74	-0,70
NPER_PRIM	**-0,72**	-0,58	0,56	X	-0,95	0,20	-0,90	-0,73
NPER_SUP	**0,77**	0,60	-0,58	-0,95	X	0,05	0,90	0,57
HOG_BAJO_UMBRAL	**-0,59**	-0,10	0,02	0,20	-0,05	X	-0,14	-0,11
RENDIMIENTO	**0,70**	0,78	-0,74	-0,90	0,90	-0,14	X	0,89
PENSIONES	**0,51**	0,74	-0,70	-0,73	0,57	-0,11	0,89	X

Cuadro 3.2: Coeficiente de correlación del IMTOT con las variables que consideramos para el modelo. Incluimos también las correlaciones de las variables explicativas entre sí.

3.2.3. Selección del mejor modelo mediante el AIC para MM

A continuación haremos una tabla en la que se incluyen los mejores modelos para explicar IMTOT junto con cAIC1, cAIC2 y el mAIC. Hemos calculado los tres AIC habiendo estimado los modelos mediante REML. Indicaremos con una "X" las variables que se incluyen en cada modelo. También incluiremos los grados de libertad de cada modelo según el AIC calculado. Por último elegiremos el mejor modelo y compararemos sus estimaciones con las estimaciones directas mediante gráficos.

	AIC			VARIABLES EXPLICATIVAS								
	cAIC1	cAIC2	mAIC	CTE	NPER_18A64	NPER_65	NPER_PRIM	NPER_SUP	HOG_BAJO_UMBRAL	RENDIMIENTO	PENSIONES	$\hat{\sigma}_u^2$
1.	234,266	236,266	234,266	X	X	X	X	X	X	X	X	0
2.	232,297	232,297	234,297	X	X	X	X	X	X	X		0
3.	231,2487	231,2487	233,2487	X	X	X		X	X	X		0
4.	231,5518	234,3907	233,8651		X	X		X	X	X		219,66
5.	230,9698	233,5498	233,5994		X			X	X	X		531,38

Cuadro 3.3: Tabla con los cinco mejores modelos para IMTOT en los que hemos indicado que variable incluimos en cada uno de ellos y hemos calculado el cAIC1, cAIC2 y mAIC. Por último la estimación de la varianza del efecto aleatorio para cada modelo.

Para la selección del mejor modelo para IMTOT se parte del modelo saturado (Modelo 1) con todas aquellas variables explicativas con las que más estaba relacionado IMTOT (ver Cuadro 3.2), se calculan los tres AIC descritos en la metodología (cAIC1,cAIC2 y mAIC) y la varianza del efecto aleatorio. Obtenemos un primer modelo sin efecto aleatorio de ocho variables explicativas en el que no todas las variables son significativas, se elimina la variable menos significativa (PENSIONES) y se vuelven

a calcular de nuevo los AIC. Obtenemos un segundo modelo de siete variables donde algunas no son significativas y en el que tampoco tenemos efecto aleatorio. Eliminamos la variable menos significativa, NPER_PRIM, se obtiene un modelo de seis variables sin efecto aleatorio. Seguimos teniendo un modelo donde alguna de sus variables no es significativa, de modo que eliminamos su variable menos significativa (CTE) obteniendo un modelo de cinco variables con una varianza del efecto aleatorio de 219,66 pero seguimos teniendo alguna variable no significativa, como es NPER_65. Eliminada esta última variable, obtenemos finalmente un modelo de cuatro variables donde todas las variables son significativas y con una varianza del efecto aleatorio de 531,38.

Aunque dentro de los cinco mejores modelos hay algunos donde no hay efecto aleatorio y otros donde si lo hay, para comparar todos los modelos utilizaremos el cAIC2; si todos los modelos no tuviesen efecto aleatorio bastaría con usar el mAIC. Como se ve en el Cuadro 3.3, el modelo con el mejor cAIC2 es el modelo 3 pero como no todas las variables son significativas, se tomará como mejor modelo el modelo 5 con todas las variables significativas; además una diferencia de dos unidades no es significativa para la selección del mejor modelo. De hecho como se observa en la Figura 3.6 no hay mucha diferencia entre los residuos del modelo 3 y del modelo 5. Se han calculado los residuos relativos, definidos del siguiente modo:

$$\hat{e} = \frac{y_{1d} - \hat{\mu}_{1d}}{\hat{\mu}_{1d}}.$$

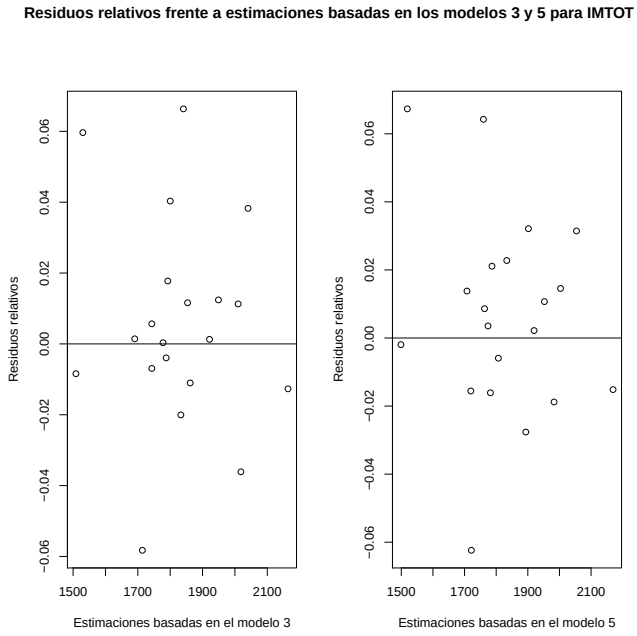

Figura 3.6: Gráfico de los residuos relativos frente a las estimaciones basadas en los modelos 3 y 5 para IMTOT.

A continuación compararemos los grados de libertad de los cinco mejores modelos para estimar IMTOT.

	Grados de libertad de los modelos con cAIC1	Grados de libertad de los modelos de cAIC2	Grados de libertad de los modelos con mAIC	$\hat{\sigma}_u^2$
1.	8	8	9	0
2.	7	7	8	0
3.	6	6	7	0
4.	5,63	7,05	6	219,66
5.	5,57	6,86	5	531.38

Cuadro 3.4: Tabla con los cinco mejores modelos para IMTOT en los que hemos indicado sus grados de libertad en función del AIC calculado y σ_u^2 de cada modelo.

Como se observa en el Cuadro 3.4 si no tenemos efecto aleatorio los grados de libertad del cAIC2 serán los mismos que el número de variables que introducimos al modelo. En cambio, si tenemos cierta varianza del efecto aleatorio los grados de libertad serán mayores, p+C donde C es una cierta cantidad de estimar σ_u^2.

El mejor modelo elegido para IMTOT es el formado por las variables explicativas NPER_18A64, NPER_SUP, HOG_BAJO_UMBRAL y RENDIMIENTO. Tiene 6,86 grados de libertad y una varianza del efecto aleatorio de 531,38, muy superior a la que tienen los otro cuatro modelos descartados. De hecho los modelos 1, 2 y 3 son modelos de efectos fijos.

La varianza del efecto aleatorio del modelo finalmente elegido (Modelo 5) tampoco es tan alta si lo comparamos con los datos que tenemos (datos con una media de 1827,5 unidades). Esto es debido a que las variables explicativas introducidas en el modelo estaban altamente relacionadas con IMTOT y por tanto ya explican por si solas gran parte del modelo sin la necesidad de introducir mucho efecto aleatorio al modelo.

VARIABLE EXPLICATIVA	ESTIMACIÓN	P-VALOR
NPER_18A64	3612,990	0
NPER_SUP	2370,586	0,00028
HOG_BAJO_UMBRAL	-1590,881	0,00171
RENDIMIENTO	-0,0394	0,01175
σ_u^2	531,38	

Cuadro 3.5: Tabla con las estimaciones para las variables explicativas del modelo elegido para IMTOT junto con sus p-valores y la varianza del efecto aleatorio del modelo.

Hemos obtenido el siguiente modelo para IMTOT:

$$\hat{\mu}_{1d} = 3612,990\, NPER\_18A64 + 2370,586\, NPER\_SUP$$

$$-1590,8881\, HOG\_BAJO\_UMBRAL - 0,0394 RENDIMIENTO + \hat{u}_d$$

$$\forall d = 1, ..., D.$$

El modelo que tenemos nos indica que cuanto mayor sea NPER_18A64 y NPER_SUP mayor será IMTOT. Y, teniendo en cuenta las dos variables anteriores, cuanto mayor sea HOG_BAJO_UMBRAL menor será IMTOT. Vemos que RENDIMIENTO tiene estimación negativa esto es debido a que la variable RENDIMIENTO estaba muy correlacionado con NPER_SUP.

A continuación haremos una tabla con las estimaciones directas, $\hat{y}_{1d}$; estimaciones basadas en el modelo, $\hat{\mu}_{1d}$; coeficiente de variación de las estimaciones directas y de las estimaciones basadas en el modelo, CV_1 y CV_2 respectivamente. Después mediante gráficos compararemos dichas estimaciones, haremos un primer gráfico (Figura 3.6) de las estimaciones directas y las basadas en el modelo frente a las áreas y un segundo gráfico boxplot (Figura 3.7) de CV_1 y CV_2 que nos permitirá saber por qué hay ciertas diferencias entre las estimaciones directas y las basadas en el modelo en cada área. En los gráficos boxplot se tomará un intervalo de 0 al 20 % ya que si un CV es mayor del 20 % ese dato no es publicable (consultar *ONS* (2014)).

Áreas	$\hat{y}_{1d}$	$\hat{\mu}_{1d}$	CV_1	CV_2
Área 1	1875,237	1833,482	3,05 %	2,34 %
Área 2	1796,260	1806,963	3,01 %	2,67 %
Área 3	1731,233	1707,668	4,30 %	2,52 %
Área 4	1963,008	1901,916	4,26 %	2,08 %
Área 5	2118.591	2053.976	2,91 %	2,32 %
Área 6	2136,663	2169,537	3,36 %	2,27 %
Área 7	1778,362	1763,179	3,26 %	2,45 %
Área 8	1872,786	1759,728	3,86 %	3,01 %
Área 9	1945,619	1982,901	3,06 %	2,12 %
Área 10	1824,403	1786,700	7,68 %	2,20 %
Área 11	1495,968	1498,878	5,55 %	3,91 %
Área 12	1614,166	1721,542	4,05 %	2,15 %
Área 13	1621,329	1519,058	6,51 %	2,69 %
Área 14	1973,362	1952,491	3,08 %	2,07 %

Área 15	1780,492	1774,214	6,16 %	1,98 %
Área 16	1692,961	1719,718	3,99 %	2,72 %
Área 17	1841,337	1893,640	2,22 %	2,30 %
Área 18	1753,132	1781,804	4,91 %	2,61 %
Área 19	2032,677	2003,507	7,61 %	2,02 %
Área 20	1924,012	1919,816	2,57 %	2,33 %

Cuadro 3.6: Tabla con las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IMTOT en cada área. Añadimos también CV_1 y CV_2.

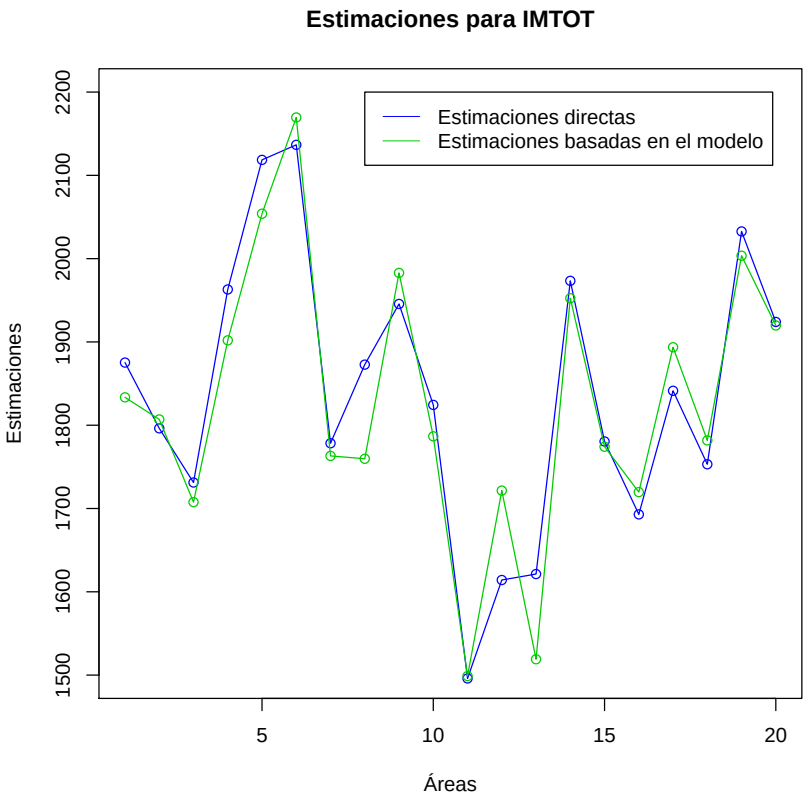

Figura 3.7: Gráfico de las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IMTOT.

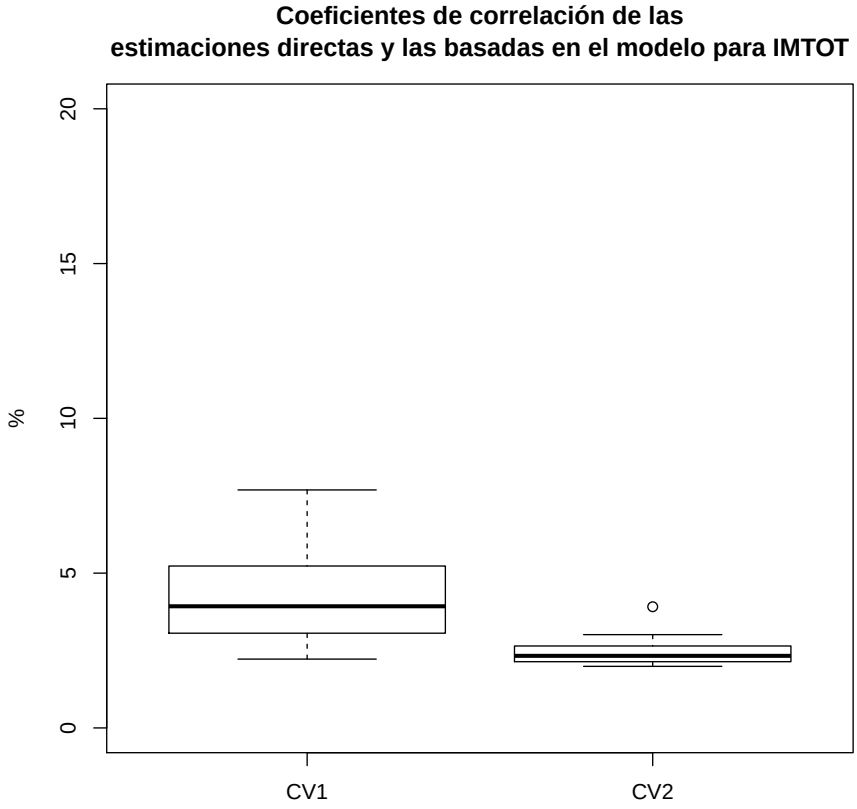

Figura 3.8: Gráfico boxplot de los coeficientes de variación para las estimaciones de IMTOT.

Como vemos en la Figura 3.6 las estimaciones basadas en el modelo son muy similares a las estimaciones directas. De todos modos las estimaciones basadas en el modelo tienen unos coeficientes de variación mucho más bajos que las estimaciones directas, lo que nos permite concluir que las estimaciones basadas en el modelo tienen una mejor precisión.

Por último veremos las estimaciones directas y las basadas en el modelo representadas en dos mapas de Galicia. Cuanto menor sea IMTOT más claro estará coloreada esa área.

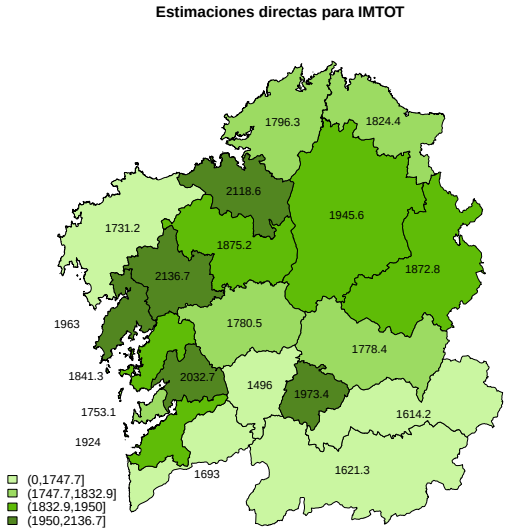

Figura 3.9: Mapa de las estimaciones directas, irán en colores más claros los IMTOT más bajos y en colores más oscuros los IMTOT más altos.

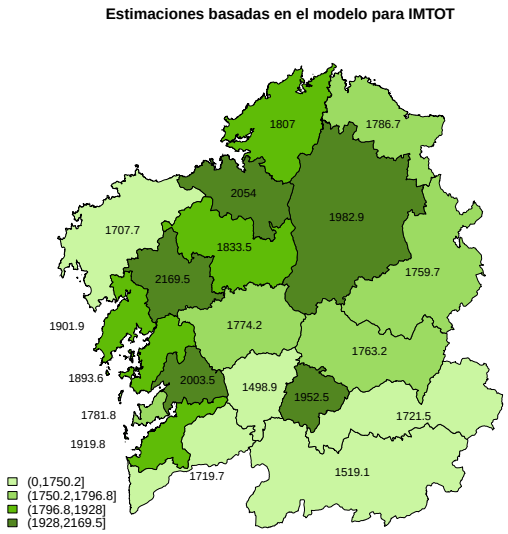

Figura 3.10: Mapa de las estimaciones basadas en el modelo, irán en colores más claros los IMTOT más bajos y en colores más oscuros los IMTOT más altos.

3.3. IM_AJENA

Ahora estamos interesados en estudiar las cuatro variables en las que se divide IMTOT: IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB. Comenzaremos con IM_AJENA, ingresos mensuales procedente de ingresos por cuenta ajena en el hogar para las veinte áreas en las que dividíamos la Comunidad de Galicia en el Capítulo 1.

El objetivo es conocer el verdadero valor de IM_AJENA, μ_{2d}:

$$\mu_{2d} = X_d\boldsymbol{\beta} + u_d.$$

Al igual que hicimos en la Sección anterior, partimos de las veinte estimaciones directas de IM_AJENA, y_{2d}, calculadas de acuerdo a la expresión (1.1) e introducidas en el siguiente modelo:

$$y_{2d} = X_d\boldsymbol{\beta} + u_d + \epsilon_d \qquad \forall d = 1, ..., D$$

donde $\epsilon_d \sim N(0, \sigma^2_{\epsilon d})$ con $\sigma^2_{\epsilon d}$ conocida y $N(0, \sigma^2_u)$ con σ^2_u desconocida.

Mediante métodos SAE trataremos de obtener un buen estimador de μ_{2d}, $\hat{\mu}_{2d}$:

$$\hat{\mu}_{2d} = X_d\hat{\boldsymbol{\beta}} + \hat{u}_d.$$

Para ello, se estudirá la normalidad de IM_AJENA, necesaria en los modelos de Fay-Herriot; correlación de IM_AJENA con las variables explicativas de la Sección 3.1.; y por último, selección del mejor modelo mediante el AIC para modelos mixtos partiendo de un modelo saturado con las variables explicativas más relacionadas con IM_AJENA.

3.3.1. Normalidad

Para comprobar la normalidad de IM_AJENA realizaremos el histograma y el correspondiente gráfico qqPlot junto con el p-valor obtenido en el contraste de normalidad Shapiro-Wilk.

Como se observa el la Figura 3.11 IM_AJENA podría seguir una distribución normal. Además en el contraste de Shapiro-Wilk hemos obtenido un p-valor de 0.2. Por tanto concluimos que IM_AJENA sigue una distribución normal.

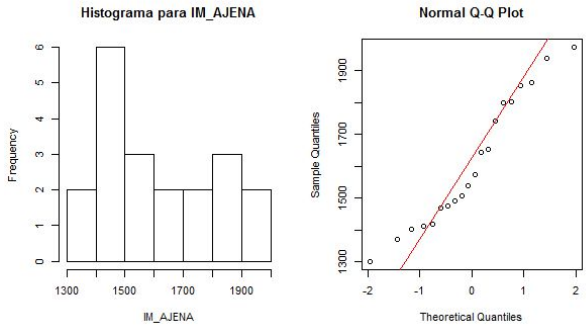

Figura 3.11: Histograma y gráfico qq-Plot del IM_AJENA.

3.3.2. Correlación

El siguiente paso para estimar μ_{2d} es ver aquellas variables más correlacionadas con IM_AJENA. Este es un primer criterio para seleccionar las seis o siete variables que posiblemente sean introducidas en el modelo final, ya que como se comentó partíamos de veintiuna variables.

Como veremos en las Figuras 3.12, 3.13 y 3.14, IM_AJENA está principalmente relacionado con: NPER_65, NPER_18 y NPER_18A64; NPER_PRIM y NPER_SUP; RENDIMIENTO y PENSIONES.

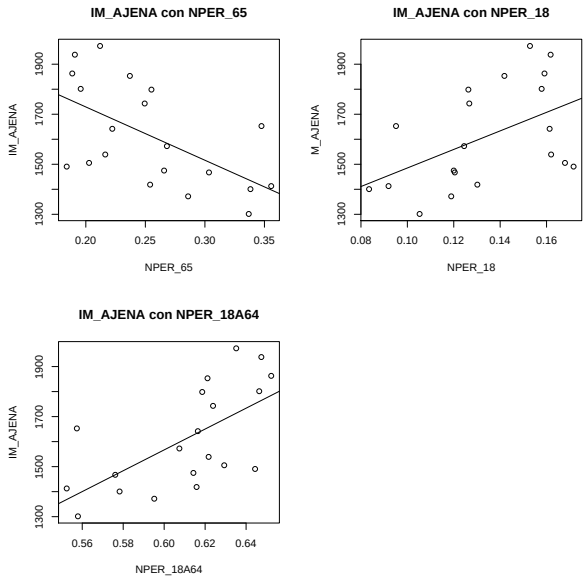

Figura 3.12: Gráfico de dispersión de IM_AJENA con EDAD: NPER_18, NPER_18A64 Y NPER_65.

Con el grupo de las variables relacionadas con la EDAD solo elegiremos para introducir en el modelo las variables NPER_65 y NPER_18A6, tomando como categoría de referencia NPER_18.

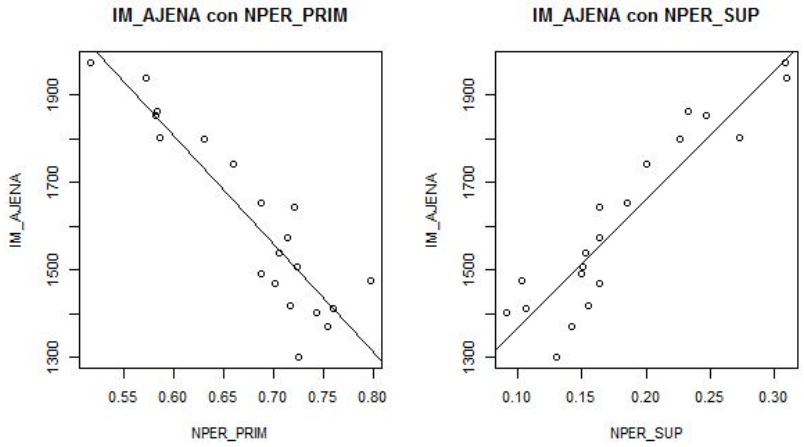

Figura 3.13: Gráficos de dispersión de IM_AJENA con NPER_PRIM y NPER_SUP.

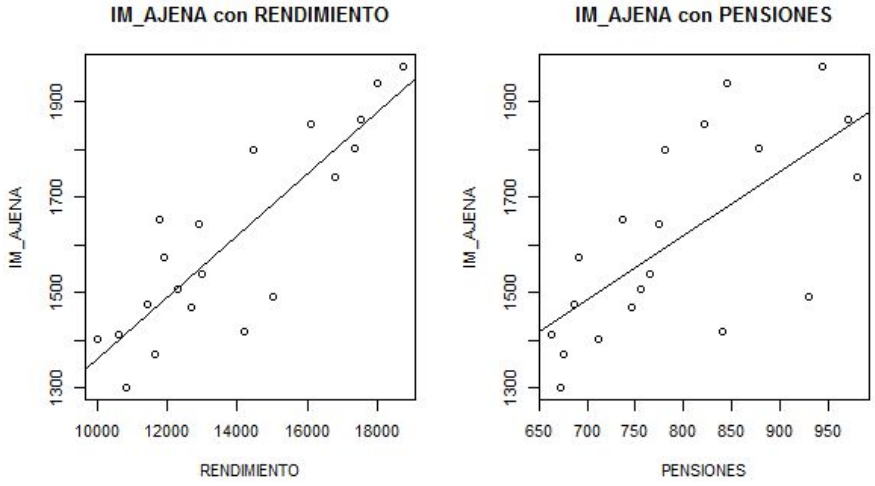

Figura 3.14: Gráficos de dispersión de IM_AJENA con RENDIMIENTO y PENSIONES.

En el Cuadro 3.7 se puede ver las correlaciones de IM_AJENA con las variables que introduciremos en el modelo en la Subsección 3.3.3. Además hemos introducido las correlaciones de las variables explicativas entre sí.

	IM_AJENA	NPER_18A65	NPER_65	NPER_PRIM	NPER_SUP	RENDIMIENTO	PENSIONES
NPER_18A64	**0,63**	X	-0,97	-0,58	0,60	0,78	0,74
NPER_65	**-0,58**	-0,97	X	0,56	-0,58	-0,74	-0,70
NPER_PRIM	**-0,90**	-0,58	0,56	X	-0,95	-0,90	-0,73
NPER_SUP	**0,92**	0,60	-0,58	-0,95	X	0,90	0,65
RENDIMIENTO	**0,86**	0,78	-0,74	-0,90	0,90	X	0,89
PENSIONES	**0,67**	0,74	-0,70	-0,73	0,65	0,89	X

Cuadro 3.7: Coeficiente de correlación del IM_AJENA con las variables que consideramos para el modelo. Incluimos también las correlaciones de las variables explicativas entre sí.

Se seleccionarán aquellas variables explicativas que tenían con IMTOT un coeficiente de correlación superior al 50 % (ver Cuadro 3.7).

3.3.3. Selección del mejor modelo mediante el AIC para MM

Al igual que hicimos para IMTOT, elegiremos el mejor modelo para explicar IM_AJENA entre los mejores modelos capaces de explicar IM_AJENA. Después compararemos las estimaciones obtenidas con las estimaciones directas mediante los coeficientes de variación. Por último representaremos ambas estimaciones en dos mapas de Galicia.

	AIC			VARIABLES EXPLICATIVAS							
	cAIC1	cAIC2	mAIC	CTE	NPER_18A64	NPER_65	NPER_PRIM	NPER_SUP	RENDIMIENTO	PENSIONES	$\hat{\sigma}_u^2$
1.	244,03	244,03	246,03	X	X	X	X	X	X	X	0
2.	242,16	242,16	244,16		X	X	X	X	X	X	0
3.	240,44	240,44	242,44		X	X	X	X	X		0
4.	238,58	238,58	240,58		X	X	X	X			0
5.	237,10	237,10	239,10		X	X		X			0
6.	238,04	238,04	240,04		X			X			0

Cuadro 3.8: Tabla con los cinco mejores modelos para IM_AJENA en los que hemos indicado que variable incluimos en cada uno de ellos y hemos calculado el cAIC1, cAIC2 y mAIC. Por último la estimación de la varianza del efecto aleatorio para cada modelo.

Para la selección del mejor modelo hemos seguido el procedimiento de IMTOT, elegimos el modelo saturado, calculamos los AIC y la varianza del efecto aleatorio obteniendo un modelo sin efecto aleatorio donde muchas de sus variables son no significativas. Eliminamos la variable menos significativa y obtenemos un segundo modelo. Seguimos el procedimiento hasta tener un modelo donde todas sus variables explicativas sean significativas (Modelo 6).

Entre todos los modelos elegidos como candidatos ninguno tiene efecto aleatorio; de modo que para la selección del mejor modelo bastaría con usar mAIC. Este método elige como mejor modelo al 5 pero la variable NPER_65 no es significativa, de modo que elegiremos finalmente al modelo 6. Además como se observa en la Figura 3.15 no hay mucha diferencia entre los residuos de los modelos 5 y 6. Se han utilizado los residuos relativos:

$$\hat{e} = \frac{y_{2d} - \hat{\mu}_{2d}}{\hat{\mu}_{2d}}.$$

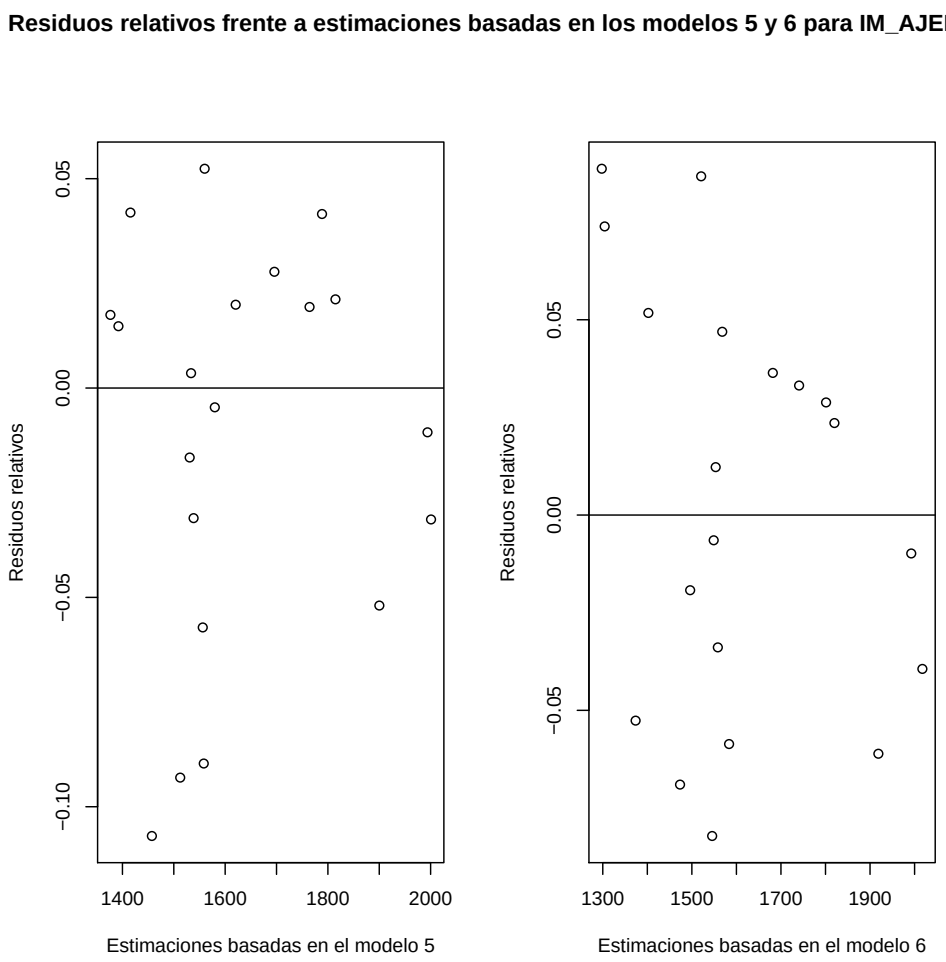

Figura 3.15: Gráfico de los residuos de Pearson frente a las estimaciones basadas en los modelos 5 y 6 para IM_AJENA.

Comentar que como hemos obtenido un modelo sin efectos aleatorios lo podríamos haber estimado sin tener en cuenta los efectos aleatorios, estimar un modelo lineal estimando los parámetros por ML, y calcular posteriormente el AIC usual. De hecho, si tenemos un modelo sin efectos aleatorios el AIC usual (habiendo estimado primero un modelo de forma lineal) coincide con el cAIC. Pero el mAIC de la librería sae no coincide con el cAIC por el motivo explicado en la metodología.

	Grados de libertad de los modelos con cAIC1	Grados de libertad de los modelos de cAIC2	Grados de libertad de los modelos con mAIC	$\hat{\sigma}_u^2$
1.	7	7	8	0
2.	6	6	7	0
3.	5	5	6	0
4.	4	4	5	0
5.	3	3	4	0
6.	2	2	3	0

Cuadro 3.9: Tabla con los cinco mejores modelos para IM_AJENA en los que hemos indicado sus grados de libertad en función del AIC calculado y σ_u^2 de cada modelo.

Si comparamos ahora los grados de libertad de los seis mejores modelos vemos que, como en ninguno de los modelos tenemos efecto aleatorio, los grados de libertad de los modelos coinciden con los grados con el número de variables introducidas en el modelo. En cambio, los grados de libertad ofrecidos para el AIC para modelos mixtos de la librería sae son $p + 1$ con p número de variables introducidas en el modelo. Esto fue explicado con más detalle en la metodología.

Finalmente, elegimos como mejor modelo para IM_AJENA a un modelo sin efectos aleatorios con variables explicativas NPER_18A64 y NPER_SUP, las cuales son significativas. Cuanto mayor sea NPER_18A64 y NPER_SUP mayor será IM_AJENA. El posible motivo por el que no teníamos efecto aleatorio es porque, como vimos en el Cuadro 3.7, NPER_SUP tenía un coeficiente de correlación de 0.92; y por ello, junto con NPER_18A64 fue prácticamente capaz de explicar IM_AJENA sin la necesidad de incluir ningún efecto aleatorio al modelo.

VARIABLE EXPLICATIVA	ESTIMACIÓN	P-VALOR
NPER_18A64	1832,597	0
NPER_SUP	2690,063	0
σ_u^2	0	

Cuadro 3.10: Tabla con las estimaciones para las variables explicativas del modelo elegido para IM_AJENA junto con sus p-valores y la varianza del efecto aleatorio del modelo.

Hemos obtenido el siguiente modelo para IM_AJENA:

$$\hat{\mu}_{2d} = 1832,597\, NPER\_18A64 + 2370,586\, NPER\_SUP + \hat{u}_d \qquad \forall d = 1, ..., D.$$

A continuación mostraremos las estimaciones directas y las estimaciones basadas en el modelo, sus coeficientes de variación CV_1 y CV_2 respectivamente, para después compararlos mediante gráficos al igual que hicimos para IMTOT.

Áreas	$\hat{y}_{2d}$	$\hat{\mu}_{2d}$	CV_1	CV_2
Área 1	1572,5	1553,4	6,79 %	2,45 %
Área 2	1742,9	1681,7	5,86 %	2,24 %
Área 3	1474,9	1402,3	11,98 %	3,11 %
Área 4	1641,7	1568,1	5,02 %	2.93 %
Área 5	1972,8	1992,4	3,41 %	3,22 %
Área 6	1938,0	2017,5	4,28 %	2,83 %

Área 7	1652,7	1520,8	6,90 %	2,23 %
Área 8	1400,6	1304,2	9,97 %	3,48 %
Área 9	1798,4	1740,7	3,96 %	2,90 %
Área 10	1418,5	1545,5	15,80 %	1,90 %
Área 11	1301,4	1373,7	15,06 %	2,21 %
Área 12	1467,3	1496,1	8.96 %	2,14 %
Área 13	1412,7	1297,7	12,60 %	2,80 %
Área 14	1853,1	1801,2	4,42 %	2,63 %
Área 15	1371,7	1473,4	10,59 %	2,34 %
Área 16	1539,0	1549,0	4,00 %	3,79 %
Área 17	1505,4	1558,2	5,19 %	3,20 %
Área 18	1490,7	1583,5	4,31 %	3,68 %
Área 19	1801,4	1918,7	9,35 %	1,93 %
Área 20	1862,9	1820,0	3,72 %	2.85 %

Cuadro 3.11: Tabla con las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_AJENA en cada área. Añadimos también CV_1 y CV_2.

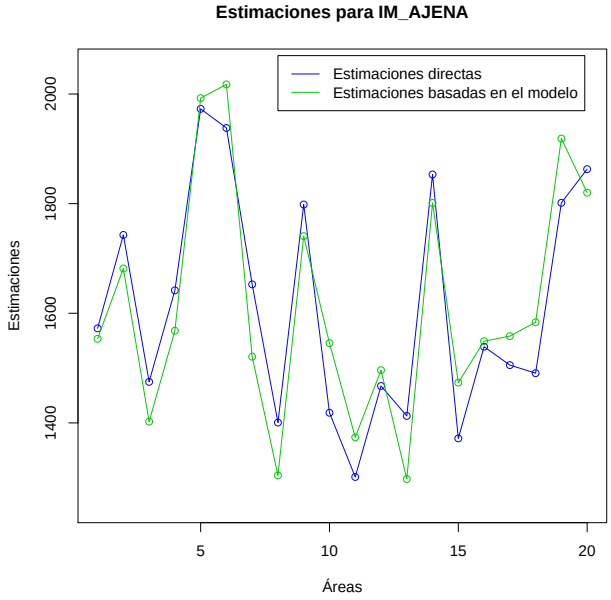

Figura 3.16: Gráfico de las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_AJENA.

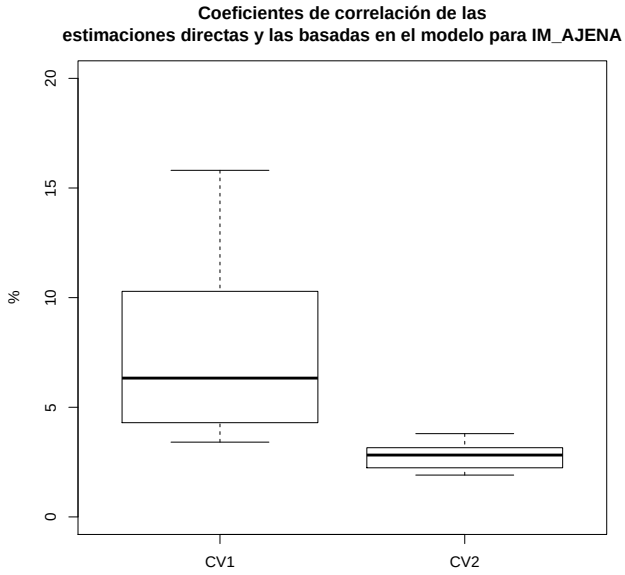

Figura 3.17: Gráfico boxplot de los coeficientes de variación para las estimaciones de IM_AJENA.

Las estimaciones basadas en el modelo son mucho más precisas que las estimaciones directas ya que tenemos que CV_2 es mucho menor que CV_1. De hecho, CV_2 está por debajo del 5 % para todas las áreas.

A Continuación veremos las estimaciones directas y las basadas en el mejor modelo representadas en los mapas de Galicia (Figuras 3.18 y 3.19). Cuanto mayor sea IM_AJENA más oscura estará coloreada esa área. Como las estimaciones directas tienen los mismos tonos de verde en casi todas las áreas respecto a las estimaciones basadas en el modelo.

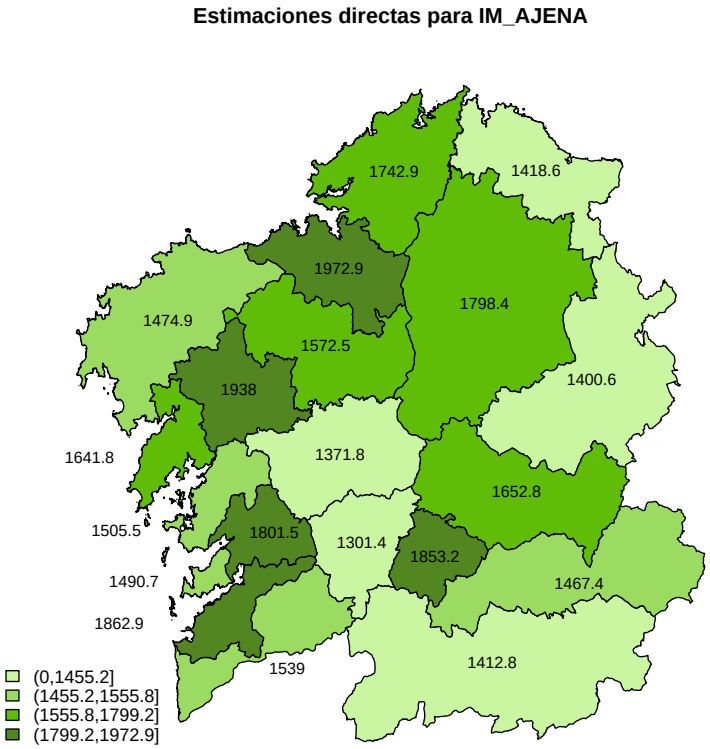

Figura 3.18: Mapa de las estimaciones directas, irán en colores más claros los IM_AJENA más bajos y en colores más oscuros los IM_AJENA más altos.

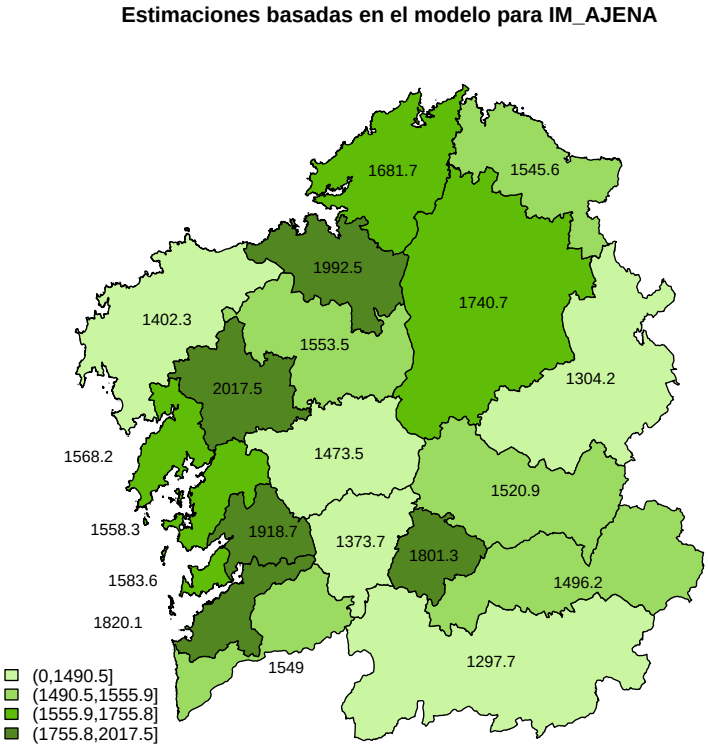

Figura 3.19: Mapa de las estimaciones basadas en el modelo, irán en colores más claros los IM_AJENA más bajos y en colores más oscuros los IM_AJENA más altos.

3.4. IM_PROPIA

La tercera variable de interés es el ingreso medio mensual procedente de ingresos por cuenta propia en el hogar, IM_PROPIA, para las veinte áreas en las que dividíamos la Comunidad de Galicia en el Capítulo 1. Es decir, se quiere conocer el verdadero valor de IM_PROPIA, μ_{3d}:

$$\mu_{3d} = X_d\boldsymbol{\beta} + u_d \qquad \forall d = 1, ..., D$$

con D=20.
Partiendo de las veinte estimaciones directas de IM_PROPIA, y_{3d}, calculadas de acuerdo a la expresión (1.1); mediante métodos SAE trataremos de obtener un buen estimador de μ_{3d}, $\hat{\mu}_{3d}$:

$$\hat{\mu}_{3d} = X_d\hat{\boldsymbol{\beta}} + \hat{u}_d \qquad \forall d = 1, .., D.$$

Al igual que hicimos para IMTOT e IM_AJENA, para estimar μ_{3d} se llevará a cabo el siguiente proceso: estudio de la normalidad de IM_PROPIA, correlación de IM_PROPIA con las variables explicativas de la Sección 3.1 y selección del mejor modelo mediante el AIC para modelos mixtos partiendo de un modelo saturado con las variables explicativas más relacionadas con IM_PROPIA.

3.4.1. Normalidad

Para comprobar la normalidad de IM_PROPIA se realizá el histograma y el correspondiente gráfico qqPlot junto con el p-valor obtenido en el contraste de normalidad Shapiro-Wilk.

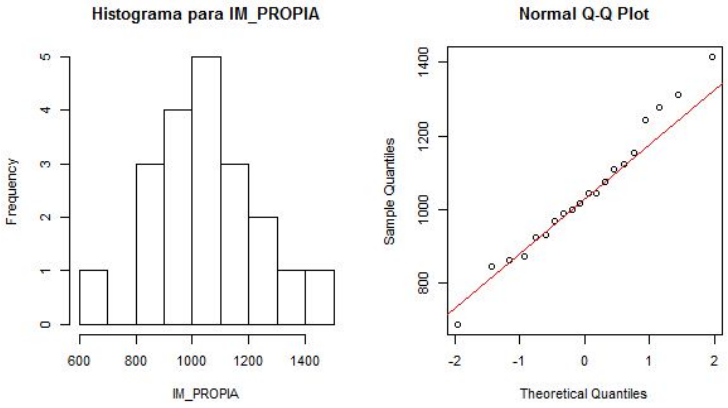

Figura 3.20: Histograma y gráfico qq-Plot del IM_AJENA.

Como se observa el la Figura 3.20 IM_PROPIA sigue claramente una distribución normal. De hecho en el contraste de Shapiro-Wilks tenemos un p-valor de 0.97.

3.4.2. Correlación

Como veremos en las Figuras 3.21, 3.22 y 3.2, el IM_PROPIA está principalmente relacionado con: NPER_65, NPER_18 y NPER_18A64; NPER_PRIM y NPER_SUP; y RENDIMIENTO.

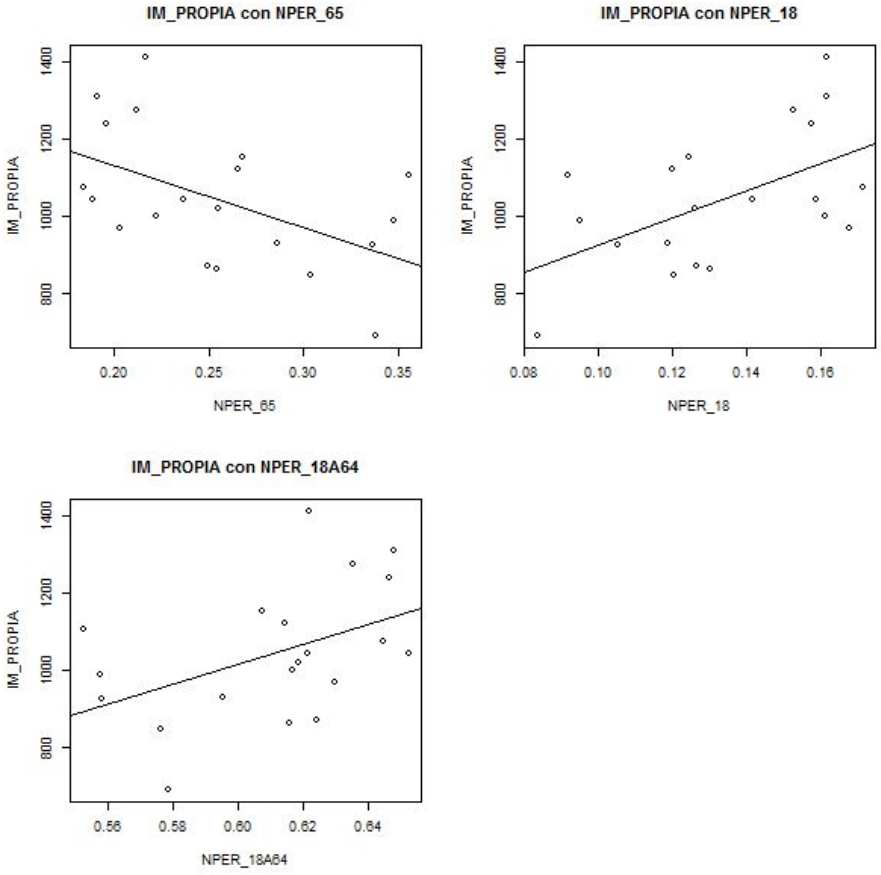

Figura 3.21: Gráfico de dispersión de IM_PROPIA con EDAD: NPER_18, NPER_18A64 Y NPER_65.

En el grupo de variables relacionadas con la EDAD se elegirán para el modelo las variables NPER_18 y NPER_65, tomando como categoría de referencia NPER_18A64.

Como vimos para IMTOT e IM_AJENA tomábamos como categoría de referencia NPER_18, en cambio para IM_PROPIA tomamos NPER_18A64. Esto dependerá de las categorías que más estén relacionadas con la respuesta y no podremos tomar al principio del estudio directamente una categoría de referencia.

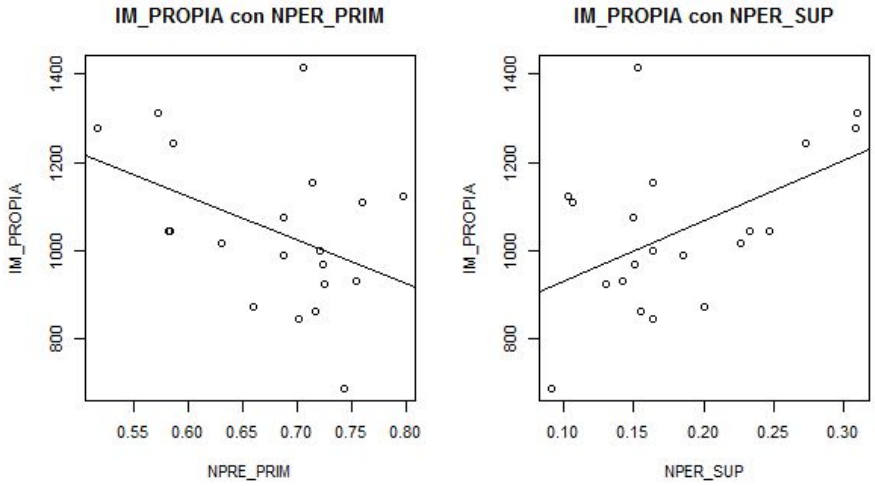

Figura 3.22: Gráficos de dispersión de IM_PROPIA con NPER_PRIM y NPER_SUP.

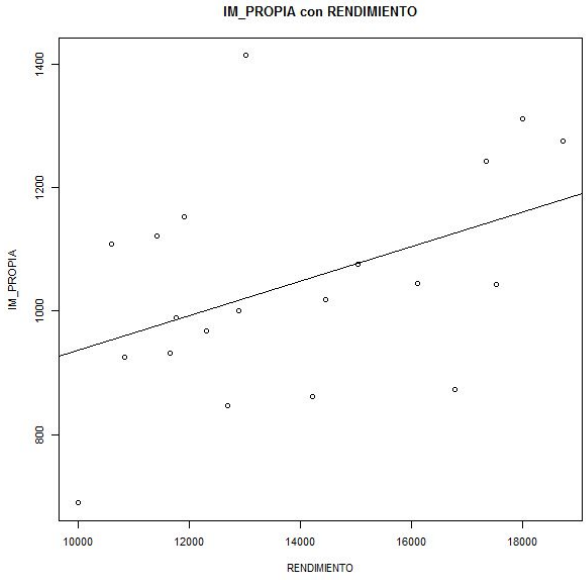

Figura 3.23: Gráficos de dispersión de IM_PROPIA con RENDIMIENTO.

En el Cuadro 3.12 se puede ver las correlaciones de IM_PROPIA con las variables que introduciremos en el modelo en la Subsección 3.4.3. Además hemos introducido las correlaciones de las variables explicativas entre sí. Como criterio, se eligieron las variables explicativas que tuviesen un coeficiente de correlación superior al 40 % con IM_PROPIA. Como vemos no podemos tomar el mismo criterio que tomábamos para

IMTOT e IM_AJENA (coeficientes superiores al 50 %) ya que IM_PROPIA no está muy relacionado con las variables explicativas de las que partimos.

	IM_PROPIA	NPER_18	NPER_65	NPER_PRIM	NPER_SUP	RENDIMIENTO
NPER_18	**0,54**	X	-0,96	-0,49	0,51	0,65
NPER_65	**-0,51**	-0,96	X	0,56	-0,58	-0,74
NPER_PRIM	**-0,41**	-0,49	0,56	X	-0,95	-0,90
NPER_SUP	**0,49**	0,51	-0,58	-0,95	X	-0,90
RENDIMIENTO	**0,43**	0,65	-0,74	-0,90	-0,90	X

Cuadro 3.12: Coeficiente de correlación del IM_PROPIA con las variables que consideramos para el modelo. Incluimos también las correlaciones de las variables explicativas entre sí.

3.4.3. Selección del mejor modelo mediante el AIC para MM

Se vuelve a seguir el mismo procedimiento para elegir el mejor modelo para explicar IM_PROPIA, se empieza por el modelo saturado con las variables elegidas en la Subsección 3.4.2 hasta conseguir un modelo con el mejor AIC y donde todas las variables son significativas. Solo el modelo 5 tiene todas las variables significativas.

	AIC			VARIABLES EXPLICATIVAS						
	cAIC1	cAIC2	mAIC	CTE	NPER_18	NPER_65	NPER_PRIM	NPER_SUP	RENDIMIENTO	$\hat{\sigma}_u^2$
1.	255,47	257,76	259,96	X	X	X	X	X	X	3185,10
2.	254,17	256,23	258,94	X	X		X	X	X	3909,64
3.	252,71	255,01	256,68		X		X	X	X	3076,74
4.	252,83	254,70	257,62		X		X	X		5089,72
5.	252,82	254,63	257,10				X	X		4864,49

Cuadro 3.13: Tabla con los cinco mejores modelos para IM_PROPIA en los que hemos indicado que variable incluimos en cada uno de ellos y hemos calculado el cAIC1, cAIC2 y mAIC. Por último la estimación de la varianza del efecto aleatorio para cada modelo.

Para IM_PROPIA tanto el cAIC1 como el cAIC2 eligen como mejor modelo aquel que tiene todas las variables significativas; en cambio mAIC eligen un modelo con más variables donde todas no son significativas. Cabe destacar que el mAIC elige como mejor modelo aquel donde tenemos menos variabilidad del efecto aleatorio (Modelo 3). Pero como estamos trabajando con modelos con efecto aleatorio donde nos interesa estudiar cada área en particular utilizaremos el cAIC2 para elegir el mejor modelo.

Si observamos los grados de libertad, cuánto mayor sea la variabilidad del efecto aleatorio también más aumentan los grados de libertad. Por ejemplo, para el Modelo 3 tenemos cuatro variables explicativas y 8,42 grados de libertad; en cambio, para el Modelo 4 con tres variables explicativas tenemos mayor número de grados de libertad, esto es debido a que también aumenta la variabilidad del efecto aleatorio.

	Grados de libertad de los modelos con cAIC1	Grados de libertad de los modelos de cAIC2	Grados de libertad de los modelos con mAIC	$\hat{\sigma}_u^2$
1.	8,90	10,05	7	3185,10
2.	8,62	9,65	6	3909,64
3.	7,27	8,42	5	3076,74
4.	8,10	9,04	4	5089.72
5.	7,33	8,34	3	4864,49

Cuadro 3.14: Tabla con los cinco mejores modelos para IM_PROPIA en los que hemos indicado sus grados de libertad en función del AIC calculado y σ_u^2 de cada modelo.

Finalmente, el mejor modelo para el IM_PROPIA es un modelo con efectos aleatorios cuyas variables explicativas son NPER_PRIM y NPER_SUP. Ambas variables explicativas son significativamente distintas de cero; además tenemos que la varianza del efecto aleatorio es de 4864,49. Los grados de libertad del modelo son de 8,3, no es un modelo muy complejo si lo comparamos con el resto de modelos descartados.

Tenemos un modelo con una variabilidad del efecto aleatorio bastante alta respecto al valor de los datos con los que estamos trabajando para IM_PROPIA, con una media de 1013,2 unidades. Esto se debe a que, como se veía en el Cuadro 3.12, los coeficientes de correlación de las variables explicativas con IM_PROPIA eran bastante bajos y no eran capaces de explicar bien IM_PROPIA.

VARIABLE EXPLICATIVA	ESTIMACIÓN	P-VALOR
NPER_PRIM	862,0348	0
NPER_SUP	2348,0721	0
σ_u^2	4864,49	

Cuadro 3.15: Tabla con las estimaciones para las variables explicativas del modelo elegido para IM_PROPIA junto con sus p-valores y la varianza del efecto aleatorio del modelo.

El modelo obtenido para conocer el verdadero valor de IM_PROPIA es:

$$\hat{\mu}_{3d} = 862,03\, NPER\_PRIM + 2348,07\, NPER\_SUP + \hat{u}_d \qquad \forall d = 1, ..., D.$$

Cuanto mayores sean NPER_PRIM y NPER_SUP mayor será IM_AJENA. Habrá que considerar también el efecto aleatorio del área a través del $\hat{u}_d$.

A continuación mostraremos (Cuadro 3.16) las estimaciones directas, $\hat{y}_{3d}$, y las basadas en el modelo, $\hat{\mu}_{3d}$ para IM_PROPIA que representaremos en los gráficos de mapas. En el mismo cuadro añadiremos los coeficientes de correlación de las estimaciones directas y de las basadas en el modelo, CV_1 y CV_2, respectivamente.

Áreas	$\hat{y}_{3d}$	$\hat{\mu}_{3d}$	CV_1	CV_2
Área 1	1152,10	1061,80	7,32 %	6,30 %
Área 2	872,78	948,73	7,36 %	6,15 %
Área 3	1121,76	998,90	8,21 %	7,39 %
Área 4	1000,18	1003,60	13,54 %	7,49 %
Área 5	1275,99	1205,47	7,49 %	6,63 %
Área 6	1310,85	1232,37	12,74 %	7,24 %
Área 7	989,5220	1005,98	5,95 %	5,49 %
Área 8	689,26	816,42	18,43 %	9,79 %
Área 9	1018,45	1057,55	10,72 %	6,90 %
Área 10	861,71	943,54	11,82 %	7,56 %
Área 11	925,73	930,05	13,31 %	8,12 %
Área 12	847,07	964,02	17,84 %	7,81 %
Área 13	1108,00	923,40	19,43 %	8,99 %

Área 14	1044,37	1064,43	7,54 %	6,29 %
Área 15	931,37	961,88	8,68 %	6,96 %
Área 16	1413,67	1005,70	15,69 %	7,65 %
Área 17	968,21	974,16	8,95 %	6,98 %
Área 18	1075,51	960,41	17,31 %	7,95 %
Área 19	1241,77	1161,14	12,93 %	7,13 %
Área 20	1043,02	1046,02	7,53 %	6,32 %

Cuadro 3.16: Tabla con las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_PROPIA en cada área. Añadimos también CV_1 y CV_2.

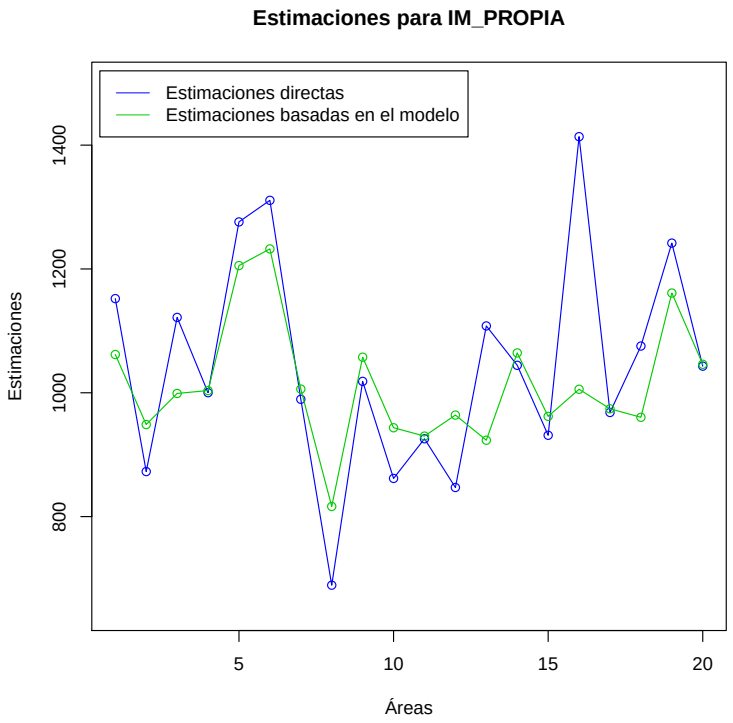

Figura 3.24: Gráfico de las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_PROPIA.

Como se observa en la Figura 3.24 las estimaciones basadas en el modelo no son muy similares a las estimaciones directas; ya que como podemos ver con el gráfico boxplot de los coeficientes de variación (Figura 3.25) las estimaciones basadas en el modelo son mucho más precisas que las estimaciones directas ya que tenemos coeficientes de correlación mucho más bajos.

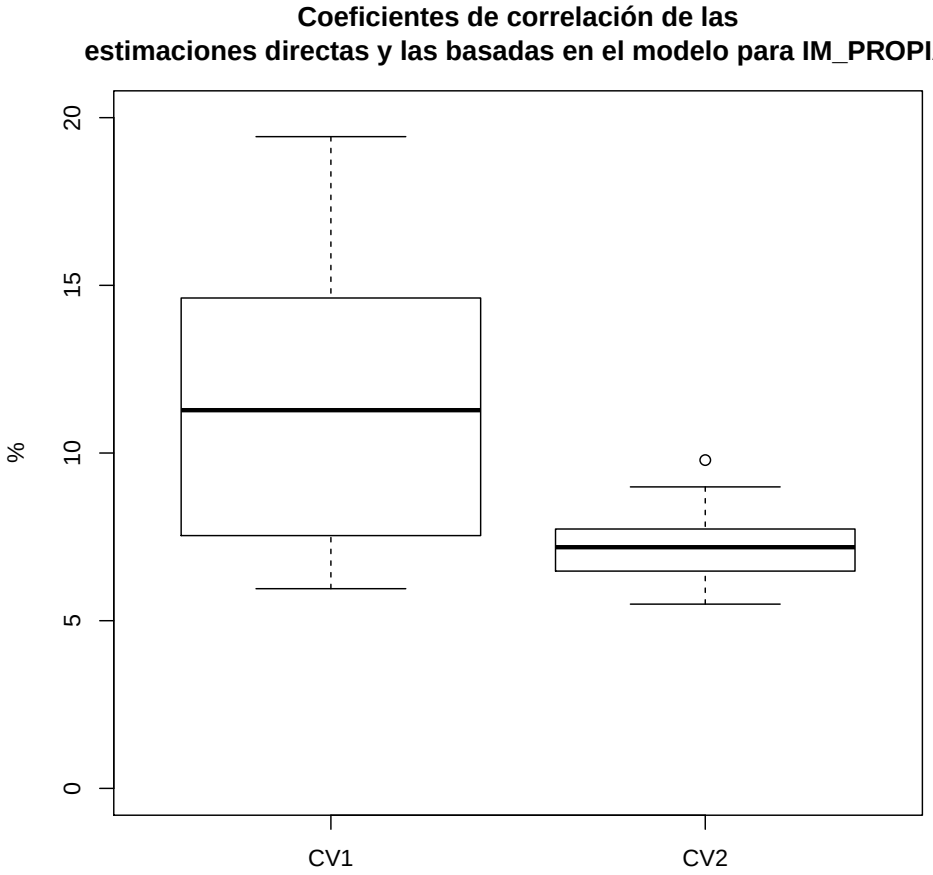

Figura 3.25: Gráfico boxplot de los coeficientes de variación para las estimaciones de IM_PROPIA.

Por último, en las Figuras 3.26 y 3.27 podemos ver dos mapas de Galicia divididos en las veinte áreas de interés en los que salen representados las estimaciones directas y las basadas en el modelo, respectivamente en cada mapa. Cuanto mayor sean esas estimaciones, más oscura aparecerá coloreada esa área.

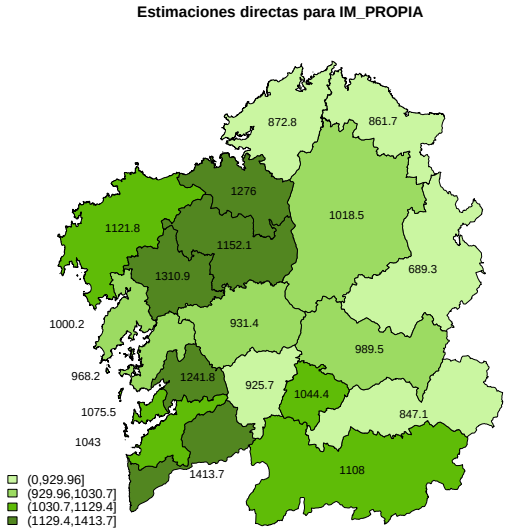

Figura 3.26: Mapa de las estimaciones directas, irán en colores más claros los IM_PROPIA más bajos y en colores más oscuros los IM_PROPIA más altos.

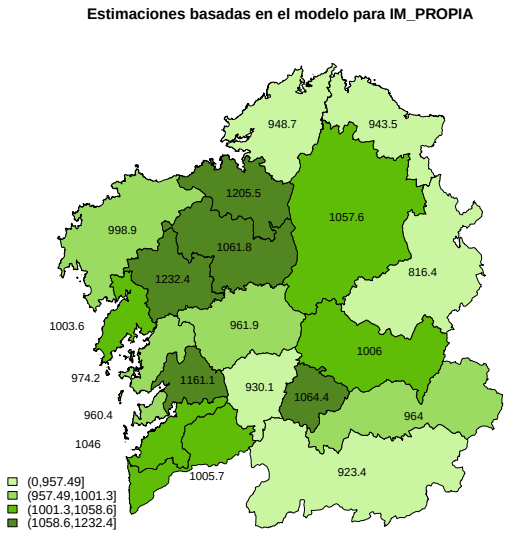

Figura 3.27: Mapa de las estimaciones basadas en el modelo, irán en colores más claros los IM_PROPIA más bajos y en colores más oscuros los IM_PROPIA más altos.

3.5. IM_CONTRIB

La cuarta variable de interés es el ingreso medio mensual en el hogar procedente de prestaciones contributivas, IM_CONTRIB, en las veinte áreas gallegas. Para ello utilizaremos, al igual que en el resto de variables respuesta, los datos del 2013.

Para conocer el verdadero valor de IM_CONTRIB, μ_{4d}:

$$\mu_{4d} = X_d\boldsymbol{\beta} + u_d \qquad \forall d = 1, ..., D \qquad D = 20.$$

Partimos de las veinte estimaciones directas de IM_CONTRIB, y_{4d}, calculadas en la expresión (1.1) e introducidas en el siguiente modelo:

$$y_{4d} = X_d\boldsymbol{\beta} + u_d + \epsilon_d \qquad \forall d = 1, ..., D$$

donde $\epsilon_d \sim N(0, \sigma^2_{\epsilon d})$ con $\sigma^2_{\epsilon d}$ conocida y $N(0, \sigma^2_u)$ con σ^2_u desconocida.

Mediante métodos SAE trataremos de obtener un buen estimador de μ_{4d}, $\hat{\mu}_{4d}$:

$$\hat{\mu}_{4d} = X_d\hat{\boldsymbol{\beta}} + \hat{u}_d.$$

En esta Sección se estudiará la normalidad de IM_CONTRIB, la correlación de IM_CONTRIB con las veintiuna variables explicativas de la Sección 3.1; y por último, se llevará a cabo la selección del mejor modelo mediante el AIC para modelos mixtos partiendo del modelo saturado con las variables explicativas más relacionadas con IM_CONTRIB.

3.5.1. Normalidad

Para comprobar la normalidad de IM_CONTRIB realizaremos el histograma y el correspondiente gráfico qqPlot junto con el p-valor obtenido en el contraste de normalidad Shapiro-Wilk. En el contraste de normalidad tenemos un p-valor de 0.24 por lo que podemos suponer que IM_CONTRIB sigue una distribución normal. Con la Figura 3.28 , histograma y correspondiente qqPlot, se reafirma la normalidad de la variable IM_CONTRIB.

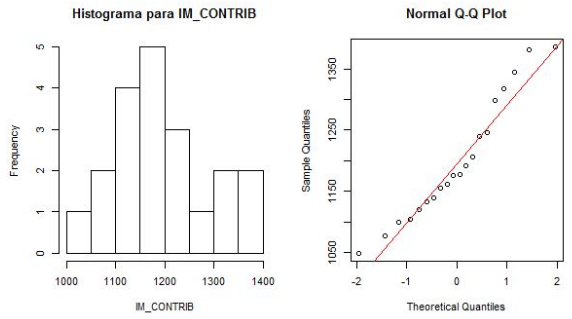

Figura 3.28: Histograma y gráfico qq-Plot del IM_CONTRIB.

3.5.2. Correlación

El siguiente paso para la estimación de μ_{4d} es ver aquellas variables más correlacionadas con IM_CONTRIB. Como criterio se tomarán las variables que tienen un coeficiente de correlación superior al 40 % (ver Cuadro 3.17).

Como veremos en las Figuras 3.29, 3.30 y 3.31, IM_CONTRIB está principalmente relacionado con: NPER_65, NPER_18 y NPER_18A64; NPER_PRIM, NPER_SEC y NPER_SUP; RENDIMIENTO y PENSIONES.

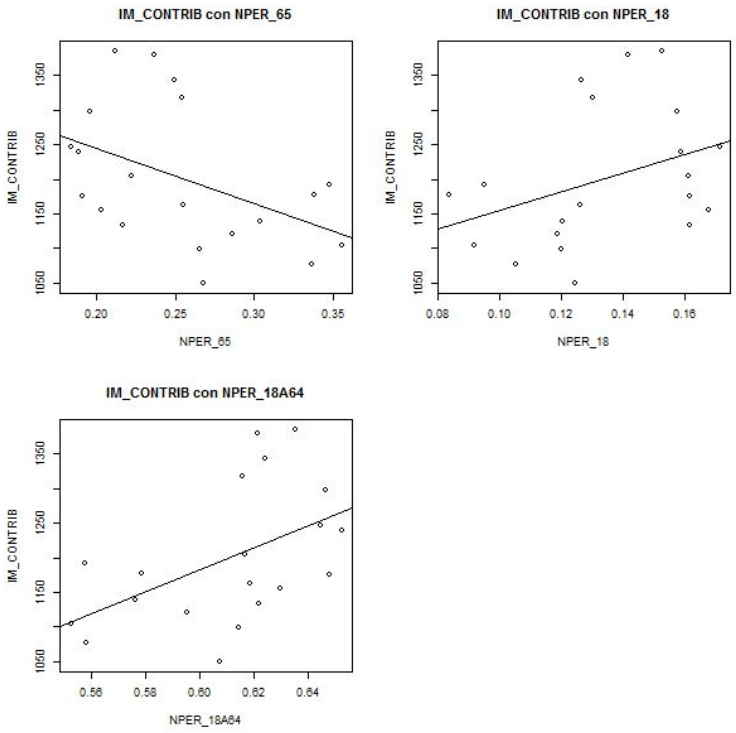

Figura 3.29: Gráfico de dispersión de IM_CONTRIB con EDAD: NPER_18, NPER_18A64 Y NPER_65.

En el grupo de las variables del grupo EDAD, a pesar de estar las tres variables muy relacionadas con IM_CONTRIB, para introducir en el modelo nos quedamos con NPER_18A64 y NPER_65 ya que de las tres son las más correlacionadas con IM_CONTRIB.

Para las variables del grupo ESTUDIOS (Figura 3.30) tomaremos como categoría de referencia NPER_SEC, introduciendo en el modelo solo NPER_PRIM y NPER_SUP y

evitar así la colinealidad. En las tres variables respuesta anteriores no teníamos este problema porque NPER_SEC no estaba prácticamente relacionado con la respuesta.

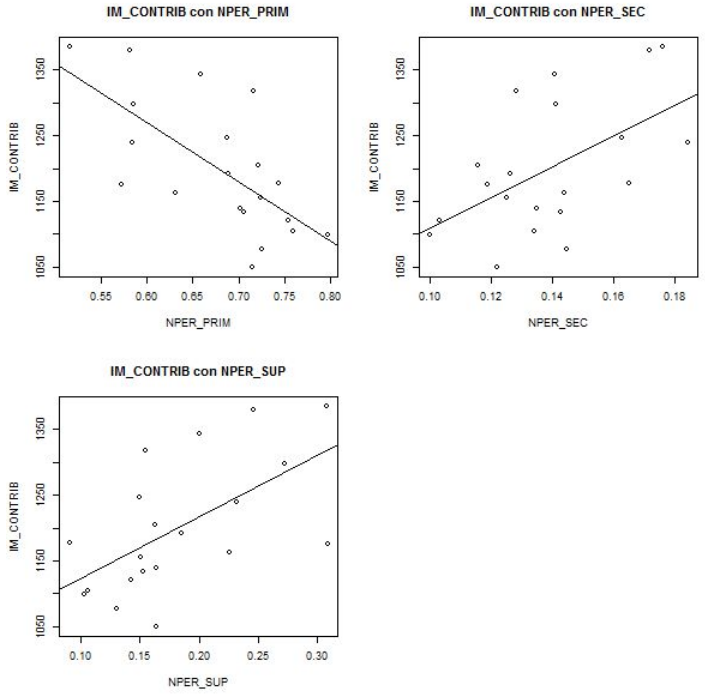

Figura 3.30: Gráficos de dispersión de IM_CONTRIB con NPER_PRIM, NPER_SEC y NPER_SUP.

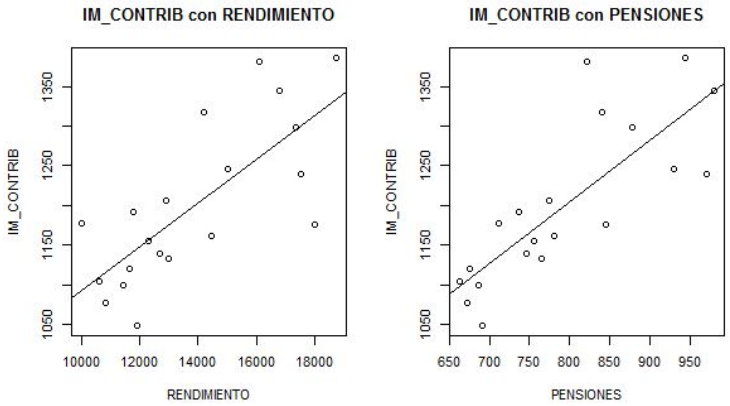

Figura 3.31: Gráficos de dispersión de IM_CONTRIB con RENDIMIENTO y PENSIONES.

En el Cuadro 3.17 se puede ver las correlaciones de IM_CONTRIB con las variables que introduciremos en el modelo en la Subsección 3.5.3. Además hemos introducido las correlaciones de las variables explicativas entre sí.

	IM_CONTRIB	NPER_18A65	NPER_65	NPER_PRIM	NPER_SUP	RENDIMIENTO	PENSIONES
NPER_18A64	**0,49**	X	-0,97	-0,58	0,60	0,78	0,74
NPER_65	**-0,44**	-0,97	X	0,56	-0,58	-0,74	-0,70
NPER_PRIM	**-0,67**	-0,58	0,56	X	-0,95	-0,90	-0,73
NPER_SUP	**0,59**	0,60	-0,58	-0,95	X	0,90	0,65
RENDIMIENTO	**0,74**	0,78	-0,74	-0,90	0,90	X	0,89
PENSIONES	**0,80**	0,74	-0,70	-0,73	0,65	0,89	X

Cuadro 3.17: Coeficiente de correlación del IM_CONTRIB con las variables que consideramos para el modelo. Incluimos también las correlaciones de las variables explicativas entre sí.

3.5.3. Selección del mejor modelo mediante el AIC para MM

Seguimos el procedimiento usado en el resto de variables, partimos del modelo saturado y vamos eliminando variables para conseguir un modelo con el AIC más pequeño y donde todas las variables sean significativas. En este caso para seleccionar el mejor modelo utilizaremos el cAIC2 ya que los seis mejores modelos tienen efecto aleatorio y en ninguno de ellos conocíamos la varianza de ese efecto aleatorio. El único modelo que cumple la hipótesis de tener el menor AIC2 es el modelo 6; además todas sus variables son significativamente distintas de cero.

	AIC			VARIABLES EXPLICATIVAS							
	cAIC1	cAIC2	mAIC	CTE	NPER_18A64	NPER_65	NPER_PRIM	NPER_SUP	RENDIMIENTO	PENSIONES	$\hat{\sigma}_u^2$
1.	227,68	228,99	236,84	X	X	X	X	X	X	X	2446,50
2.	226,51	227,95	234,4	X		X	X	X	X	X	2140,90
3.	225,36	226,91	232,32	X		X	X	X		X	1912,09
4.	226,14	225,84	230,05	X		X	X			X	1652,60
5.	223,61	225,30	229,42	X		X				X	1784,83
6.	223,04	224,80	228,45	X						X	1787,26

Cuadro 3.18: Tabla con los cinco mejores modelos para IM_CONTRIB en los que hemos indicado que variable incluimos en cada uno de ellos y hemos calculado el cAIC1, cAIC2 y mAIC. Por último la estimación de la varianza del efecto aleatorio para cada modelo.

Los grados de libertad para todas los modelos son bastante altos en relación con el número de variables que introducimos al modelo. Esto tiene relación con la variabilidad del efecto aleatorio que introducimos en cada uno de los modelos desde el 1 hasta el modelo finalmente elegido, el 6.

	Grados de libertad de los modelos con cAIC1	Grados de libertad de los modelos de cAIC2	Grados de libertad de los modelos con mAIC	$\hat{\sigma}_u^2$
1.	12,76	13,71	8	2446,50
2.	10,80	12,52	7	2140,90
3.	10,84	11,61	6	1912,09
4.	9,76	10,61	5	1652,60
5.	9,44	10,29	4	1784,83
6.	8,86	9,74	3	1787,26

Cuadro 3.19: Tabla con los cinco mejores modelos para IM_CONTRIB en los que hemos indicado sus grados de libertad en función del AIC calculado y σ_u^2 de cada modelo.

El mejor modelo para el IM_CONTRIB es un modelo con efectos aleatorios cuyas variables explicativas son CTE y PENSIONES, ambas significativamente distintas de cero. La varianza del efecto aleatorio es de 1787,264. No es un modelo muy complejo si lo comparamos con el resto de modelos descartados. De hecho es el modelo menos complejo.

La variabilidad del efecto aleatorio del modelo elegido es relativamente alta en relación con los datos que tenemos, datos con una media de 1193,4 unidades. En el Cuadro 3.17 se vio que la variable PENSIONES ya era capaz de explicar un 80 % de IM_CONTRIB, el resto se explicará con la CTE y con la ayuda de un efecto aleatorio con una variabilidad de 1787,26.

Obtenemos el siguiente modelo para explicar IM_CONTRIB:

$$\hat{\mu}_{4d} = 571,49\,\text{CTE} + 0,738\,\text{PENSIONES} + \hat{u}_d \qquad \forall d = 1, ..., D.$$

Habiendo fijado la CTE, la variable PENSIONES tiene un peso positivo en IM_CONTRIB, bastante lógico; aunque también debemos considerar el efecto del área a través del $\hat{u}_d$.

VARIABLE EXPLICATIVA	ESTIMACIÓN	P-VALOR
CTE	571,4994	0
PENSIONES	0,783722	0
σ_u^2	1787,26	

Cuadro 3.20: Tabla con las estimaciones para las variables explicativas del modelo elegido para IM_CONTRIB junto con sus p-valores y la varianza del efecto aleatorio del modelo.

A continuación mostraremos las estimaciones directas, las estimaciones basadas en el modelo finalmente elegido y sus respectivos coeficientes de variación, CV_1 y CV_2.

Áreas	$\hat{y}_{4d}$	$\hat{\mu}_{4d}$	CV_1	CV_2
Área 1	1048,7	1082,9	4,31 %	3,42 %
Área 2	1343,86	1341,6	3,06 %	2,71 %
Área 3	1098,5	1106	5,21 %	3,70 %
Área 4	1205,3	1182,7	8,07 %	3.70 %
Área 5	1385,9	1349,9	2,93 %	2,63 %
Área 6	1175,1	1206,7	3,84 %	3,00 %
Área 7	1191,4	1165,6	4,47 %	3,35 %
Área 8	1177,6	1141,7	6,15 %	3,76 %
Área 9	1162,10	1171,62	3,19 %	2,76 %
Área 10	1317,2	1247,4	6,38 %	3,46 %

Área 11	1076,6	1086,6	3,62 %	3,16 %
Área 12	1139,6	1148,1	3,76 %	3,08 %
Área 13	1103,9	1095,3	5,16 %	3,78 %
Área 14	1380,7	1282,8	3,69 %	2,96 %
Área 15	1120,0	1105,6	6,42 %	3,97 %
Área 16	1132,4	1156,8	4,82 %	3,38 %
Área 17	1154,3	1159,7	4,25 %	3,24 %
Área 18	1246,0	1289,8	6,80 %	3,53 %
Área 19	1298,7	1268,7	575 %	3,39 %
Área 20	1239,6	1278,2	2.86 %	2,59 %

Cuadro 3.21: Tabla con las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_CONTRIB en cada área. Añadimos también CV_1 y CV_2.

Como se observa en la Figura 3.32 aunque las estimaciones directas son muy similares a las estimaciones basadas en el modelo, podría haber ciertas diferencias en las áreas 7, 8, 10 y en torno al área 15 donde también hay ciertas diferencias en sus coeficientes de correlación. Aún así, como se observa en la Figura 3.33 ambos coeficientes de correlación, CV_1 y CV_2, son muy similares siendo más bajos los asociados a las estimaciones basadas en el modelo. Podemos concluir que las estimaciones basadas en el modelo son más precisas que las estimaciones directas.

En las Figuras 3.34 y 3.35 aparecen representados los mapas de Galicia junto con las estimaciones directas y las basadas en el modelo para cada área, ya escritas en el Cuadro 3.21. Cuanto más bajas son los valores de las estimaciones más claro es el color de ese área y para estimaciones con valores mas altos de IM_CONTRIB más oscuro aparece coloreada ese área. Como vemos el mapa de las estimaciones directas tiene prácticamente para todas las áreas tono similar al mapa de las estimaciones basadas en el modelo.

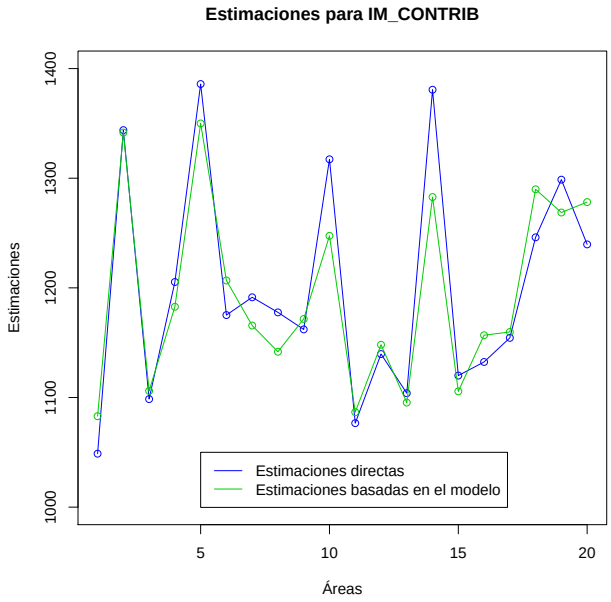

Figura 3.32: Gráfico de las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_CONTRIB.

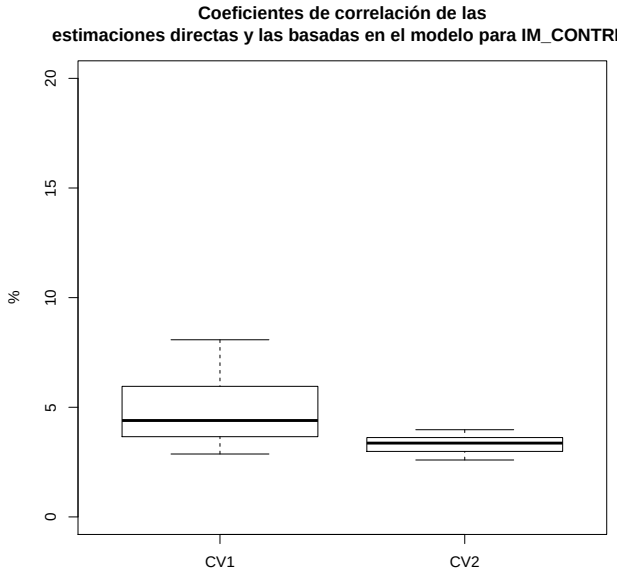

Figura 3.33: Gráfico boxplot de los coeficientes de variación para las estimaciones de IM_CONTRIB.

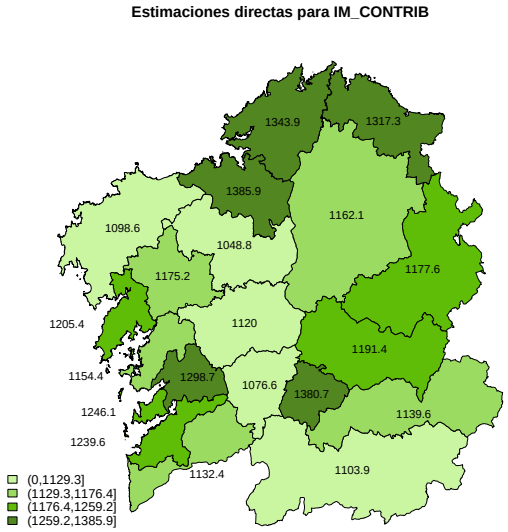

Figura 3.34: Mapa de las estimaciones directas, irán en colores más claros los IM_CONTRIB más bajos y en colores más oscuros los IM_CONTRIB más altos.

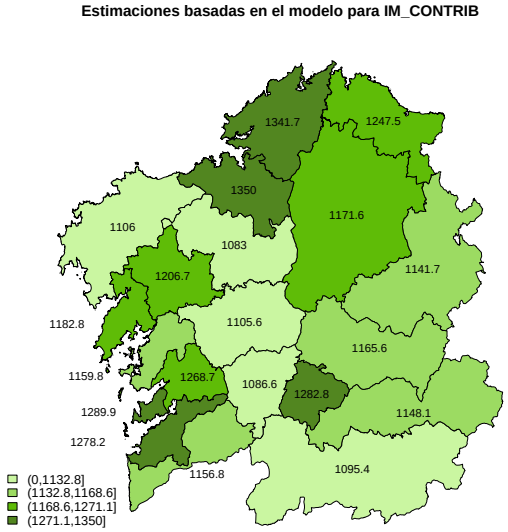

Figura 3.35: Mapa de las estimaciones basadas en el modelo, irán en colores más claros los IM_CONTRIB más bajos y en colores más oscuros los IM_CONTRIB más altos.

3.6. IM_NO_CONTRIB

La última variable de estudio es el ingreso medio mensual en el hogar procedente de prestaciones no contributivas en las veinte áreas gallegas es IM_NO_CONTRIB.

Partiendo de las veinte estimaciones directas de IM_NO_CONTRIB, y_{5d}, una para cada área. Se tratará de estimar el verdadero valor de IM_NO_CONTRIB, μ_{5d}, mediante un estimador más preciso que el directo, el estimador basado en el modelo:

$$\hat{\mu}_{5d} = X_d\hat{\boldsymbol{\beta}} + \hat{u}_d \qquad \forall d = 1, .., D.$$

Se utilizaran métodos de estimación en áreas pequeñas. Para ello, se comprobará la normalidad de la variable de interés, IM_NO_CONTRIB, necesaria para los modelos Fay-Herriot; después veremos aquellas variables más correlacionadas con la variable respuesta, las seleccionadas serán introducidas en el modelo saturado con el que comenzaremos la selección del mejor modelo mediante el criterio del AIC en la Subsección 3.6.2.

3.6.1. Normalidad

Para comprobar la normalidad de IM_NO_CONTRIB se fijará un nivel de significación del 1 %. Realizaremos el histograma y el correspondiente gráfico qqPlot (Figura 3.36) junto con el p-valor obtenido en el contraste de normalidad Shapiro-Wilk obteniendo un p-valor de 0.013. Podemos concluir que IM_NO_CONTRIB sigue una distribución normal.

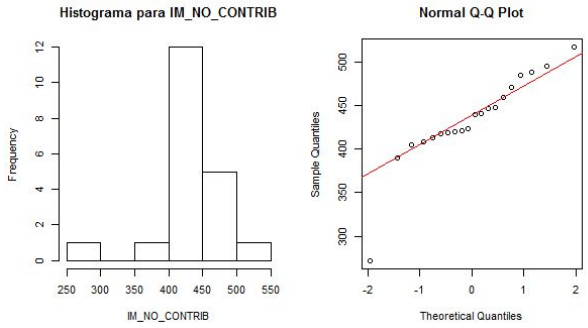

Figura 3.36: Histograma y gráfico qq-Plot del IM_NO_CONTRIB.

3.6.2. Correlación

El siguiente paso para la estimación de μ_{5d} es ver aquellas variables más correlacionadas con IM_NO_CONTRIB.

A pesar de que IM_NO_CONTRIB no está altamente relacionada con ninguna variable, como veremos en las Figuras 3.37, 3.38 y 3.39, las variables que tienen mayor correlación con IM_NO_CONTRIB son: NPER, y, en consecuencia, UC; las variables relacionadas con el HOG_UMBRAL_POBREZA (HOG_BAJO_UMBRAL y HOG_SOBRE_UMBRAL); y las variables del grupo NACIONALIDAD (NPER_ESP y NPER_NO_ESP).

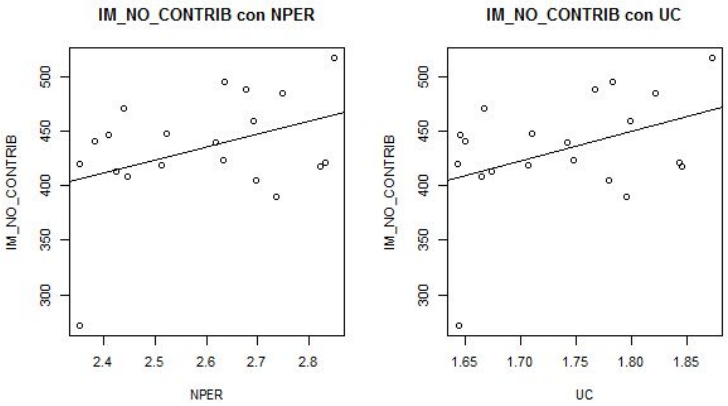

Figura 3.37: Gráfico de dispersión de IM_NO_CONTRIB con NPER y UC.

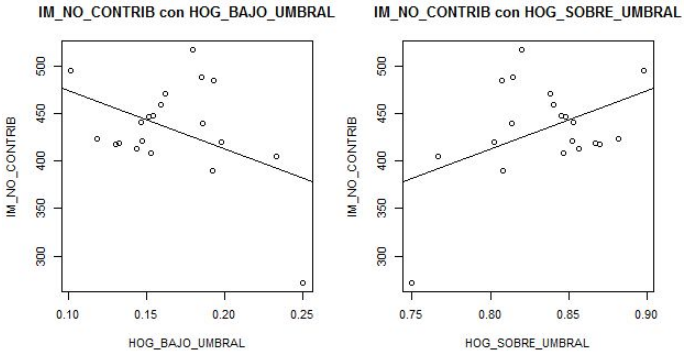

Figura 3.38: Gráficos de dispersión de IM_NO_CONTRIB con HOG_BAJO_UMBRAL y HOG_SOBRE_UMBRAL.

Como HOG_BAJO_UMBRAL es el opuesto de HOG_SOBRE_UMBRAL, tomaremos como categoría de referencia IM_SOBRE_UMBRAL, introduciendo al modelo IM_BAJO_UMBRAL.

Para el grupo de variables NACIONALIDAD ocurre lo mismo, tenemos que tomar como categoría de referencia NPER_NO_ESP.

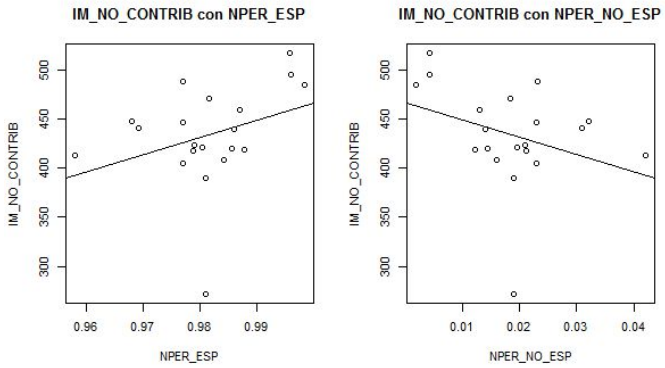

Figura 3.39: Gráficos de dispersión de IM_NO_CONTRIB con NPER_ESP y NPER_NO_ESP.

En el Cuadro 3.22 se puede ver las correlaciones de IM_NO_CONTRIB con las variables que introduciremos en el modelo en la Subsección 3.5.3. Además hemos introducido las correlaciones de las variables explicativas entre sí. En este caso hemos elegido aquellas variables cuya correlación con IM_NO_CONTRIB es superior al 30 %.

	IM_NO_CONTRIB	NPER	UC	HOG_BAJO_UMBRAL	NPER_ESP
NPER	**0,38**	X	0,98	-0,07	0,39
UC	**0,40**	0,98	X	-0,08	0,44
HOG_BAJO_UMBRAL	**-0,43**	-0,07	-0,08	X	0,06
NPER_ESP	**0,30**	0,39	0,44	0,06	X

Cuadro 3.22: Coeficiente de correlación del IM_NO_CONTRIB con las variables que consideramos para el modelo. Incluimos también las correlaciones de las variables explicativas entre sí.

3.6.3. Selección del mejor modelo mediante el AIC para MM

Seguimos el procedimiento usado en el resto de variables, partimos del modelo saturado y vamos eliminando variables hasta conseguir un modelo con el AIC más pequeño y donde todas las variables sean significativas.

	AIC			VARIABLES EXPLICATIVAS					
	cAIC1	cAIC2	mAIC	CTE	NPER	UC	HOG_BAJO_UMBRAL	NPER_ESP	$\hat{\sigma}_u^2$
1.	213,55	215,66	218,44	X	X	X	X	X	541,99
2.	212,14	214,50	216,14	X		X	X	X	421,95
3.	211,1	213,58	214,70	X			X	X	383,51
4.	210,51	212,96	213,95				X	X	406,59

Cuadro 3.23: Tabla con los cinco mejores modelos para IM_NO_CONTRIB en los que hemos indicado que variable incluimos en cada uno de ellos y hemos calculado el cAIC1, cAIC2 y mAIC. Por último la estimación de la varianza del efecto aleatorio para cada modelo.

Partimos de un modelo saturado (Modelo 1) con la CTE y las cuatro variables más relacionadas con IM_NO_CONTRIB. Vamos eliminando variables hasta encontrar el modelo con el menor AIC y donde todas sus variables sean significativas. Como los cuatro mejores modelos para conocer el verdadero valor de IM_NO_CONTRIB (Cuadro 3.23) tienen efecto aleatorio con variabilidad desconocida, para la selección del mejor modelo utilizaremos el cAIC2. El modelo seleccionado es el 4; coincide que aquel modelo que tiene el menor cAIC2 tiene todas sus variables significativamente distintas de cero.

	Grados de libertad de los modelos con cAIC1	Grados de libertad de los modelos de cAIC2	Grados de libertad de los modelos con mAIC	$\hat{\sigma}_u^2$
1.	8,85	9,90	6	541,99
2.	7,49	8,67	5	421,45
3.	6,52	7,74	4	383,51
4.	5,88	7,11	3	406,59

Cuadro 3.24: Tabla con los cinco mejores modelos para IM_NO_CONTRIB en los que hemos indicado sus grados de libertad en función del AIC calculado y σ_u^2 de cada modelo.

El mejor modelo para el IM_NO_CONTRIB es un modelo con efectos aleatorios cuyas variables explicativas son HOG_BAJO_UMBRAL y NPER_ESP. Estas dos variables estaban poco correlacionadas con IM_NO_CONTRIB (Cuadro 3.22), por lo que fue necesario introducir un efecto aleatorio que fuese capaz de explicar la variabilidad entre las áreas que no pudo ser explicada por la parte fija del modelo. Este efecto aleatorio tiene una distribución normal de media cero y varianza 406,59; una variabilidad relativamente alta si la comparamos con los datos que tenemos, con una media de 423,1 unidades. Los grados de libertad obtenidos son 7,11 por lo que tenemos un modelo de complejidad similar al resto de modelos descartados.

VARIABLE EXPLICATIVA	ESTIMACIÓN	P-VALOR
HOG_BAJO_UMBRAL	-609,9758	0,02
NPER_ESP	534,3473	0
σ_u^2	406,59	

Cuadro 3.25: Tabla con las estimaciones para las variables explicativas del modelo elegido para IM_NO_CONTRIB junto con sus p-valores y la varianza del efecto aleatorio del modelo.

Tenemos el siguiente modelo:

$$\hat{\mu_{5d}} = -609,97\, HOG\_BAJO\_UMBRAL + 534,34\, NPER\_ESP + \hat{u}_d.$$

A continuación mostraremos las estimaciones directas, las estimaciones basadas en el modelo finalmente elegido y sus coeficientes de variación, CV_1 y CV_2 respectivamente. Las estimaciones directas y las basadas en el modelo estarán representadas frente a cada área en la Figura 3.40. El gráfico boxplot de los coeficientes de variación nos permitirá comparar ambas estimaciones(Figura 3.41).

Áreas	$\hat{y}_{4d}$	$\hat{\mu}_{4d}$	CV_1	CV_2
Área 1	517,8	427,8	15,93 %	5,42 %
Área 2	471,2	435,7	8,06 %	5,32 %
Área 3	485,1	431,2	7,76 %	5,52 %
Área 4	417,9	437,9	9,12 %	5.56 %
Área 5	447	433,7	8,09 %	5,35 %
Área 6	423,7	442,2	6,91 %	5,43 %
Área 7	413,6	421,8	8,93 %	5,54 %
Área 8	495,7	472,8	12 %	6,04 %
Área 9	419,6	437,7	6,79 %	5,27 %
Área 10	477,7	424,5	17,39 %	5,43 %
Área 11	271,8	348,6	13,57 %	8,44 %
Área 12	440,6	429,3	15,19 %	5,47 %
Área 13	419,7	408,8	9.46 %	5,94 %
Área 14	408,31	420, 5	5,01 %	4,69 %

Área 15	459,4	431,2	21,98 %	5,30 %
Área 16	405,4	383,3	12,47 %	7,44 %
Área 17	421,6	430,1	7.1 %	5,29 %
Área 18	390,1	401,6	7.52 %	5,71 %
Área 19	488,7	416,9	12,37 %	5,70 %
Área 20	440,3	427,1	4,45 %	4,53 % %

Cuadro 3.26: Tabla con las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_NO_CONTRIB en cada área. Añadimos también CV_1 y CV_2.

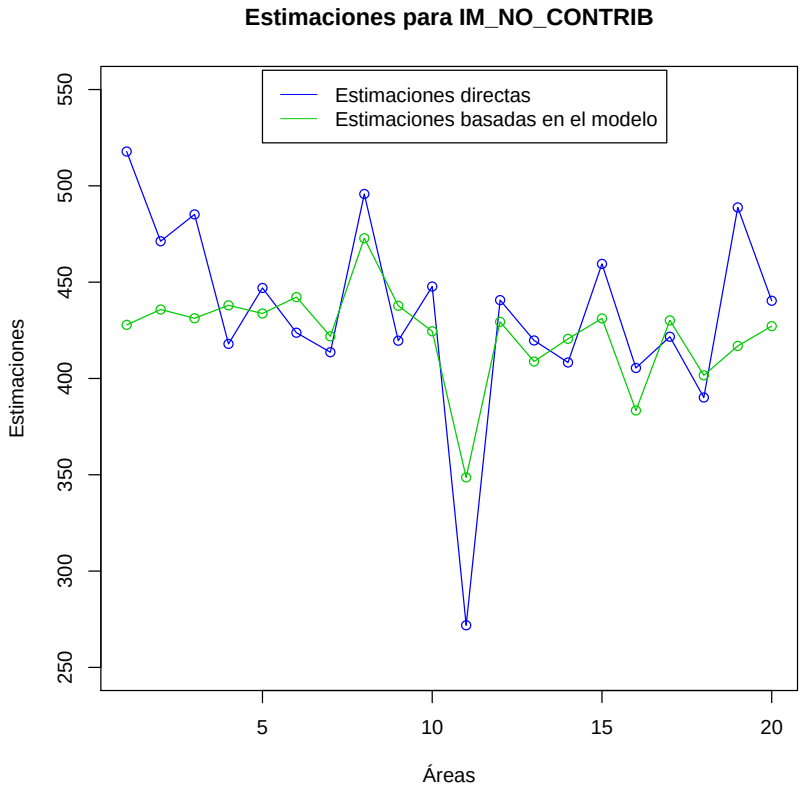

Figura 3.40: Gráfico de las estimaciones directas junto con las estimaciones basadas en el mejor modelo para IM_NO_CONTRIB.

Como podemos observar en las Figuras 3.40 y 3.41 las estimaciones basadas en el modelo no son muy similares a las directas. Además las estimaciones basadas en el modelo son mucho más precisas que las estimaciones directas ya que los coeficientes de variación son más pequeños.

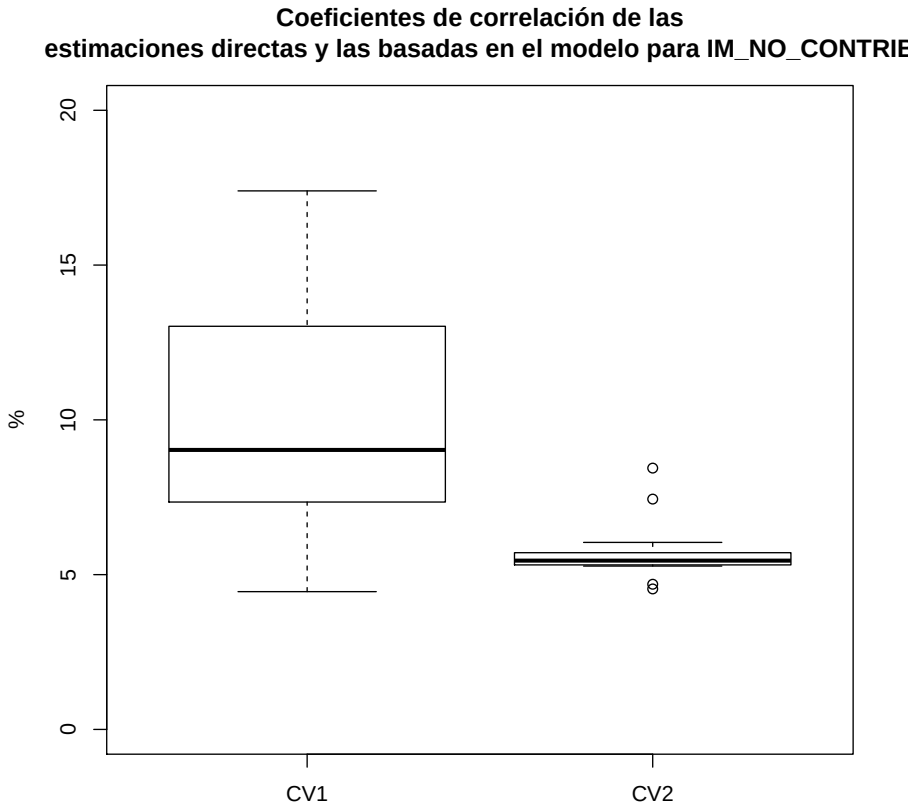

Figura 3.41: Gráfico boxplot de los coeficientes de variación para las estimaciones de IM_NO_CONTRIB.

Por último mostraremos los mapas de Galicia para las estimaciones directas y las basadas en el modelo para IM_NO_CONTRIB. Cuanto mayor sea el valor de la estimación más oscura estará coloreada esa área.

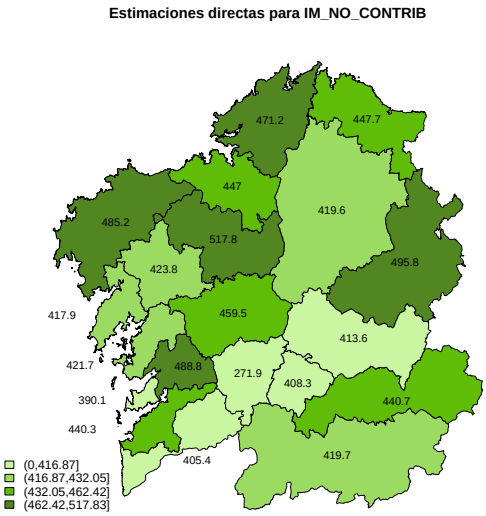

Figura 3.42: Mapa de las estimaciones directas, irán en colores más claros los IM_NO_CONTRIB más bajos y en colores más oscuros los IM_NO_CONTRIB más altos.

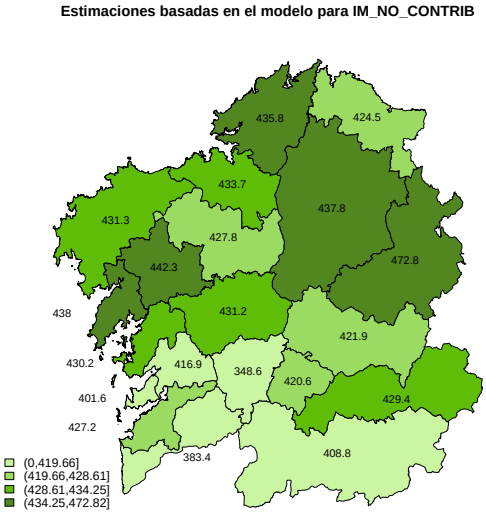

Figura 3.43: Mapa de las estimaciones basadas en el modelo, irán en colores más claros los IM_NO_CONTRIB más bajos y en colores más oscuros los IM_NO_CONTRIB más altos.

3.7. Estimaciones basadas en el modelo de IMTOT formadas por combinación de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB.

Como se comentó en el Capítulo 1, IMTOT es combinación de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB. A petición del IGE se quiere volver a combinar las estimaciones basadas en el modelo de las cuatro variables es que se divide IMTOT para obtener unas nuevas estimaciones que posiblemente sean incluso más precisas que las basadas en el modelo para IMTOT. Calcularemos estas nuevas estimaciones, a las que llamaremos **estimaciones combinadas.**

Teníamos un tamaño de muestra de 4540, 1715, 5159 y 3127 para IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB, respectivamente. De modo que las estimaciones combinadas de IMTOT vienen dadas por la siguiente expresión:

$$\frac{4540\,\hat{\mu}_{2d} + 1715\,\hat{\mu}_{3d} + 5149\,\hat{\mu}_{4d} + 3127\,\hat{\mu}_{5d}}{9188}$$

Estas nuevas estimaciones serán comparadas con las estimaciones directas y las basadas en el modelo de IMTOT en el siguiente gráfico.

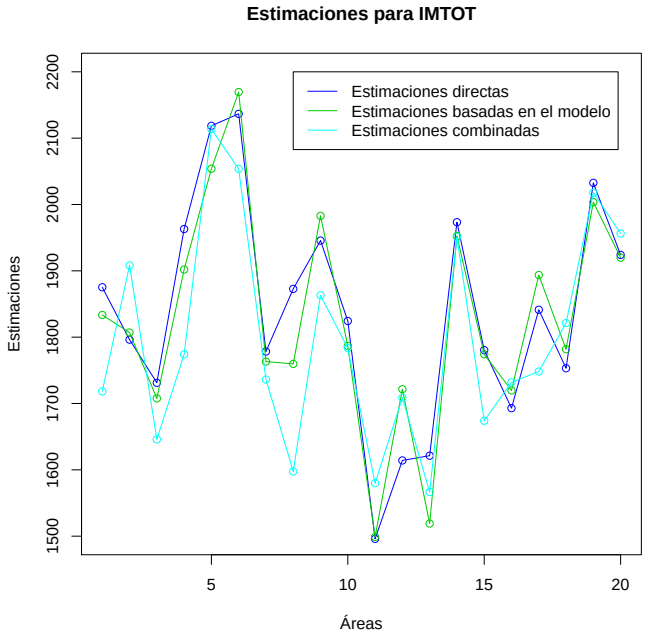

Figura 3.44: Estimaciones directas, estimaciones basadas en el modelo y estimaciones combinadas de IMTOT para el año 2013.

Para saber si estas nuevas estimaciones son mejores que las anteriores debemos calcular las varianzas de las estimaciones, necesarias para conocer los nuevos coeficientes de correlación a los que llamaremos CV_3.

Para el calculo de las varianzas de las nuevas estimaciones debemos saber, en primer lugar, si IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB son variables independientes. Pero como han sido calculadas utilizando los mismos hogares, están correlacionadas por lo que no son independientes. Como no conocemos las covarianzas de esas cuatro variables entre sí, para calcular la varianza de las estimaciones combinadas, supondremos que IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB son variables independientes y por tanto las covarianzas son igual a 0.

Por tanto, suponiendo que IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB son variables independientes, las varianzas de las estimaciones combinadas serán igual a:

$$\mathrm{Var}(\text{Estimaciones combinadas}) = \mathrm{Var}\Big(\frac{4540\,\hat{\mu}_{2d} + 1715\,\hat{\mu}_{3d} + 5149\,\hat{\mu}_{4d} + 3127\,\hat{\mu}_{5d}}{9188}\Big) =$$

$$= \frac{\mathrm{Var}(4540\,\hat{\mu}_{2d} + 1715\,\hat{\mu}_{3d} + 5149\,\hat{\mu}_{4d} + 3127\,\hat{\mu}_{5d})}{9188^2} =$$

$$= \frac{4540^2\,\mathrm{Var}(\hat{\mu}_{2d}) + 1715^2\,\mathrm{Var}(\hat{\mu}_{3d}) + 5149^2\,\mathrm{Var}(\hat{\mu}_{4d}) + 3127^2\,\mathrm{Var}(\hat{\mu}_{5d})}{9188^2} \quad \forall d = 1, ..., D$$

De modo que los nuevos coeficientes de variación vienen dados por la siguiente expresión:

$$\mathrm{CV}_3 = \frac{\sqrt{\mathrm{Var}(\text{Estimaciones combinadas})}}{\text{Estimaciones combinadas}}$$

los cuales serán comparados con los coeficientes de correlación de las estimaciones basadas en el modelo de IMTOT calculadas en la Sección 3.2.

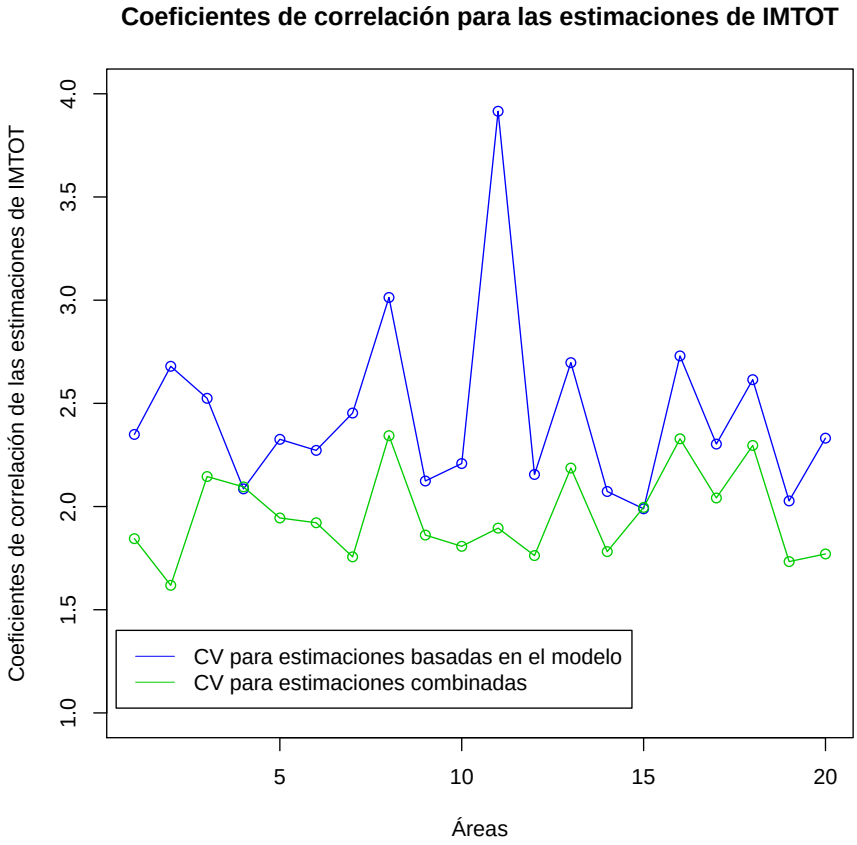

Figura 3.45: Coeficientes de correlación de las estimaciones basadas en el modelo y las combinadas para IMTOT para cada una de las veinte áreas de Galicia.

Como vemos en la Figura 3.45 los coeficientes de correlación asociados a las estimaciones combinadas son inferiores a los asociados a las estimaciones basadas en el modelo. Por lo que las nuevas estimaciones son mucho más precisas que las directas y que las basadas en el modelo. El único área donde coinciden los coeficientes de correlación son en las áreas 4 y 15. Hay que recordar que CV_3 es una aproximación ya que no estamos considerando las covarianzas.

Capítulo 4

Evolución de los cinco tipos de ingresos desde el 2007 al 2013

El objetivo de este trabajo era conocer el verdadero valor de IMTOT desde el año 2007 al 2013. El procedimiento utilizado fue tomar los datos del año 2013 para obtener un modelo que fuese capaz de obtener esos ingresos (Capítulo 3) y después aplicarlo al resto de los años. En este capítulo se calcularán las estimaciones del resto de años y se compararán a lo largo del tiempo en cada área de Galicia, considerando el período 2007 al 2013. Se calcularán las siguientes estimaciones:

- Estimaciones directas de IMTOT en el periodo 2007-2013.
- Estimaciones basadas en el modelo de IMTOT desde los años 2007 al 2013 con el modelo ajustado para el año 2013, es decir, usando las variables explicativas NPER_18A64, NPER_SUP, HOG_BAJO_UMBRAL y RENDIMIENTO.
- Estimaciones combinadas de IMTOT desde el año 2007 al 2013 utilizando las estimaciones basadas en el modelo de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB de dichos años. Para conocer las estimaciones basadas en el modelo de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB para el periodo 2007-2013 se utilizarán los modelos ajustados para el año 2013 calculados en el capítulo 3 y después se aplicarán al resto de los años.

  Es decir, las variables explicativas utilizadas para los cuatro últimos ingresos para dicho periodo de años son:

  - IM_AJENA: NPER_18A64, NPER_SUP.
  - IM_PROPIA: NPER_PRIM, NPER_SUP.
  - IM_CONTRIB: CTE, PENSIONES.
  - IM_NO_CONTRIB: HOG_BAJO_UMBRAL, NPER_ESP.

En primer lugar mostraremos las estimaciones directas, las basadas en el modelo y las sintéticas de IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB. Por último, las estimaciones directas, las basadas en el modelo, las sintéticas y las combinadas de IMTOT.

Las **estimaciones sintéticas** nos permitirán obtener estimaciones más suavizadas y vienen dadas por la siguiente expresión:

$$y_{\text{sint}} = X\hat{\beta}.$$

Donde X es la matriz de las covariables y $\hat{\beta}$ las estimaciones de β habiendo estimado los modelos con efecto aleatorio. Como veremos, con las estimaciones sintéticas, obtendremos unas estimaciones más suavizadas a lo largo de los años.

Comentar que cuando nos referimos a las siete principales áreas gallegas son: A Coruña, Santiago, Ferrol, Lugo, Ourense, Pontevedra y Vigo.

IM_AJENA

En esta sección calcularemos las estimaciones directas, las basadas en el modelo y las sintéticas de IM_AJENA considerando el periodo 2007 al 2013. Esta es la primera variable necesaria para conocer las estimaciones combinadas de IMTOT y que, como vimos en el Capítulo 3, son más precisas que las estimaciones basadas en el modelo para IMTOT.

Las estimaciones directas han sido calculadas directamente de los resultados obtenidos en la ECV para cada año para IM_AJENA. Las estimaciones basadas en el modelo se han calculando tomando como modelo el ajustado para el año 2013 y como variable respuesta las estimaciones directas relativas a cada año. Es decir, para calcular las estimaciones basadas en el modelo de IM_AJENA para los años 2007 al 2013 utilizaremos como variables explicativas NPER_18A64 y NPER_SUP. Las estimaciones basadas en el modelo de IM_AJENA será utilizadas para calcular las estimaciones combinadas de IMTOT.

A continuación mostraremos las tres estimaciones nombradas mediante gráficos. Haremos cuatro figuras (4.1, 4.2, 4.3 y 4.4), una para cada provincia. Dentro de cada provincia se harán los gráficos de las áreas correspondientes a esa provincia; esto nos permitirá comparar los resultados dentro de cada provincia, y después entre provincias.

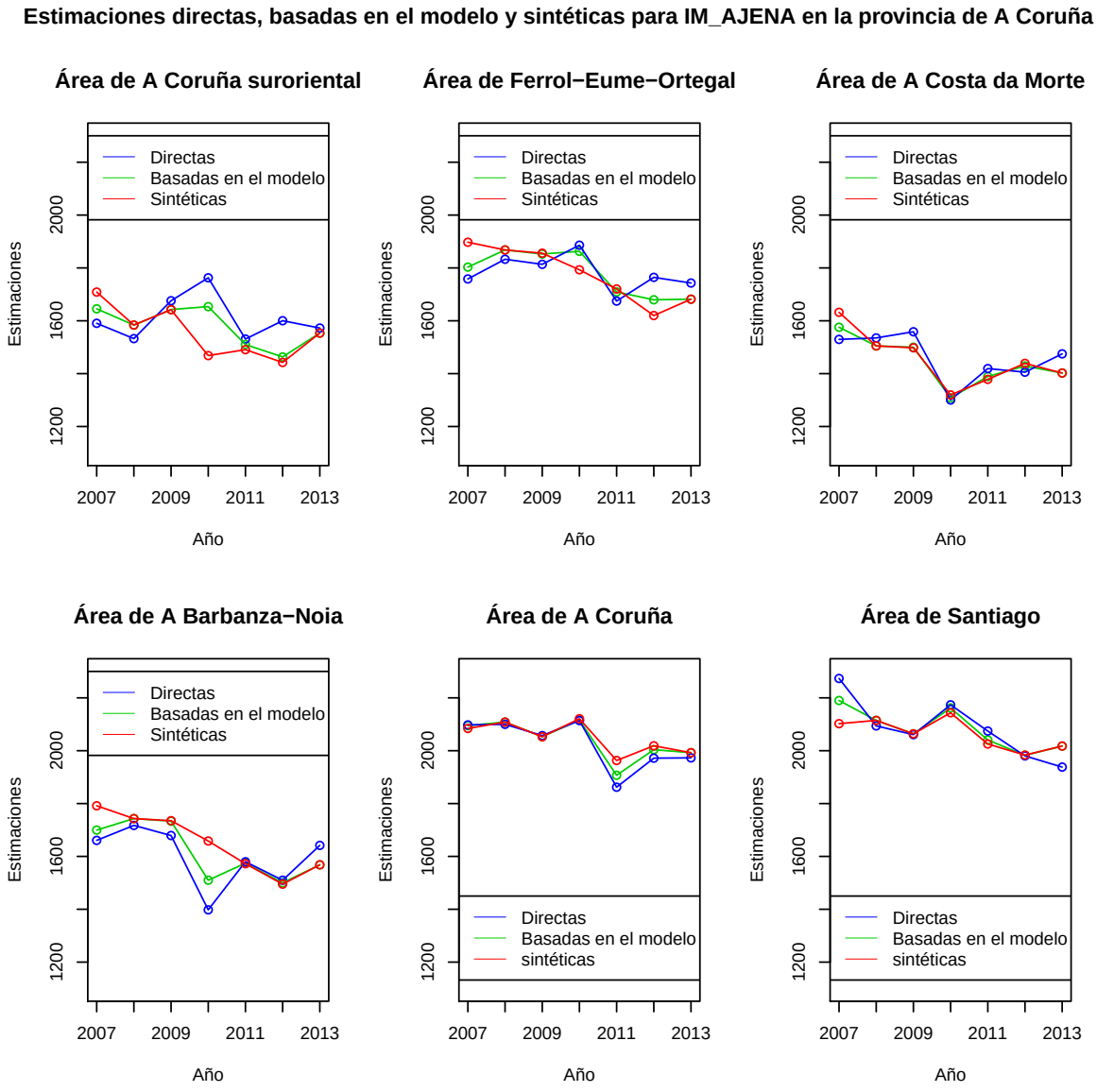

Figura 4.1: Estimaciones directas, basadas en el modelo y sintéticas para IM_AJENA en la provincia de A Coruña.

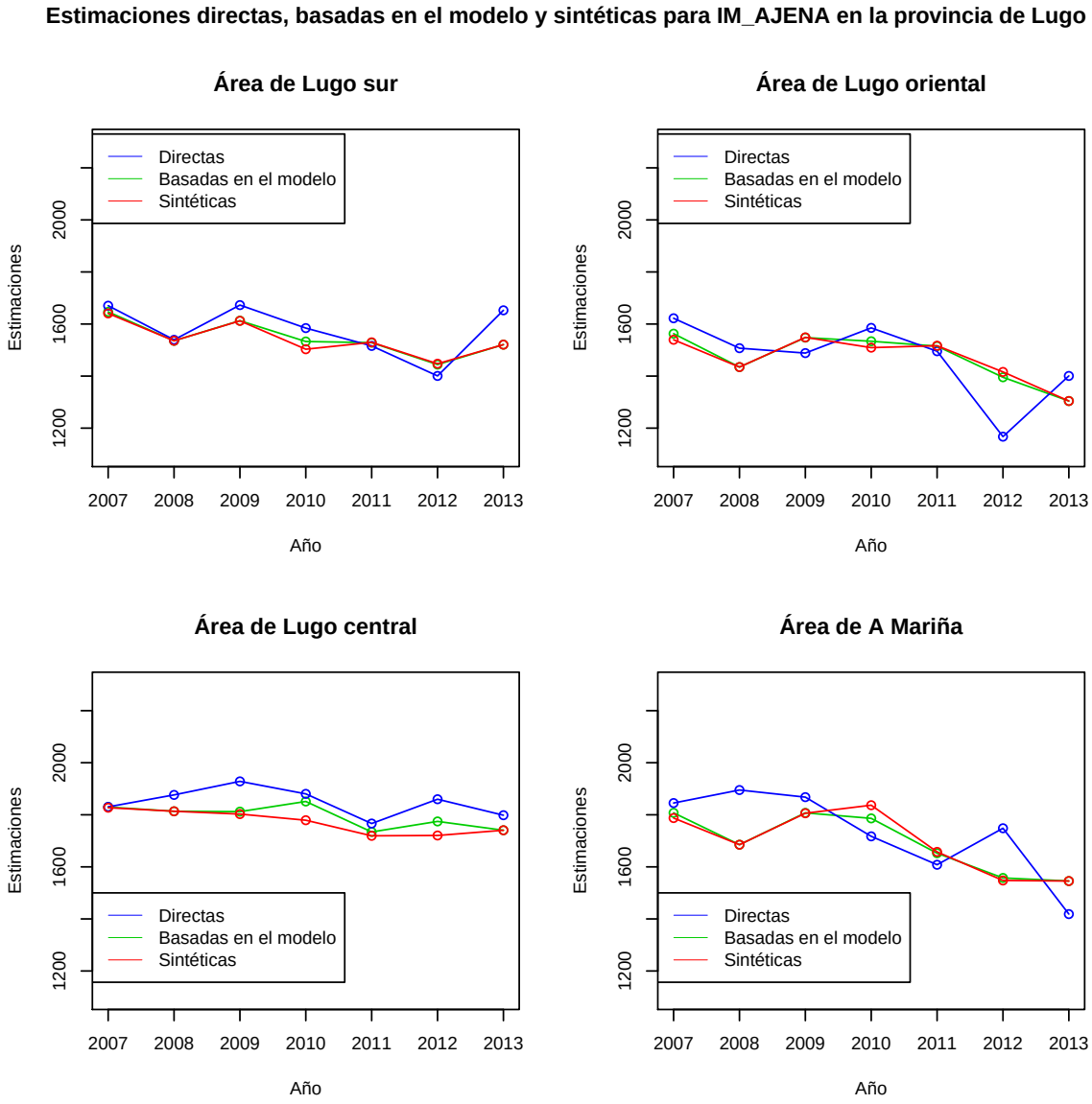

Figura 4.2: Estimaciones directas, basadas en el modelo y sintéticas para IM_AJENA en la provincia de Lugo.

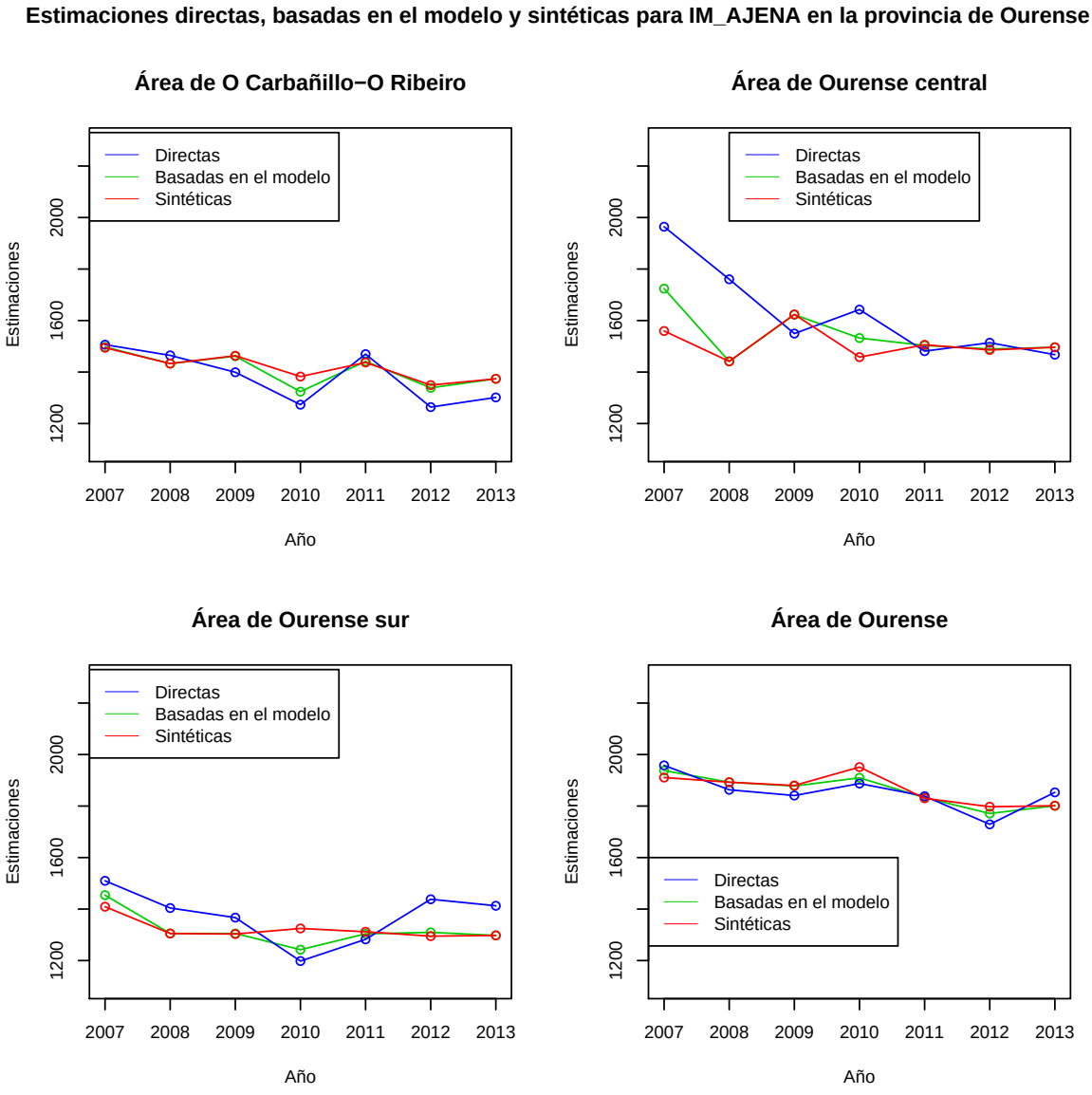

Figura 4.3: Estimaciones directas, basadas en el modelo y sintéticas para IM_AJENA en la provincia de Ourense.

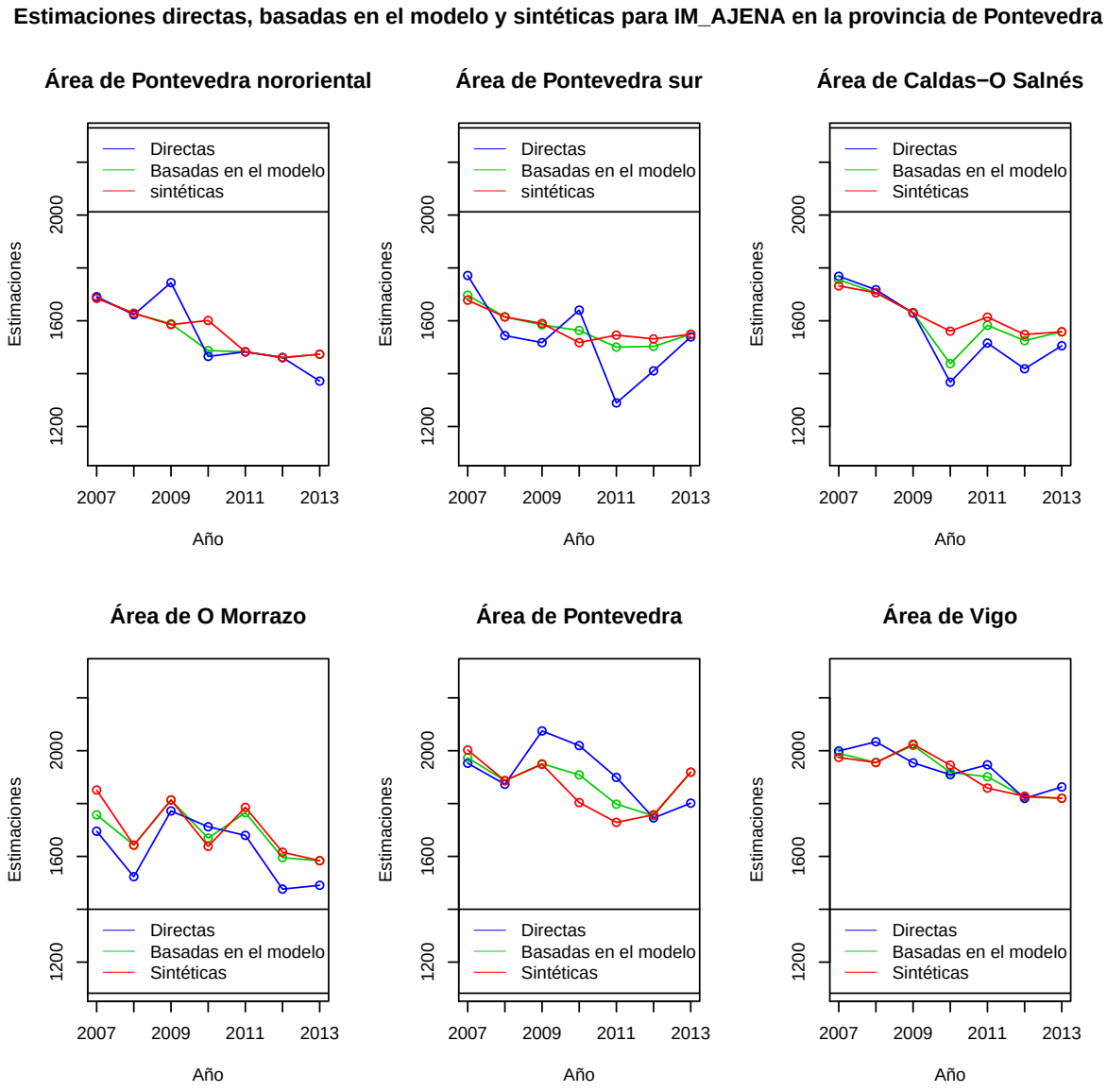

Figura 4.4: Estimaciones directas, basadas en el modelo y sintéticas para IM_AJENA en la provincia de Pontevedra.

Para IM_AJENA los resultados obtenidos de las estimaciones basadas en el modelo nos indican que para las provincias de A Coruña, Lugo y Ourense en el año 2013 los ingresos se han reducido aproximadamente 100 unidades respecto al año 2007. En cambio en la provincia de Pontevedra se ha producido una reducción de 200 unidades. En general desde el año 2007 al 2013 los ingresos procedentes de ingresos por cuenta ajena, IM_AJENA, se han reducido.

En la provincia de A Coruña las áreas con más ingresos son A Coruña y Santiago, con ingresos superiores a 2000 unidades, y el área de Ferrol con ingresos de aproximadamente 1800 unidades. Por otro lado, en la provincia de Lugo las áreas con más ingresos son Lugo central y A Mariña con unos ingresos cercanos a los 1800 unidades. En la provincia de Ourense los ingresos más altos los tiene el área de Ourense con unos ingresos similares a los del área de Ferrol. Por último, en la provincia de Ponvedra las áreas con mayores ingresos son sus dos áreas principalente, Vigo y Pontevedra con ingresos entre 1800 y 2000 unidades según el año.

En general, las áreas con mayores ingresos mensuales procedentes de ingresos por cuenta ajena en el hogar, IM_AJENA, son a Coruña, Santiago de Compostela, Ourense, Pontevedra y Vigo; cinco de las siete principales ciudades grandes.

Comentar que las estimaciones sintéticas no son muy suaves comparadas con las basadas en el modelo, esto es debido a que para IM_AJENA no habíamos introducido mucho efecto aleatorio al modelo.

IM_PROPIA

Al igual que hicimos para IM_AJENA, compararemos las estimaciones obtenidas para IM_PROPIA para cada área a lo largo de los años del 2007 al 2013.

Calcularemos las estimaciones directas , las basadas en el modelo utilizando el modelo ajustado para el año 2013 y aplicándolo al resto de años y, por último, las estimaciones sintéticas. Para las estimaciones basadas en el modelo utilizaremos como variables explicativas NPER_PRIM y NPER_SUP

En las figuras 4.5, 4.6, 4.7 y 4.8 representaremos las tres estimaciones para cada provincia, dividiendo cada provincia en áreas.

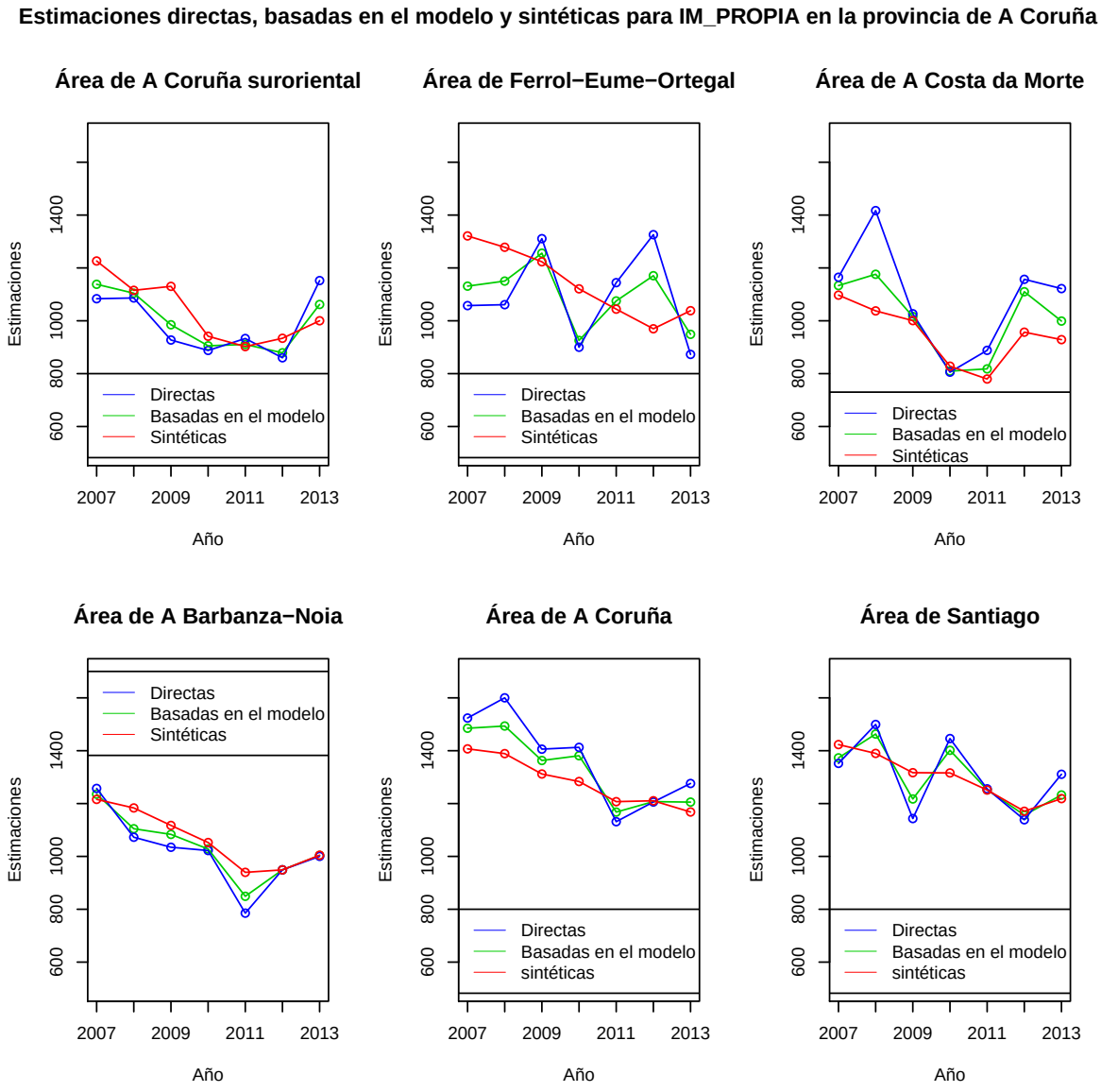

Figura 4.5: Estimaciones directas, basadas en el modelo y sintéticas para IM_PROPIA en la provincia de A Coruña.

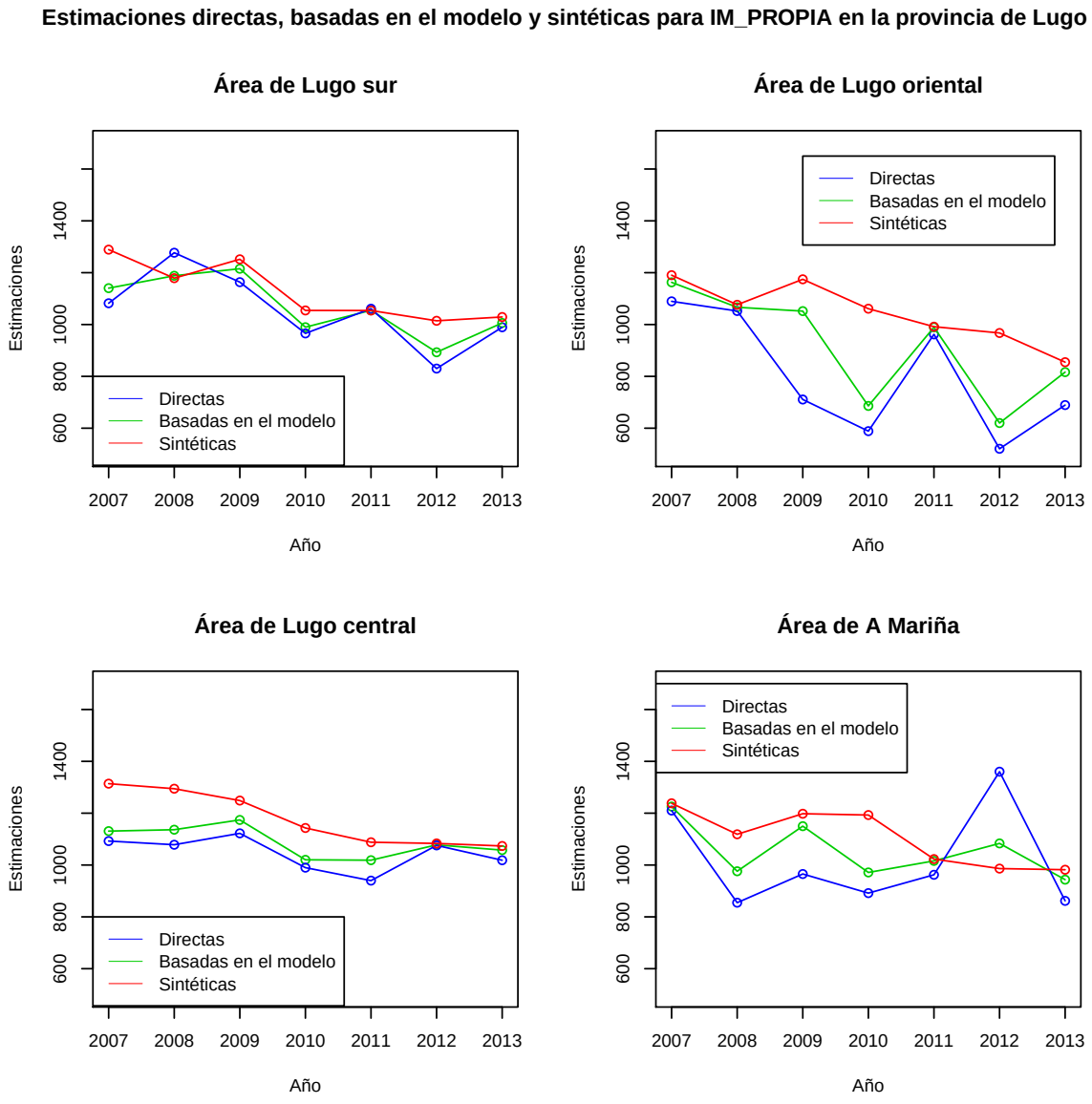

Figura 4.6: Estimaciones directas, basadas en el modelo y sintéticas para IM_PROPIA en la provincia de Lugo.

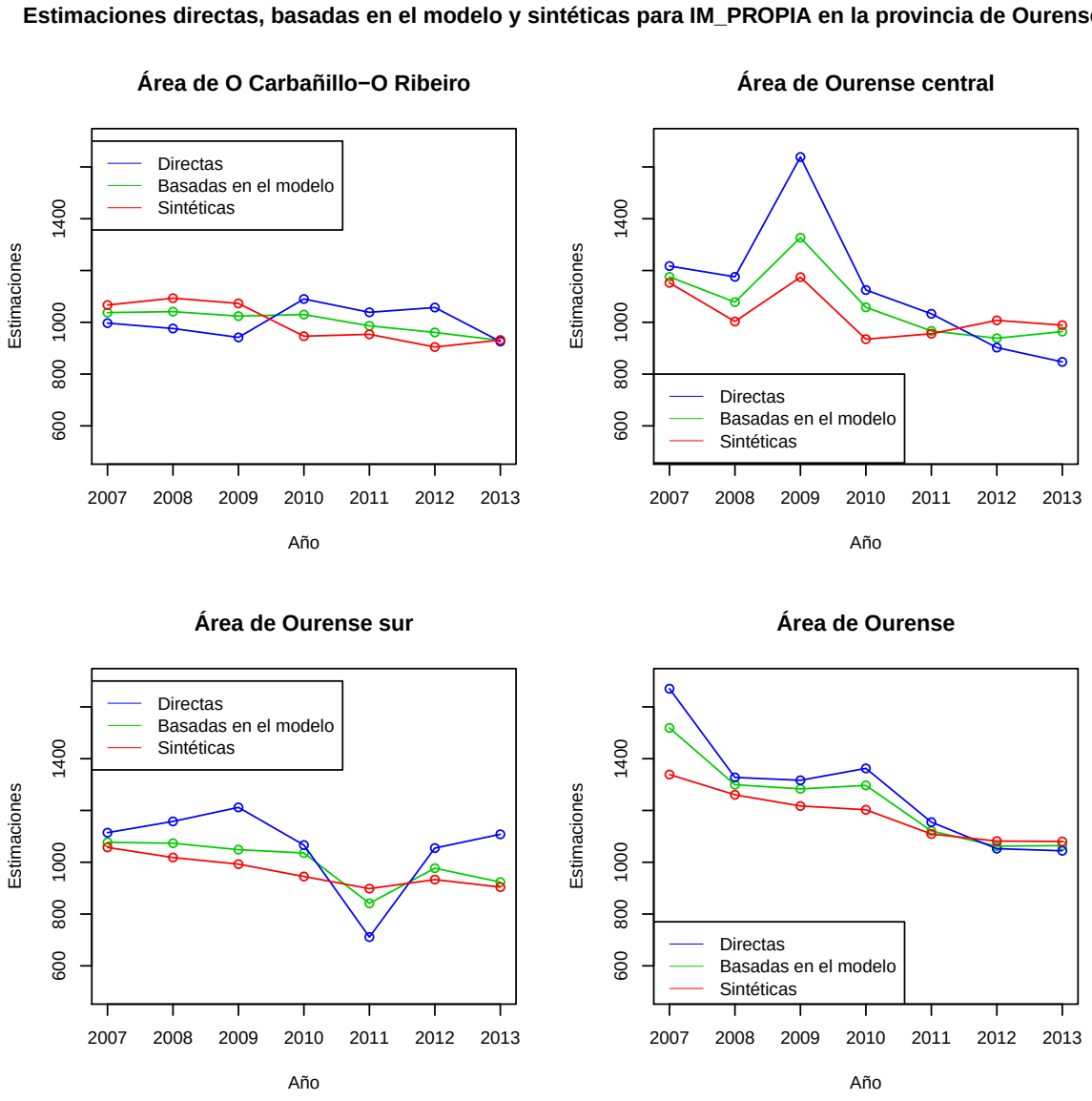

Figura 4.7: Estimaciones directas, basadas en el modelo y sintéticas para IM_PROPIA en la provincia de Ourense.

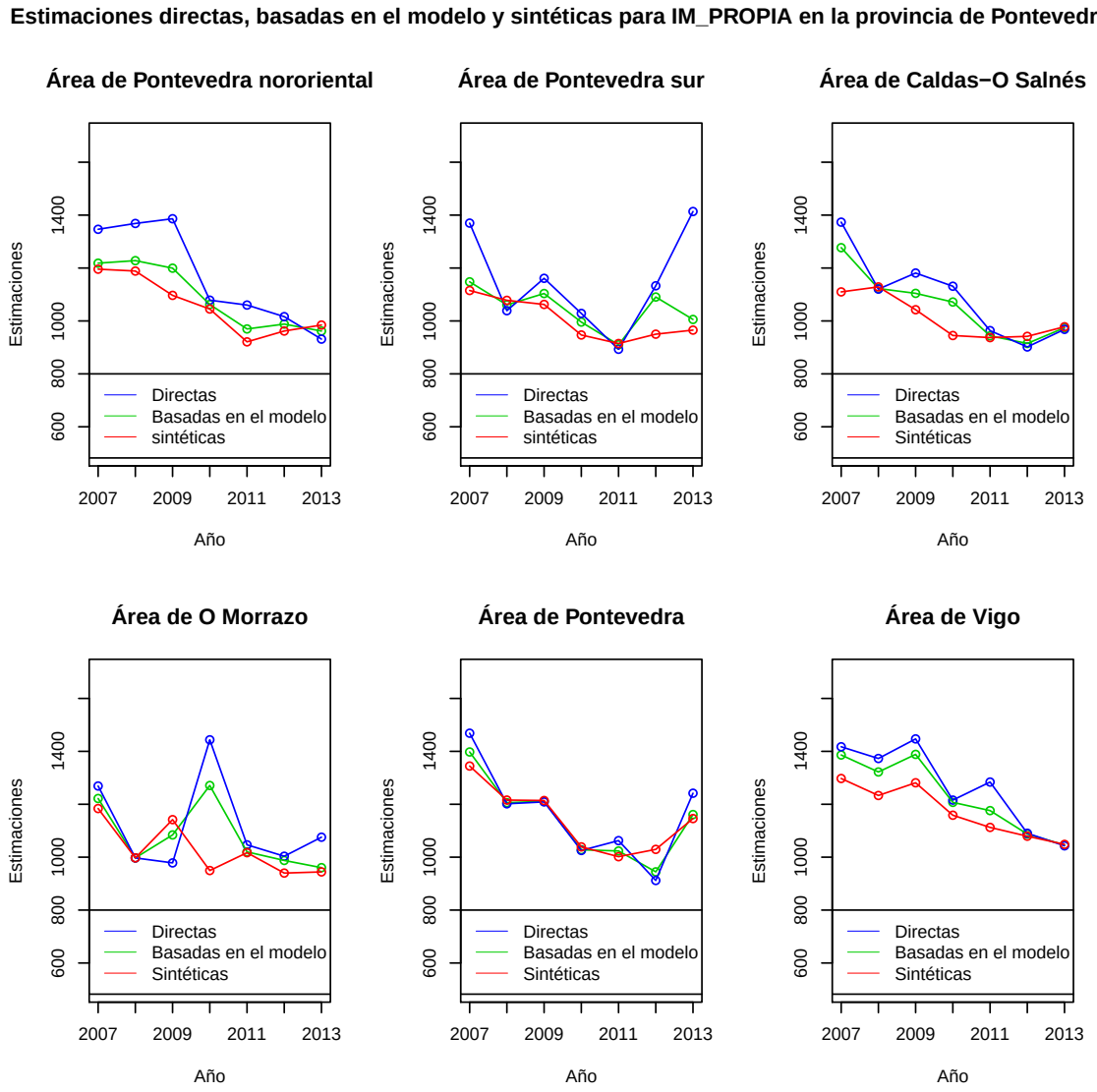

Figura 4.8: Estimaciones directas, basadas en el modelo y sintéticas para IM_PROPIA en la provincia de Pontevedra.

Al igual que ocurría para IM_AJENA, en IM_PROPIA las estimaciones obtenidas nos indican que los ingresos se han reducido desde el año 2007 al 2013. También observamos que, en general, las estimaciones varían bastante de un año a otro para cada área; principalmente en las áreas de O Morrazo y Ferrol en el año 2010. En el IGE suelen tener problemas con el área de O Morrazo. Las áreas con IM_PROPIA más alto son A Coruña, Santiago, Ourense y Vigo.

Comentar además, que para este tipo de ingreso las estimaciones sintéticas suavizan mucho a las estimaciones basadas en el modelo.

IM_CONTRIB

Calculamos las estimaciones directas, las sintéticas y las basadas en el modelo para IM_CONTRIB para los años 2007 al 2013. Para las basadas en el modelo utilizaremos como variables explicativas la CTE y PENSIONES. Las estimaciones serán representadas en las Figuras 4.9, 4.10, 4.11 y 4.12 para las provincias de A Coruña, Lugo, Ourense y Pontevedra, respectivamente.

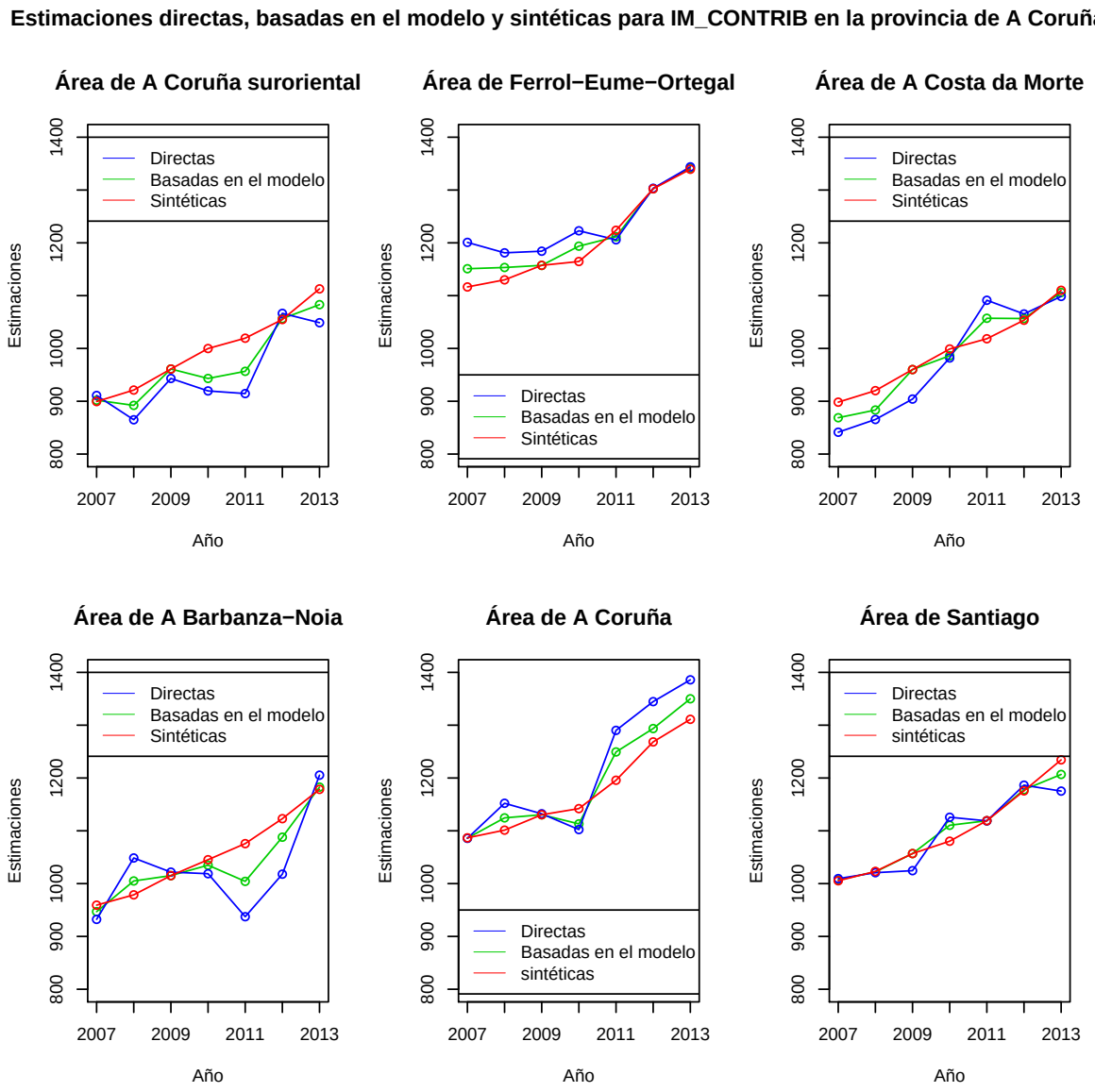

Figura 4.9: Estimaciones directas, basadas en el modelo y sintéticas para IM_CONTRIB en la provincia de A Coruña.

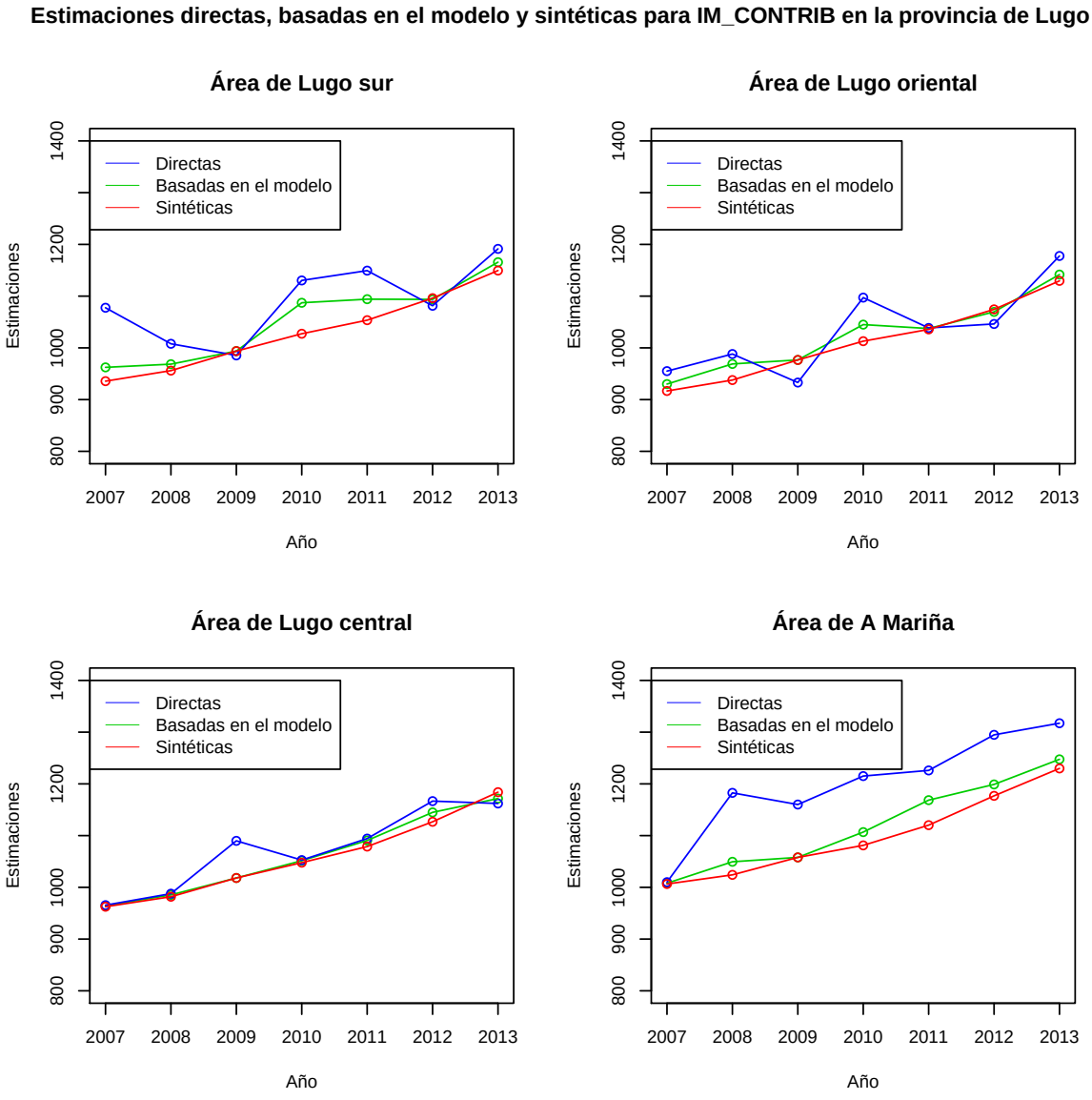

Figura 4.10: Estimaciones directas, basadas en el modelo y sintéticas para IM_CONTRIB en la provincia de Lugo.

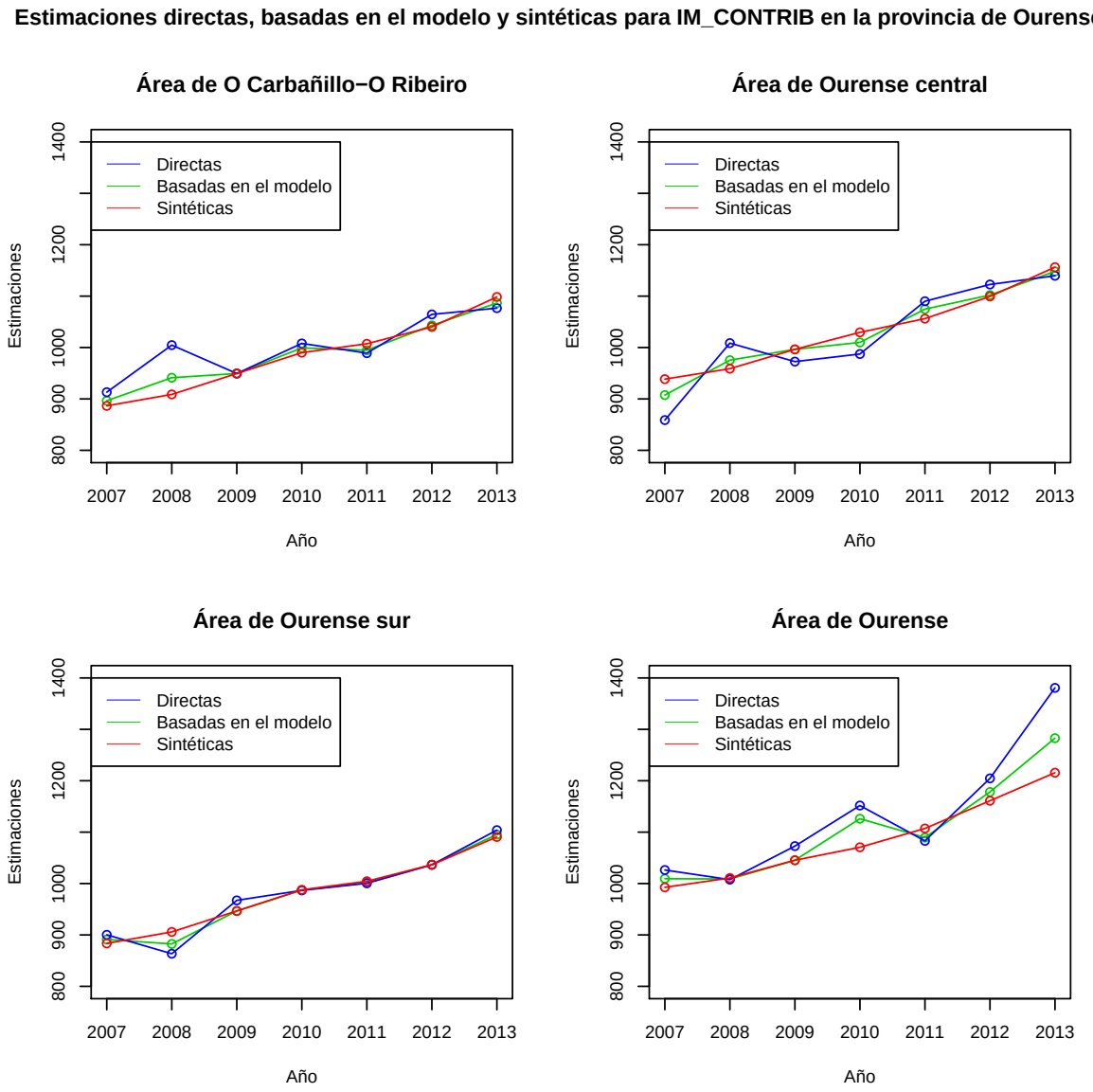

Figura 4.11: Estimaciones directas, basadas en el modelo y sintéticas para IM_CONTRIB en la provincia de Ourense.

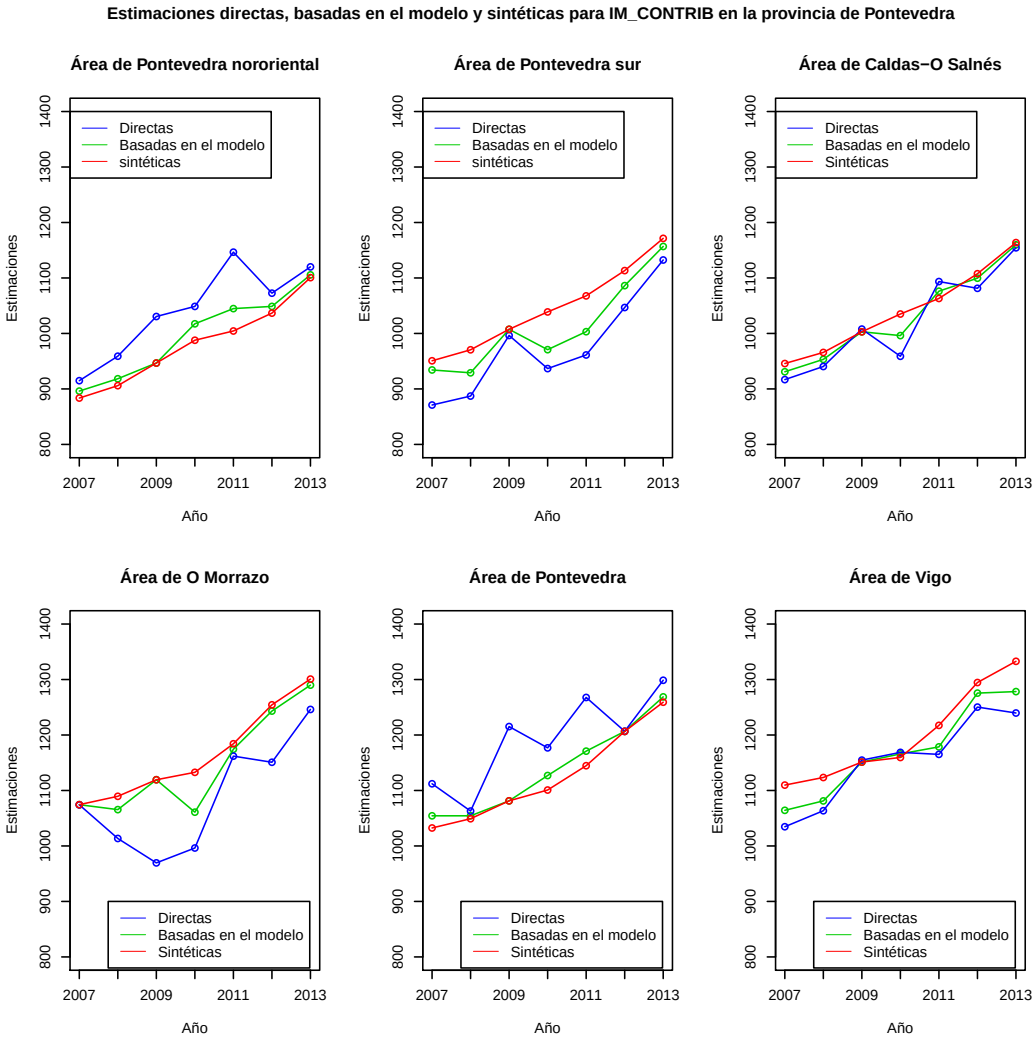

Figura 4.12: Estimaciones directas, basadas en el modelo y sintéticas para IM_CONTRIB en la provincia de Pontevedra.

Para los ingresos mensuales en el hogar procedentes de prestaciones contributivas, IM_CONTRIB, destaca la tendencia creciente en todas las áreas a medida que aumenta los años. En esta variable teníamos también cierta variabilidad del efecto aleatorio ($\sigma_u^2 = 1787$) por lo que como era lógico las estimaciones sintéticas suavizarán a las estimaciones directas y a las basadas en el modelo.

Las áreas con mayores ingresos de este tipo son las siete principales áreas gallegas. El área de Ferrol sale muy beneficiada en este tipo de ingresos ya que hay pensiones muy altas de la gente jubilada de Endesa y Navantia. En la Figura 3.35 vimos que para el año 2013 este área es el que tiene más ingresos de este tipo, después de Coruña.

IM_NO_CONTRIB

En las Figuras 4.13, 4.14, 4.15 y 4.16 veremos las estimaciones directas, las basadas en el modelo y las sintéticas desde el año 2007 al 2013 para las veinte áreas en las que se divide la Comunidad de Galicia para IM_NO_CONTRIB. Las cuatro figuras corresponden a las cuatro provincias gallegas, que a su vez están divididas en áreas. Para obtener las estimaciones basadas en el modelo de dicho periodo se han utilizado como variables explicativas HOG_BAJO_UMBRAL y NPER_ESP.

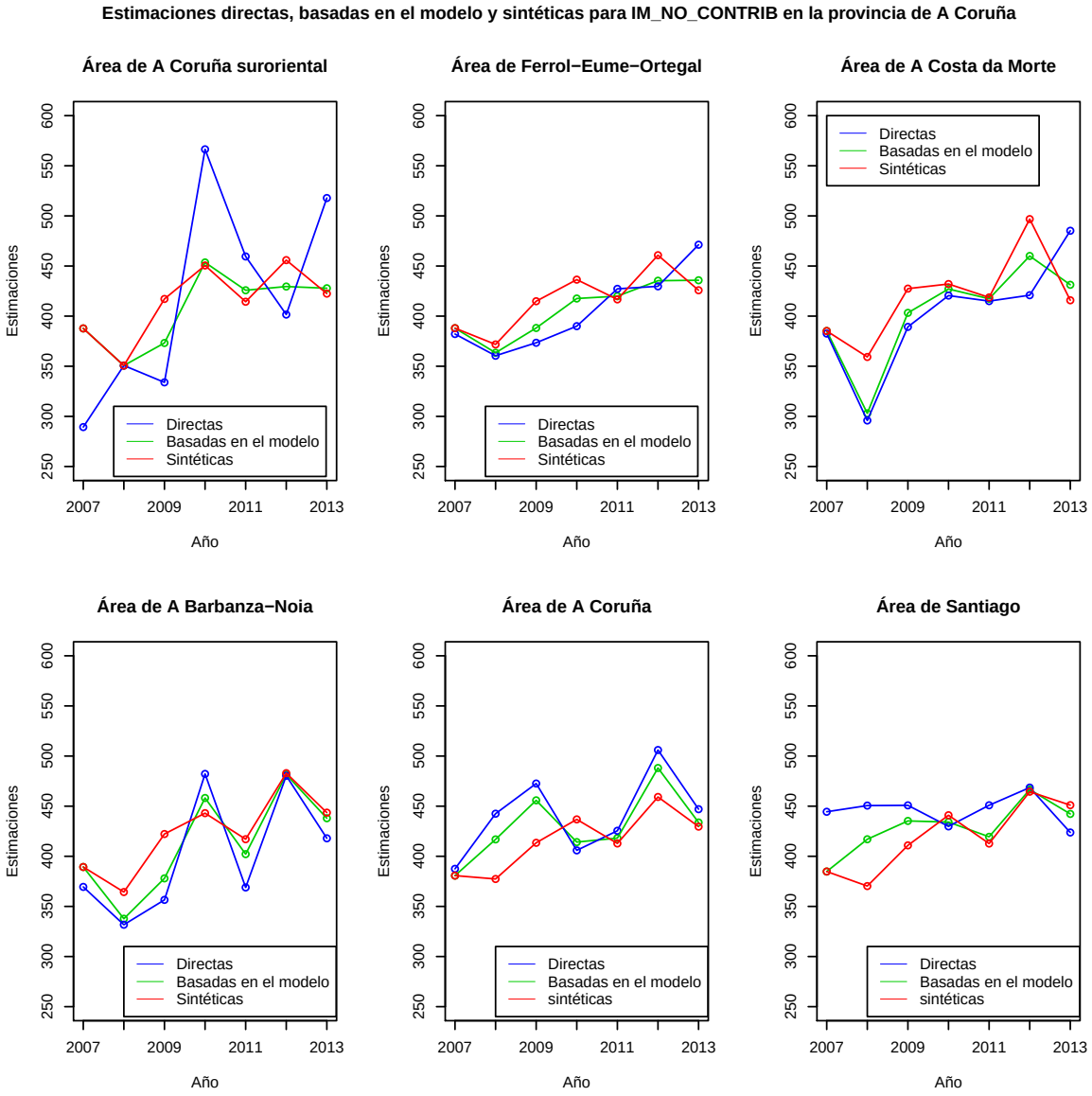

Figura 4.13: Estimaciones directas, basadas en el modelo y sintéticas para IM_NO_CONTRIB en la provincia de A Coruña.

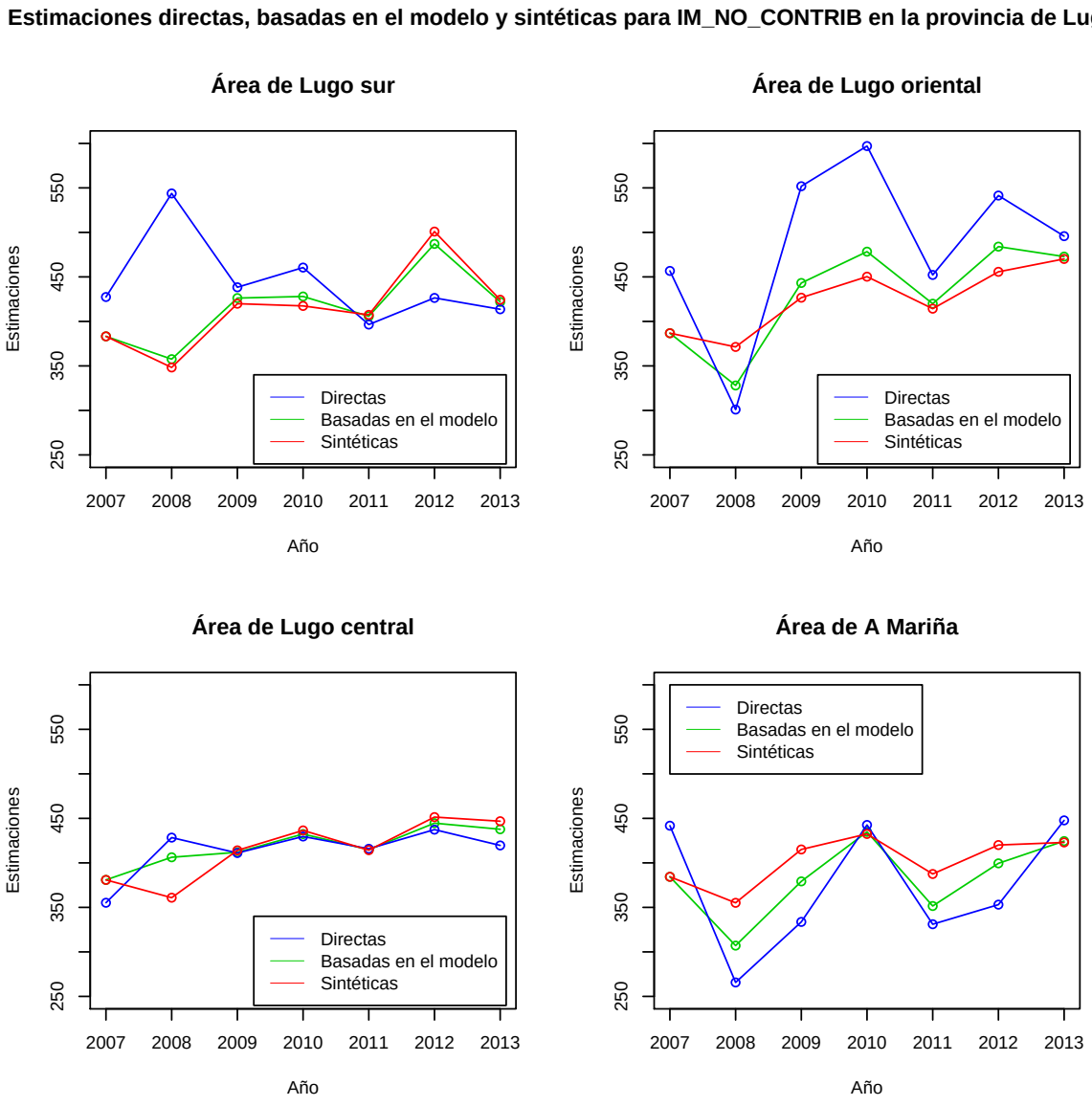

Figura 4.14: Estimaciones directas, basadas en el modelo y sintéticas para IM_NO_CONTRIB en la provincia de Lugo.

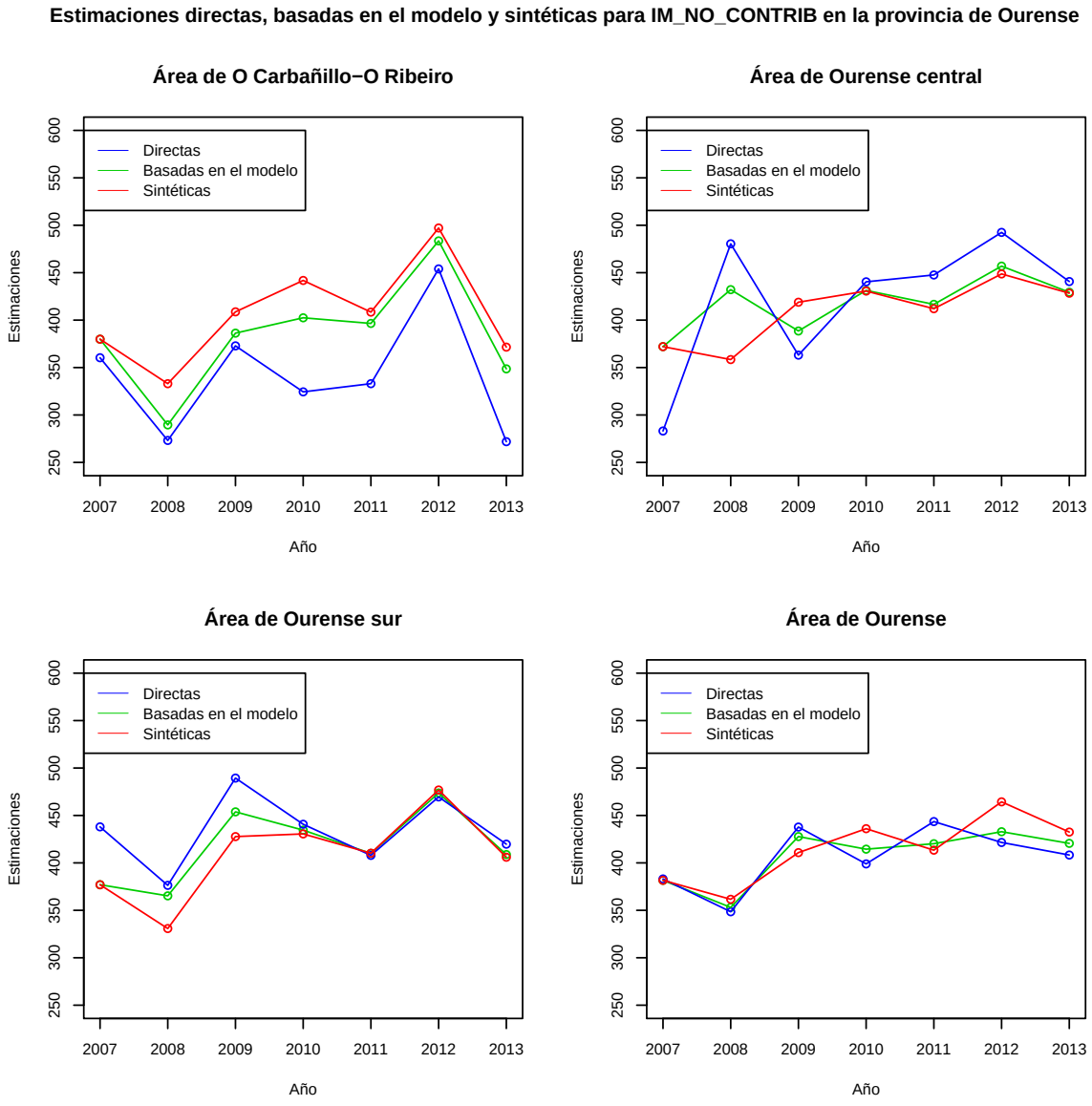

Figura 4.15: Estimaciones directas, basadas en el modelo y sintéticas para IM_NO_CONTRIB en la provincia de Ourense.

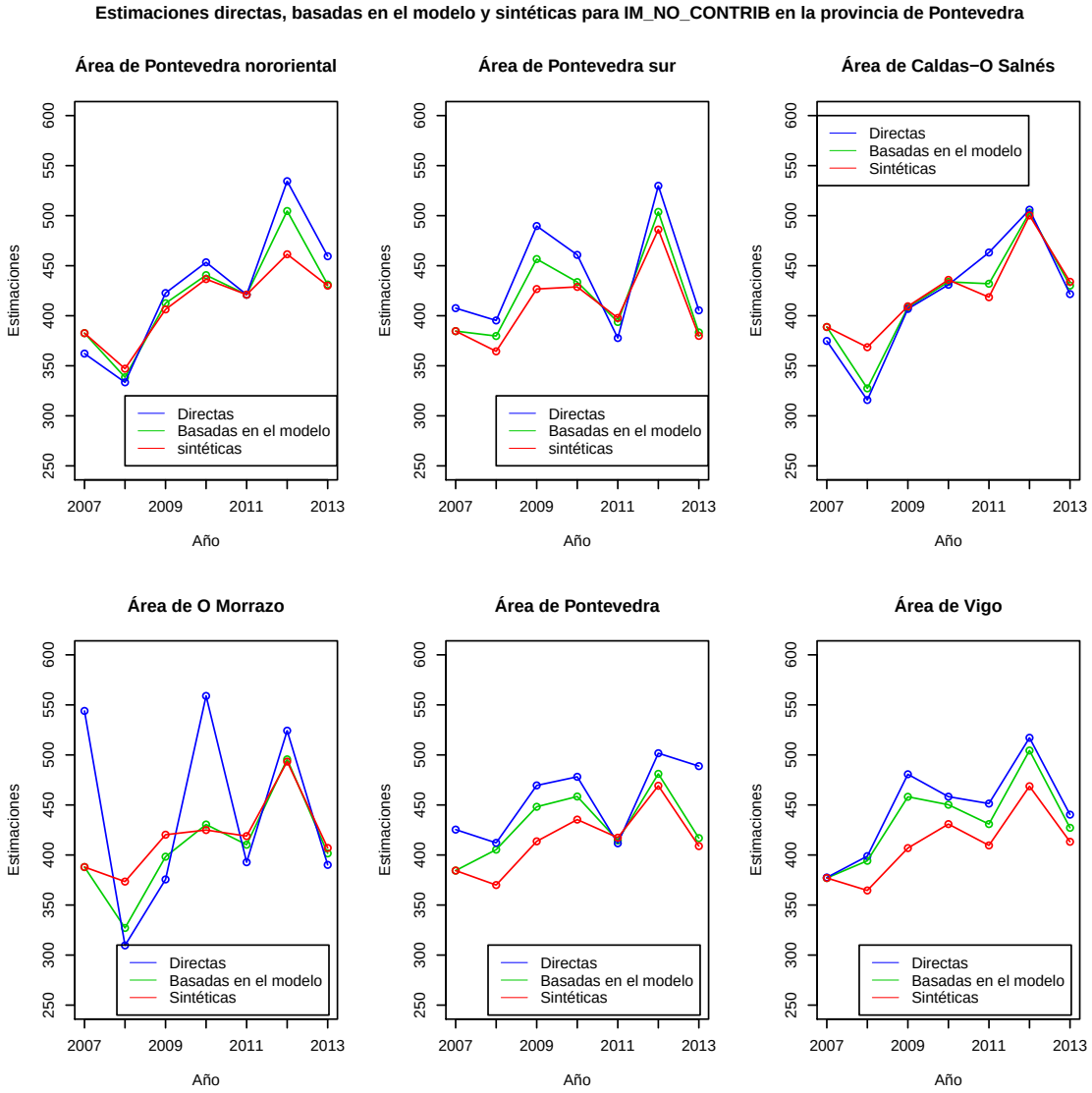

Figura 4.16: Estimaciones directas, basadas en el modelo y sintéticas para IM_NO_CONTRIB en la provincia de Pontevedra.

Los valores de las estimaciones de IM_NO_CONTRIB son muy similares para todas las áreas, con una media de 413,12 unidades. Las estimaciones directas son mucho más variables de un año a otro que las sintéticas y las basadas en el modelo. Las estimaciones más suaves respecto a las otras son las sintéticas.

Por ejemplo, para el año 2013 (Figura 3.43) las áreas con mayores ingresos de este tipo son A Barbanza-Noia, Santiago, Ferrol, Lugo central y Lugo oriental. Vemos que existen áreas de las no principales que tienen ingresos altos, puede ser por la existencia de personas muy mayores que no tienen pensiones contributivas.

IMTOT

Por último representaremos las estimaciones directas, las basadas en el modelo, las sintéticas y las combinadas para IMTOT (Figuras 4.17, 4.18, 4.19 y 4.20) para las veinte áreas gallegas. Cada provincia estará representada en una figura diferente.

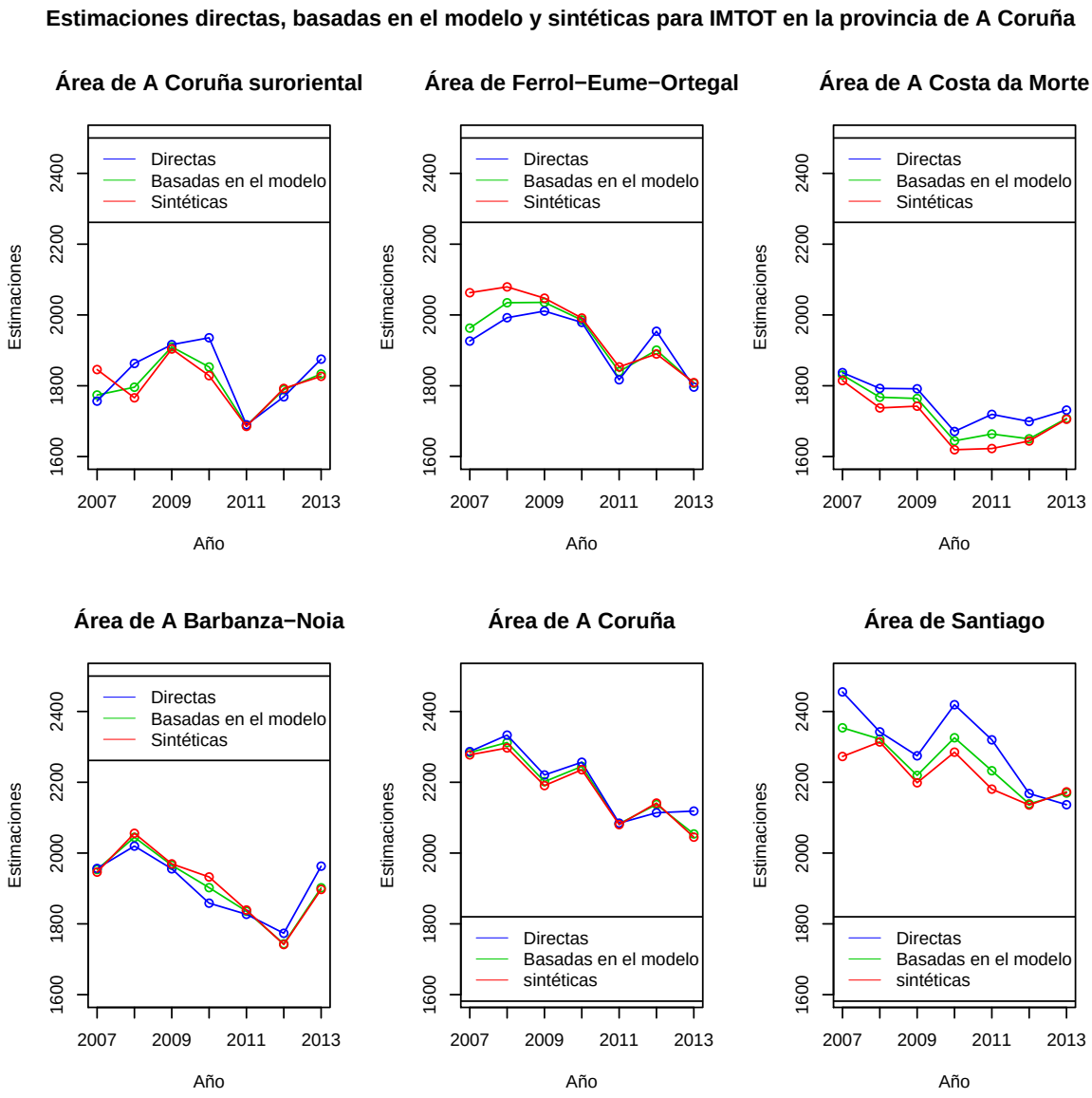

Figura 4.17: Estimaciones directas, basadas en el modelo y sintéticas para IMTOT en la provincia de A Coruña.

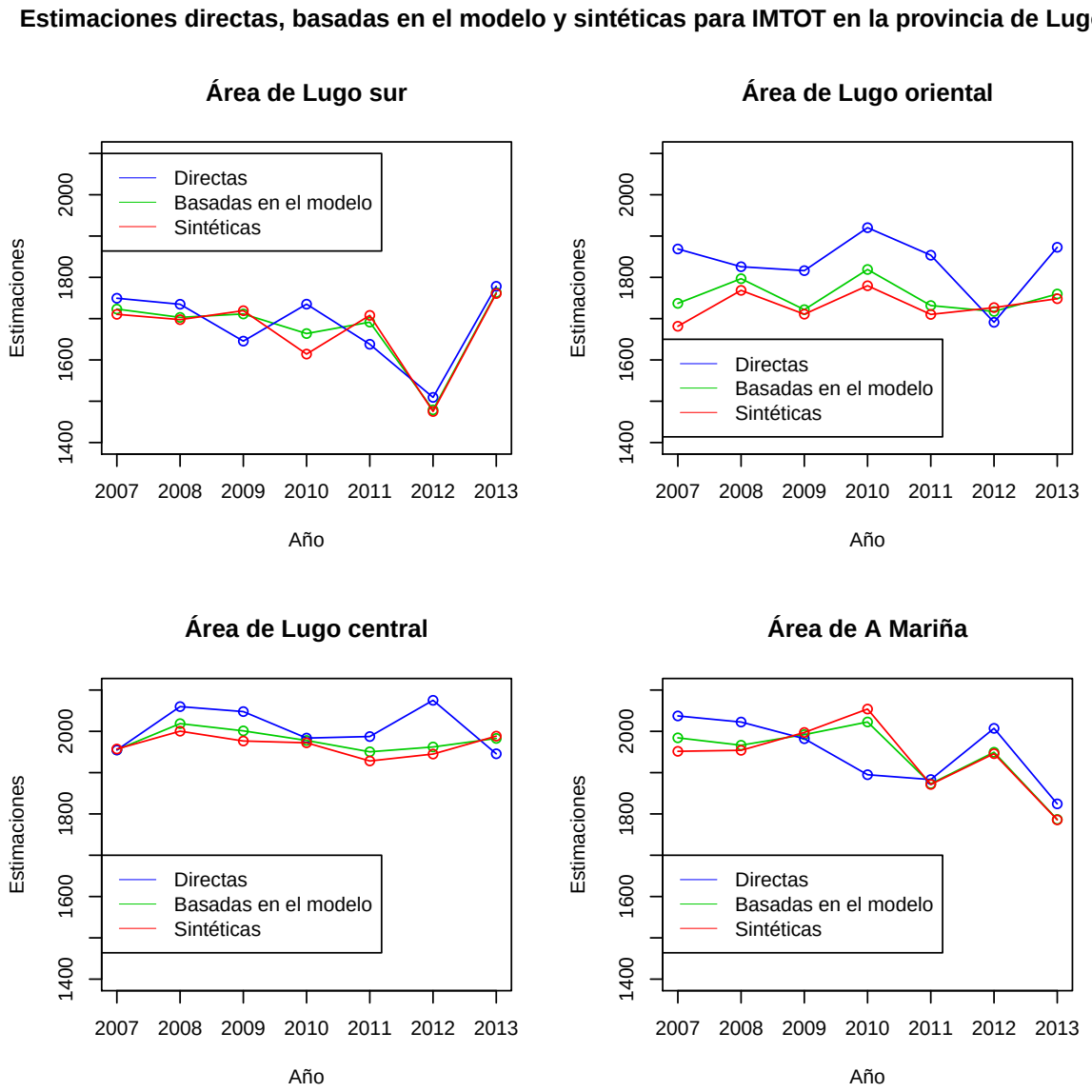

Figura 4.18: Estimaciones directas, basadas en el modelo y sintéticas para IMTOT en la provincia de Lugo.

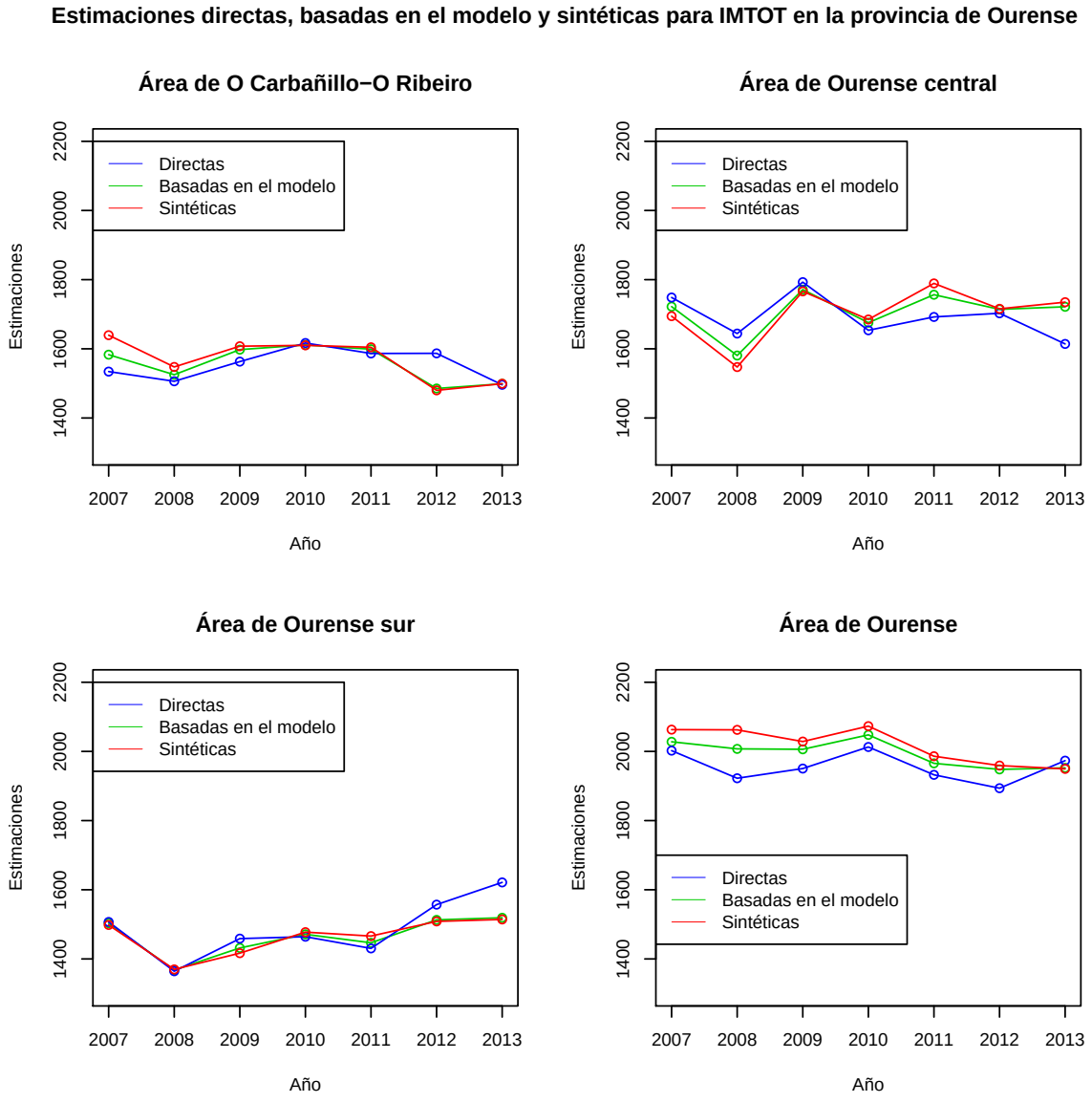

Figura 4.19: Estimaciones directas, basadas en el modelo y sintéticas para IMTOT en la provincia de Ourense.

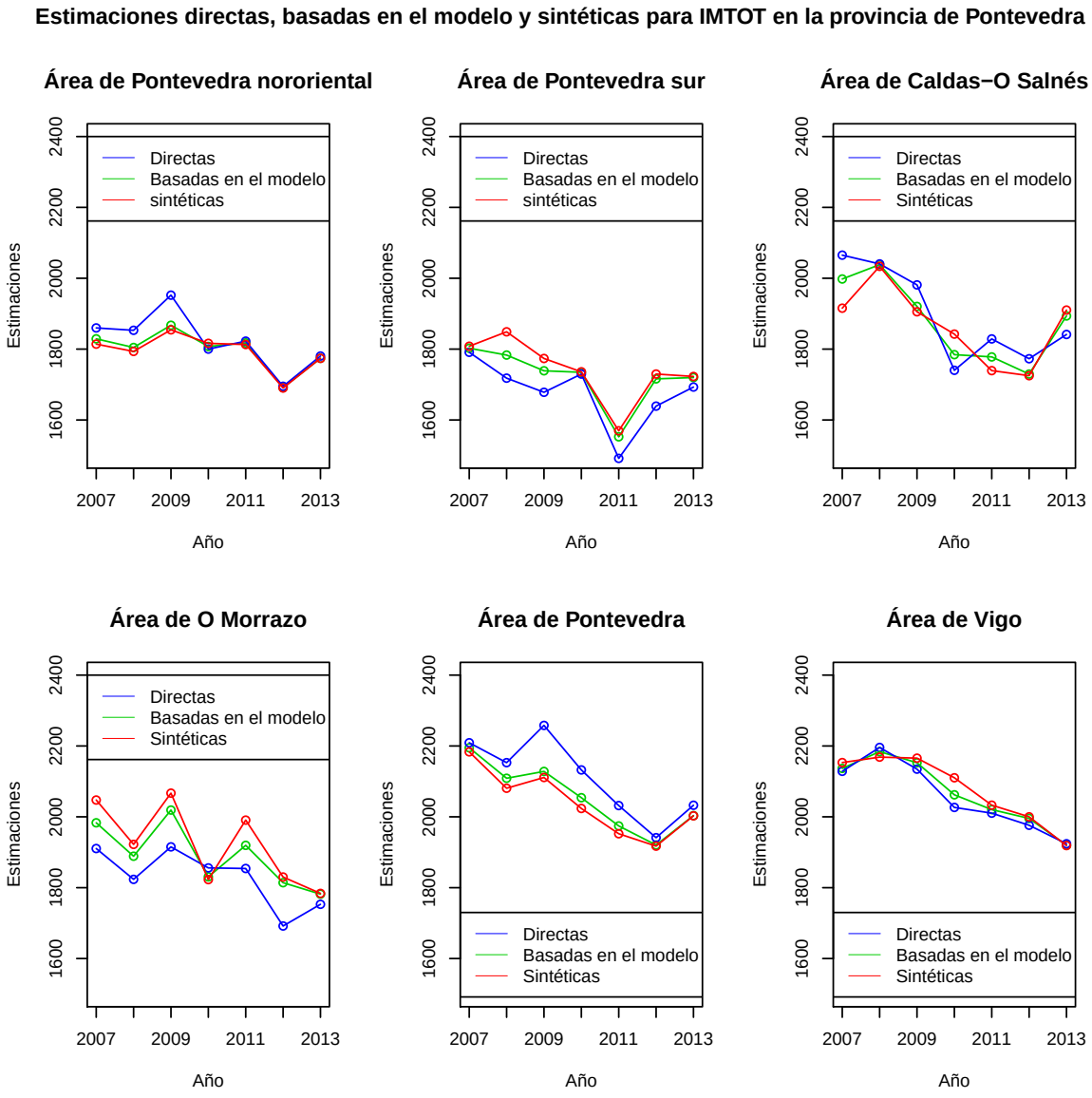

Figura 4.20: Estimaciones directas, basadas en el modelo y sintéticas para IMTOT en la provincia de Pontevedra.

Para las provincias de A Coruña y Pontevedra vemos una tendencia decreciente de IMTOT en todas las áreas a medida que aumentan los años. En cambio para las provincias de Lugo y Ourense los ingresos son más o menos estables de un año para otro; excepto en Lugo sur en el año 2012.

Las áreas con mayor ingreso total, IMTOT, son cuatro de las siete áreas principales gallegas: A Coruña, Santiago, Pontevedra y Vigo. El área con IMTOT más bajo es en Ourense sur. El área donde más varían los ingresos es en O Morrazo.

Como resumen de todas las variables, si tenemos poco efecto aleatorio o no tenemos (IMTOT e IM_AJENA) las estimaciones sintéticas no son tan suaves si las comparamos con las estimaciones directas y las basadas en el modelo. En cambio, si tenemos bastante efecto aleatorio (IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB) las estimaciones sintéticas suavizan mucho a las otras dos estimaciones.

Capítulo 5

Conclusiones

En este trabajo se ha tratado de estimar el ingreso medio mensual total en el hogar, IMTOT, para cada una de las veinte áreas en las que se divide Galicia. Para ello se han utilizado métodos de estimación en áreas pequeñas (SAE), en concreto los modelos de Fay-Herriot. En áreas pequeñas, el método utilizado para seleccionar aquellas variables que finalmente se introducen en el modelo de Fay-Herriot para IMTOT fue el criterio del AIC para modelos mixtos, en particular el AIC condicional. El AIC condicional se utiliza cuando estamos interesados en estudiar cada área en particular y no la población en general, como es nuestro caso.

Para la estimación de IMTOT se compararon los estimadores de IMTOT basados en el modelo y basados en el diseño. Por otro lado, se calcularon las estimaciones obtenidas al combinar las estimaciones basadas en el modelo de los cuatro ingresos en que se divide IMTOT: IM_AJENA, IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB. Estas nuevas estimaciones, denotadas como estimaciones combinadas, fueron mucho más precisas que las estimaciones directas e incluso que las basadas en el modelo.

A efectos metodológicos se vio que para modelos mixtos en áreas pequeñas existen dos AIC, un AIC marginal (mAIC) utilizado cuando es de interés la población en general, y un AIC condicional (cAIC) utilizado cuando se tiene como objetivo analizar cada área en particular. El mAIC utiliza la verosimilitud marginal y el cAIC la verosimilitud condicional; ambos AIC tienen el mismo parámetro de penalización. En este trabajo se tiene como objetivo estudiar cada área en particular por lo que se hizo uso del cAIC.

Fueron muchos los autores que trataron de estudiar dicho cAIC; en particular *Vaida y Blanchard* (2005) y *Han* (2013). El cAIC ofrecido por *Vaida y Blanchard* (2005) es utilizado cuando conocemos la variabilidad del efecto aleatorio, algo bastante improbable en la práctica; en cambio el cAIC de *Han* (2013) se utiliza cuando no conocemos dicha variabilidad y estimamos los parámetros del modelo por REML o ML. Para elegir el mejor modelo, en este trabajo se ha utilizado el cAIC de *Han* (2013) ya que no se

conocía la variabilidad del efecto aleatorio.

El AIC que utiliza la librería *sae* para modelos mixtos no es correcto ya que solo considera el caso de estudiar la población en general, utiliza la verosimilitud marginal, además utiliza como parámetro de penalización $p + 1$ sin tener en cuenta la estimación de σ_u^2. En general, este AIC tiende a elegir el modelo con menor variabilidad del efecto aleatorio.

En cuanto a la metodología, otro punto a destacar es que si tenemos mucha variabilidad del efecto aleatorio en el modelo (IM_PROPIA, IM_CONTRIB e IM_NO_CONTRIB), las estimaciones sintéticas serán mucho más suaves comparadas con las basadas en el modelo en relación con los modelos donde o no hay efecto aleatorio o su variabilidad es muy pequeña (IMTOT e IM_AJENA).

A efectos prácticos se vio que las áreas con mayores ingresos totales, IMTOT, son las que contienen a las siete principales ciudades gallegas. Dentro de ellas las de mayores ingresos son A Coruña y Santiago con una media 2188 y 2251 euros por hogar, respectivamente. El área con menores ingresos es Ourense sur, con ingresos inferiores a 1600 euros. El único área donde teníamos muchas variaciones anuales era en el área de O Morrazo.

Para los ingresos por cuenta ajena y propia, IM_AJENA y IM_AJENA, también las áreas que contienen a las ciudades son las que tienen mayores ingresos de estos tipos. Para IM_CONTRIB las áreas con mayores ingresos son Ferrol y A Coruña. Y por último, para IM_NO_CONTRIB, tenemos áreas de las no principales con ingresos altos; por ejemplo en las áreas de A Barbanza y Lugo Oriental. Aún así, para IM_NO_CONTRIB los ingresos son muy similares para todas las áreas.

Los resultados obtenidos por la selección del mejor modelo mediante el cAIC han sido comprobados por uno de los métodos clásicos de selección de variables como es el "método backward".

Apéndice A

Código de los AIC para modelos mixtos utilizados en la memoria

A continuación mostraremos los códigos utilizados para calcular los AIC de *Vaida y Blanchard* (2005) y *Han* (2013) en R. Se mostrará, por ejemplo, los AIC para el modelo finalmente elegido para IMTOT para el año 2013; en el resto de modelos se calculan de manera análoga.

```
#datos: Conjunto de datos con las variables respuesta, variables
explicativas y varianza del error muestral para cada una de las cinco
variables respuesta

#datos$ftot: valor de IMTOT para las veinte áreas gallegas en el año 2013.

#datos$nper18a64: valor de NPER_18A65 para las veinte áreas gallegas en el
año 2013.

#datos$nperestudos_sup: valor de NPER_SUP para las veinte áreas gallegas en
el año 2013.

#datos$fogar_baixo_limiarpob: valor de HOG_BAJO_UMBRAL para las veinte áreas
gallegas en el año 2013.

#datos$rendimiento: valor de RENDIMIENTO para las veinte áreas gallegas en
el año 2013.

#datos$var1: σ²_cd para IMTOT.

Modelo:
mod=eblupFH(ftot ~ 0+nper18a64+nperestudos_sup+fogar_baixo_limiarpob+
rendimiento, vardir=var1,data=datos)
```

```
################################
#cAIC Vaida y Blanchard (2005)
################################

#Matriz diagonal de unos de tamaño 20x20:
Z=matrix(diag(20),20,20)

#Matriz de las varianzas del efecto aleatorio, Σ̂_u:
Vu1=mod15$fit$refvar*Z

#Matriz de las varianzas de los errores muestrales, Σ̂_ϵ:
V1=diag(datos$var1,20,20)

V=V1+Vu1

#Matriz hat:
H=X%*%(solve(t(X)%*%solve(V)%*%X))%*%t(X)%*%solve(V)+
Vu1%*%t(Z)%*%(solve(V))- Vu1%*%t(Z)%*%(solve(V))
%*%X%*%(solve(t(X)%*%(solve(V))%*%X))%*%t(X)%*%solve(V)

#Traza de la matriz hat:
ro=sum(diag(H))

#Verosimilitud condicional:
L=-20/2*log(2*pi)-1/2*log(det(V1))-1/2*t(datos$ftot-mod15$eblup)
%*%(solve(V1))%*%(datos$ftot-mod15$eblup)

# cAIC de Vaida y Blanchard (2005)
cAIC_VB=-2*L+2*ro; cAIC_VB



#######################
# cAIC Han (2013)
#######################


#Factores utilizados para calcular el término de penalización del cAIC de
Han (2013)

P=Z-X%*%solve(t(X)%*%solve(V)%*%X)%*%t(X)%*%solve(V)
```

```
rs=solve(V)%*%(P)%*%as.matrix(datos$ftot)

T=(solve(V)%*%P)**2

KK=sum(diag(T))-2*t(rs)%*%solve(V)%*%P%*%rs

#Término penalización:
KKK=ro-2*((KK)**-(1))*t(rs)%*%solve(V)%*%P%*%V1%*%solve(V)%*%P%*%rs

#cAIC de Han (2013):
AIC_Han=-2*L+2*KKK; AIC_Han
```

Bibliografía

[1] Akaike, H. (1974) *A New Look at the Statistical Model Identification.* IEEE Transactions on Automatic Control 19 (6): 716-723.

[2] Battese G. E. and Coelli T.J. (1988) *Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data.* Journal of Econometrics 38, 387-399.

[3] Burnham and Anderson (2002) *Model Selection and Multimodel Inference*.

[4] Fay R. E. and Herriot R. A. (1979) *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data*.

[5] Fisher, R.A. (1925) *Statistical Methods for Research Workers (1st edition*. Oliver and Boyd. Edinburgh

[6] González-Manteiga W. et al. (2008) *Multivariante Fay-Herriot model*. Computational Statistics and Data Analysis 52, pp 5242-5252.

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

[8] Hartley, H.O. y Rao, J.N.K. (1967) *Maximum Likelihood estimation for the mixed analysis of variance model*. Biometrika, 54:93-108

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

[10] Jiang et al. (2008). *Fence methods for mixed model selection.* Annals of Statistics. 36, 1669-1692.

[11] Lange and Rian (1989) *Assessing Normality in Random Effects Models*. The Annals of statistics, pp 624-642.

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

[13] ONS (2004): *Labour force survey user guide.* Tecnica Report Vol 6, Office for National Statistics, United Kingdon.

[14] Pan, Z. and Lim, D.Y. (2005). *Goodness-of-fit methods for generalized linear mixed models.* Biometrics 61, 1000-1009.

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

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

[17] Prasad and Rao (1990) *The estimation of the mean squared error of small-area estimation*. Journal of the American Statistical Association.

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

[19] Schwarz, Gideon E. (1978). *Estimating the dimension of a model.* Annals of Statistics 6 (2): 461-464.

[20] Smith, D. M., Robertson, B., and Diggle. P. J. (1996) *Object-oriented Software for the Analysis of Longitudinal Data in S.* Tecnical Report MA 96192. Dept. of Mathematics and Statistics, University of Lancaster.

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