Trabajo Fin de Máster

Uso de métodos de estimación en áreas pequeñas para mejorar las estimaciones proporcionadas por la EPA

Alexandro Aneiros Batista

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Uso de métodos de estimación en áreas pequenas para mellorar as estimacións proporcionadas pola EPA

Título en español: Uso de métodos de estimación en áreas pequeñas para mejorar las estimaciones proporcionadas por la EPA

English title: Usage of the methodology in small areas at the area level for improving official statistics estimators in the LFS

Modalidad: Modalidad A

Autor/a: Alexandro Aneiros Batista, Universidade da Coruña

Director/a: María José Lombardía Cortiña, Universidade da Coruña; Esther López Vizcaíno, Universidad de Santiago de Compostela

Tutor/a: , ; ,

Breve resumen del trabajo:

Este estudio explora el uso de modelos mixtos, en particular los de Fay-Herriot, para la estimación en áreas pequeñas con datos de la Encuesta de Población Activa. Se abordan enfoques univariantes y bivariantes aplicados a datos composicionales, evaluando su precisión con información del INE en el marco del CIDMEFEO. Se discuten sus aportes metodológicos, ventajas y posibles aplicaciones en la estadística oficial.

Recomendaciones:

Otras observaciones:

Don/doña María José Lombardía Cortiña, Catedrática en Estadística e I.O. de la Universidade da Coruña, don/doña Esther López Vizcaíno, Doctora en Estadística e I.O. de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Uso de métodos de estimación en áreas pequeñas para mejorar las estimaciones proporcionadas por la EPA

fue realizado bajo su dirección por don/doña Alexandro Aneiros Batista para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, Don/doña María José Lombardía Cortiña, don/doña Esther López Vizcaíno y don/doña Alexandro Aneiros Batista

√ sí ☐ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En A Coruña, a 03 de Junio de 2025.

El/la director/a:
Don/doña María José Lombardía Cortiña

LOMBARDIA CORTIÑA MARIA JOSE - 33322690Z
Firmado digitalmente por LOMBARDIA CORTIÑA MARIA JOSE - 33322690Z
Fecha: 2025.06.02 09:38:10 +02'00'

El/la director/a:
Don/doña Esther López Vizcaíno

MARIA ESTHER LOPEZ VIZCAINO - 76575190W
Firmado digitalmente por MARIA ESTHER LOPEZ VIZCAINO - 76575190W
Fecha: 2025.06.02 18:08:13 +02'00'

El/la autor/a:
Don/doña Alexandro Aneiros Batista

ANEIROS BATISTA ALEXANDRO - 47388770F
Firmado digitalmente por ANEIROS BATISTA ALEXANDRO - 47388770F
Fecha: 2025.06.01 15:48:33 +02'00'

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un Capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumen

Resumen en español

En el presente trabajo se llevará a cabo un estudio detallado sobre la aplicabilidad de los modelos mixtos a nivel de área dentro del marco de la estimación en áreas pequeñas, un ámbito clave de la inferencia en poblaciones finitas cuyo propósito es mejorar las estimaciones en dominios con muestras reducidas. Para ello, se abordarán las bases metodológicas fundamentales de este campo, incluyendo las técnicas de muestreo y estimación directa, con especial énfasis en su implementación en la Encuesta de Población Activa y en el *benchmarking*.

El análisis se centrará en los modelos mixtos lineales como base para el estudio de los modelos de Fay-Herriot, tanto en su versión univariante como multivariante. En este contexto, se hará hincapié en la formulación bivariante y su aplicabilidad a datos composicionales, considerando los desafíos que ello conlleva en términos de modelización e inferencia.

Asimismo, se evaluará el desempeño de estos modelos en la estimación de diversas variables laborales y socioeconómicas de interés para el Instituto Nacional de Estadística, utilizando datos reales proporcionados por dicho organismo en el marco de la Línea 7 de investigación del CIDMEFEO. A lo largo del estudio, se analizará el impacto de estas metodologías en la mejora de la precisión de las estimaciones y su posible aplicación en estudios oficiales. Finalmente, se discutirán las ventajas y limitaciones de los modelos propuestos, así como su potencial para extenderse a otros ámbitos de la estadística oficial y aplicada.

Palabras clave: Modelos de Fay-Herriot, *Benchmarking*, EPA, Transformación aditiva logística.

English abstract

In this study, a detailed analysis will be conducted on the applicability of area-level mixed models within the framework of small area estimation, a key field in finite population inference aimed at improving estimates in domains with small sample sizes. To this end, the fundamental methodological foundations of this field will be addressed, including sampling techniques and direct estimation, with particular emphasis on their implementation in the Labour Force Survey and *benchmarking*.

The analysis will focus on linear mixed models as a basis for studying Fay-Herriot models, both in their univariate and multivariate versions. In this context, special attention will be given to the bivariate formulation and its applicability to compositional data, considering the challenges it entails in terms of modeling and inference.

Furthermore, the performance of these models will be evaluated in the estimation of various labor and socio-economic variables of interest to the National Institute of Statistics, using real data provided by the same institution as part of Research Line 7 of CIDMEFEO. Throughout the study, the impact of these methodologies on improving estimation accuracy and their potential application in official studies will be analyzed. Finally, the advantages and limitations of the proposed models will be discussed, as well as their potential to be extended to other areas of official and applied statistics.

Keywords: Fay-Herriot Models, Benchmarking, LFS, Additive Log-ratio Transformation.

Capítulo 1

Introducción

Desde la aparición de la necesidad de contabilizar bienes y registrar posesiones, tanto en el ámbito público como privado, y tanto a nivel individual como colectivo, las matemáticas han desempeñado un papel fundamental en la evolución de los sistemas administrativos y económicos. Existen registros que evidencian el empleo de sistemas de contabilidad desde la antigua Sumeria, alrededor del 3500 a.C., en forma de tablillas de arcilla utilizadas para llevar inventarios y controlar transacciones económicas (Katz, 2009). Con el transcurso de los siglos, se desarrollaron métodos empíricos para cuantificar magnitudes de interés, tales como medidas de tendencia central o dispersión, con el objetivo de evaluar los recursos disponibles y mejorar la gestión administrativa.

El desarrollo de la estadística permitió la formalización de estos procedimientos, estableciendo metodologías matemáticas rigurosas que culminaron en la conformación de instituciones dedicadas a la recopilación, análisis y difusión de datos de interés socioeconómico. En la actualidad, se dispone de organismos oficiales, como el Instituto Nacional de Estadística (INE) en España, que garantizan la producción de estadísticas de carácter público, así como de entidades privadas que elaboran informes estadísticos con fines comerciales, especialmente en sectores como el financiero o el asegurador.

En este contexto, emerge la estadística pública u oficial, definida como la producción de información estadística de interés general con el propósito de orientar la toma de decisiones tanto en el ámbito gubernamental como en el sector privado. Esta información es generada a nivel nacional por organismos como el INE y, en el ámbito supranacional, por instituciones como EUROSTAT, la Oficina Europea de Estadística. La estadística pública se fundamenta en disciplinas metodológicas clave, entre las que destacan el muestreo y la regresión, así como en métodos avanzados de estimación, como la estimación en áreas pequeñas (más conocido en inglés como *Small Area Estimation* y bajo las siglas SAE), que constituye el núcleo de la presente investigación.

En términos operativos, las encuestas diseñadas para la obtención de datos poblacionales emplean técnicas de muestreo que permiten la estimación de parámetros de interés con una precisión aceptable. No obstante, la magnitud de la población, los costos asociados a la recolección de información y las limitaciones inherentes a los diseños muestrales pueden generar situaciones en las que ciertos grupos o áreas geográficas no sean adecuadamente representados en la muestra. En algunos casos, el tamaño muestral dentro de un subconjunto específico es demasiado reducido para obtener estimaciones directas fiables, es decir, aquellas obtenidas exclusivamente a partir de la muestra sin el uso de modelos auxiliares.

Para abordar esta problemática, se recurre a estimaciones indirectas basadas en modelos estadísticos que integran información proveniente de otras fuentes. Dichos modelos permiten mejorar la precisión de las estimaciones al incorporar variables auxiliares que poseen una fuerte correlación con la variable de interés, o, por lo menos, ayudan al modelo en el ajuste dentro del procedimiento de selección de variables (Lombardía et al. 2017, 2018). Por ejemplo, en estudios sobre pobreza, la renta media por hogar es una variable auxiliar de alta relevancia, mientras que en análisis de empleo y contratación, la proporción de población extranjera puede desempeñar un papel clave. En el ámbito de la estimación

en áreas pequeñas, estas variables auxiliares suelen proceder de registros administrativos oficiales, tales como los proporcionados por la Tesorería General de la Seguridad Social (TGSS), o de censos poblacionales elaborados por el INE (Rao y Molina 2015).

La propia denominación de estimación en áreas pequeñas plantea la cuestión de qué se considera un área pequeña. En términos generales, se define un área pequeña, o dominio, como un subconjunto de la población cuya representación muestral es insuficiente para realizar estimaciones directas con la precisión requerida (Jiang y Lahiri 2006). A menudo, estos subconjuntos corresponden a unidades geográficas, como municipios o provincias, donde el diseño muestral no garantiza una cobertura adecuada. No obstante, el concepto de área pequeña no se limita exclusivamente a divisiones geográficas, sino que puede incluir grupos poblacionales específicos definidos por características sociodemográficas, como la intersección de sexo, comunidad autónoma y nacionalidad extranjera (Jiang y Lahiri 2006). En tales casos, la insuficiencia de datos muestrales obliga a recurrir a métodos de estimación indirecta para obtener inferencias estadísticas robustas.

La metodología de estimación en áreas pequeñas se torna especialmente útil en situaciones en las que el diseño muestral ha sido concebido para proporcionar estimaciones a un determinado nivel de agregación, como el autonómico, pero se requiere información más detallada a niveles inferiores, como el provincial. Dado que el diseño original no permite obtener estimaciones directas fiables a estos niveles más desagregados, se hace necesario recurrir a la modelización estadística para generar estimaciones indirectas con una precisión aceptable (Pfeffermann 2013).

Con todo, la estimación en áreas pequeñas constituye una herramienta metodológica esencial en la estadística oficial, permitiendo la generación de estimaciones confiables en contextos donde los métodos tradicionales resultan insuficientes. Su desarrollo y aplicación han sido impulsados por la creciente demanda de información estadística detallada, lo que ha conducido a avances metodológicos significativos en la modelización estadística y el uso de fuentes de datos auxiliares (Pfeffermann 2013; Rao y Molina 2015). En este sentido, su integración en la producción estadística oficial representa un elemento clave para la mejora de la precisión y utilidad de los indicadores sociodemográficos y económicos.

Capítulo 2

Estado del Arte y metodología

La estimación en áreas pequeñas constituye un campo fundamental en estadística, orientado a la obtención de estimaciones precisas para subpoblaciones con tamaños de muestra reducidos. La literatura especializada ha desarrollado un marco teórico y metodológico extenso, como evidencian las revisiones de Rao (1999, 2008), Rao y Choudhry (1995), Ghosh y Rao (1994), Pfeffermann (2002, 2013), o Jiang y Lahiri 2006, destacando estos últimos por ser más introductorios y accesibles al público con poco conocimiento en la rama de la regresión con modelos mixtos. Adicionalmente, los trabajos de Rao (2003), Rao y Molina (2015), y especialmente Morales et al. (2021), ofrecen un análisis detallado de los enfoques basados en modelos, mientras que Pratesi (2016) enfatiza aplicaciones prácticas en ciencias sociales.

Los métodos de SAE pueden clasificarse en tres enfoques principales: basados en el diseño, asistidos por modelos y fundamentados en modelos. Los métodos basados en el diseño dependen exclusivamente de la muestra recolectada y buscan minimizar el sesgo mediante el uso de datos auxiliares sin imponer modelos específicos. En este contexto, los estimadores sintéticos incorporan información poblacional externa para mejorar la precisión (Rao 2003). No obstante, estos enfoques pueden generar sesgos si se incumplen los supuestos sobre los datos auxiliares.

Los enfoques asistidos por modelos integran modelos de regresión para mejorar la eficiencia de la estimación, manteniendo propiedades de insesgadez aproximada bajo el diseño. Ejemplo de ello es el estimador de regresión generalizada (Särndal et al. 1992), que incorpora covariables para reducir la variabilidad.

Por otro lado, las estrategias basadas en modelos asumen explícitamente procesos generadores de datos, lo que permite la incorporación de dependencias espaciales, temporales y jerárquicas. Los modelos a nivel de unidad, como el modelo de regresión con errores anidados (Battese et al. 1988), integran datos de encuestas y auxiliares a nivel individual. Extensiones de este enfoque incluyen predictores empíricos óptimos (Molina y Rao 2010) y enfoques de M-cuantiles (Tzavidis et al. 2008; Bugallo et al. 2024b). Los modelos a nivel de área, que operan con datos agregados, han sido ampliamente estudiados, destacando el modelo de Fay-Herriot (Fay y Herriot 1979) (FH) para la estimación de medias de dominio. Extensiones recientes incluyen componentes temporales (Marhuenda et al. 2013) y espaciales (Chandra et al. 2017), así como modelos lineales mixtos generalizados para resultados de conteos y proporciones (López-Vizcaíno et al. 2015; Boubeta et al. 2016). Estos avances evidencian la flexibilidad de SAE para abordar distintas estructuras de datos y problemáticas de investigación.

El desarrollo de modelos mixtos multivariantes ha sido crucial en SAE, particularmente en la extensión del modelo FH. Huang y Bell (2004), González-Manteiga et al. (2008) y Benavent y Morales (2016) han explorado la estimación de parámetros de interés en modelos multivariantes. Más recientemente, Krause et al. (2022b) han propuesto modelos multivariantes penalizados aplicados a datos de consumo de alcohol. Además, Burgard et al. (2021) y Esteban et al. (2020) han extendido estos modelos a la estimación con datos composicionales en áreas pequeñas.

En el ámbito de los modelos generalizados, los modelos de Poisson y binomiales han sido amplia-

mente estudiados para la estimación de conteos y proporciones. En este sentido, Boubeta et al. (2016a, 2017, 2023), Faltys et al. (2022) o Diz-Rosales et al. (2024) han propuesto predictores de SAE basados en modelos a nivel de área con modelos lineales mixtos generalizados. No obstante, una limitación común de estos modelos es su incapacidad para manejar datos con exceso de ceros. Un enfoque para abordar esta limitación consiste en ajustar un modelo FH tras una transformación, seguido de una estimación de varianza no nula cuando el valor observado es cero (Berg y Fuller, 2012). Otra alternativa es el uso de modelos inflados en cero, que permiten modificar la probabilidad de la variable objetivo para capturar estructuras de datos con alta frecuencia de ceros (Bugallo et al. 2024a).

2.1. Modelos a Nivel de Área

Los modelos a nivel de unidad representan una herramienta estadística de gran potencia para describir una variable objetivo, siempre que se ajusten adecuadamente a los datos observados. No obstante, su aplicación práctica conlleva una serie de restricciones importantes que limitan su potencial predictivo. En particular, para el cálculo de los predictores empíricos óptimos lineales insesgados (EBLUPs, por sus siglas en inglés), que serán tratados en detalle en la sección 3.3, de parámetros lineales a nivel de dominio bajo un modelo mixto lineal a nivel de unidad, se requiere un archivo auxiliar con los promedios de dominio de las variables auxiliares seleccionadas. Esta exigencia impone una reducción sustancial del conjunto de variables disponibles, dado que muchas de ellas no cuentan con estimaciones agregadas confiables a nivel de dominio, lo cual merma significativamente la capacidad explicativa del modelo (Morales et al. 2021). A esto se suma el hecho de que dichos promedios suelen derivarse de registros administrativos, que en muchos casos no utilizan las mismas definiciones, metodologías o periodos de referencia que las encuestas muestrales. Esta disparidad entre fuentes es una causa habitual de incompatibilidad que puede introducir sesgos en la estimación y afectar la validez de los resultados.

Cuando, además, se requiere estimar parámetros no lineales mediante predictores empíricos óptimos (EBPs) bajo el mismo tipo de modelo, la situación se torna aún más compleja, pues es necesario contar con un archivo censal que contenga los valores individuales de las variables auxiliares seleccionadas, lo cual representa un obstáculo considerable para la mayoría de las oficinas estadísticas, especialmente en países donde el acceso a datos censales detallados y de alta calidad es limitado (Rao y Molina 2015). Si bien existe una alternativa metodológica en aquellos casos en que las variables auxiliares son categóricas, es decir, que adoptan un número finito de valores discretos, y se dispone del tamaño poblacional de los dominios cruzados por categorías, este enfoque, de inspiración ANOVA, sólo resulta adecuado en ciertas circunstancias. Los modelos basados exclusivamente en variables categóricas tienden a mostrar un desempeño predictivo limitado, y su aplicabilidad práctica se ve aún más restringida cuando se introducen variables continuas. En tal caso, la implementación del procedimiento de estimación por *bootstrap* exige la generación de la población completa en cada iteración, lo que conlleva un costo computacional que puede hacer inviable su utilización en estudios reales.

Frente a estas limitaciones, cuando no se dispone de información auxiliar individual confiable, pero sí existen datos agregados a nivel de área obtenidos de registros administrativos, los modelos pueden reformularse a nivel de área. Este enfoque fue establecido formalmente por Fay y Herriot en su artículo de 1979, donde propusieron un modelo lineal mixto con efectos aleatorios para estimar el ingreso per cápita promedio en pequeñas localidades de Estados Unidos (Fay y Herriot 1979). Desde entonces, el EBLUP basado en este modelo se ha consolidado como una de las metodologías más empleadas en la inferencia para áreas pequeñas, gracias a su capacidad para combinar información directa proveniente de encuestas con predictores construidos a partir de variables auxiliares agregadas (Jiang y Lahiri 2006; Pfeffermann 2013). Aunque este enfoque presenta ciertas limitaciones, entre ellas, la dependencia de la precisión de las estimaciones directas y la calidad de los agregados auxiliares utilizados como predictores, constituye una alternativa metodológica sólida y, en muchos casos, más factible, especialmente en contextos donde la información microestadística resulta escasa, fragmentaria o de calidad cuestionable.

El uso del modelo FH implica una pérdida de información derivada de la agregación de datos a

nivel de unidad. No obstante, presenta varias ventajas indiscutibles que lo hacen un candidato a tener en cuenta dentro de la SAE: permite la utilización de un mayor número de variables auxiliares; elimina la restricción de los modelos a nivel de unidad de requerir las mismas variables auxiliares tanto en el archivo muestral de la encuesta como en los registros administrativos externos; y ofrece buenos resultados en comparación con los modelos a nivel de unidad cuando el número de áreas pequeñas es grande y los tamaños muestrales son reducidos (Rao y Molina 2015; Morales et al. 2021).

A lo largo de los años, el modelo FH ha sido ampliamente estudiado y extendido en múltiples direcciones, lo que ha permitido su aplicación en contextos cada vez más diversos. Una de las principales líneas de investigación se ha centrado en la estimación del error cuadrático medio (MSE, por sus siglas en inglés) del EBLUP. En este sentido, se han desarrollado diversos métodos y aproximaciones, entre los que destacan los trabajos pioneros de Prasad y Rao (1990), artículo de referencia y base de la estimación del error, seguidos por las aportaciones de Datta et al. (2000), así como de Das et al. (2004). Posteriormente, Hall y Maiti (2006a, 2006b), Esciulescu y Fuller (2015) y Slud y Maiti (2006) propusieron mejoras en la estimación del MSE por métodos de remuestreo *bootstrap*, tanto por el lado paramétrico como por el no paramétrico, mientras que González-Manteiga et al. (2010) y Datta et al. (2011) introdujeron metodologías que optimizan la eficiencia del EBLUP. Asimismo, Kubokawa (2011) exploró aproximaciones alternativas para mejorar la precisión de las estimaciones.

Esta línea de trabajo ha evolucionado hasta abordar la construcción de intervalos de confianza y pruebas de hipótesis, como se observa en los estudios de Diao et al. (2014), Molina et al. (2015) y Marhuenda et al. (2016). Además, Jiang y Tang (2011) examinaron el impacto que tienen los estimadores de los parámetros del modelo en la eficiencia del predictor empírico.

Otro aspecto clave en el desarrollo del modelo FH ha sido su extensión para abordar diferentes problemáticas. En particular, se han propuesto modificaciones en su formulación bayesiana, como los modelos transformados paramétricos de Sugasawa y Kubokawa (2015) basándose en modelos con transformación del tipo Box-Cox y la propia estimación del parámetro de Box-Cox, y los enfoques semiparamétricos generalizados introducidos por Polettini (2017). Adicionalmente, una dificultad recurrente en la aplicación del modelo es la presencia de errores de medición en las variables auxiliares. Para mitigar este problema, diversos estudios han propuesto soluciones metodológicas específicas, incluyendo los trabajos de Ybarra y Lohr (2008), Arima et al. (2015), Datta et al. (2018) y Burgard et al. (2020). Por otro lado, la selección óptima del modelo es un factor determinante en su desempeño, razón por la cual Marhuenda et al. (2014) y Lombardía et al. (2017, 2018) exploraron la adaptación del criterio de información de Akaike para el contexto del modelo FH.

Finalmente, aunque no se vayan a detallar, se han desarrollado numerosas extensiones que incluyen la consideración de distribuciones no normales, la incorporación de métodos robustos y semiparamétricos, la inferencia simultánea y el diseño de procedimientos bayesianos jerárquicos, lo que ha permitido ampliar significativamente el ámbito de aplicación del modelo FH en la estimación para áreas pequeñas (Tzavidis et al. 2018; Diz-Rosales et al. 2023; Bugallo et al. 2024b). En efecto, y al respecto de la no normalidad y robustez, en los modelos FH, aunque por defecto se asuma normalidad tanto en los errores independientes del modelo como en los efectos aleatorios, como veremos en el apartado donde se detallará la teoría sobre la que se sustenta este trabajo, Jiang (1996, 1997, 1998) demostró que los estimadores son robustos incluso bajo hipótesis de no normalidad.

Al respecto de los modelos mixtos a nivel de área con datos composicionales, al ser un campo nuevo de aplicación, no existen numerosos artículos. Sin embargo, se destaca el artículo de Esteban et al. (2020), donde se exponen dichos modelos aplicados al estudio de variables laborales, como son las estimaciones de totales de ocupados, parados, inactivos y menores de 16 años a nivel municipal en Galicia. A su vez, emplea un *bootstrap* paramétrico para la estimación del MSE. En el caso de ser a nivel individuo, un modelo similar se emplea en Esteban et al. (2023).

2.2. Objetivos

Este estudio se enfocará en el desarrollo y aplicación de modelos a nivel de área, con especial énfasis en los modelos de Fay-Herriot (FH). En particular, se abordará la extensión de estos modelos al ámbito de los datos composicionales, considerando su formulación bivariante, ya que su generalización al caso multivariante es inmediata. El tratamiento del caso univariante se considera inmediato, dado que cualquier variable única puede ser interpretada dentro de un marco composicional al tomar como referencia su complementario. Esta perspectiva permite englobar dicho caso dentro del mismo enfoque metodológico, asegurando así una coherencia conceptual en el análisis y la teoría subyacente.

Se llevará a cabo un estudio detallado sobre la aplicabilidad de los modelos FH a los datos provenientes de la Encuesta de Población Activa (EPA), con el objetivo de estimar diversos indicadores socioeconómicos y laborales de interés. En particular, se considerarán las estimaciones de los totales de parados, ocupados e inactivos, así como la tasa de desempleo y la cuantificación de la población joven de entre 16 y 24 años que ni estudia ni trabaja (denominados «ninis»).

Más concretamente, se detallan a continuación los objetivos propuestos por el INE, los cuales conforman gran parte de este trabajo.

1. **Objetivo 1: Totales de personas paradas, ocupadas e inactivas, así como la tasa de desempleo a nivel municipal.** Este objetivo tiene como objetivo un estudio para aquellos municipios con más de 16.000 habitantes (según el Padrón de Habitantes) a lo largo de los ocho trimestres de 2021 y 2022. En este caso, se emplearon modelos FH bivariantes aplicados a datos composicionales.

2. **Objetivo 2: Número de jóvenes «ninis» a nivel provincial,** también a lo largo de los ocho trimestres de 2021 a 2022. En este objetivo se empleó un modelo univariante del tipo FH.

3. **Objetivo 3: Totales de personas paradas, ocupadas e inactivas y tasas de desempleo, desglosada por sexo y por las cinco principales nacionalidades extranjeras en España: Marruecos, Colombia, Venezuela, Reino Unido y Rumanía, a nivel autonómico** y también a lo largo de los ocho trimestres de 2021 a 2022. El modelo empleado es el mismo que en el primer objetivo, pero habiendo realizado un tratamiento de ceros más cuidadoso debido a la aparición de cruces, muchos de ellos con apenas muestra o ninguna.

No obstante, se advierte que la presente labor centrará su atención de manera preeminente en los dos primeros objetivos, pues el tercero no es sino un caso particular del primero, en el que se han empleado modelos idénticos. En efecto, no se abordará dicho objetivo de manera directa, sino que se hará tan solo una sucinta referencia, en la medida en que constituye un escenario en el que la aparición de valores nulos puede tornarse un obstáculo insoslayable para la correcta estimación y ajuste de los modelos y parámetros de interés, los cuales, dicho sea de paso, son los mismos que en el primer objetivo.

En todo momento, se adoptará como criterio de referencia para la confiabilidad de las estimaciones aquel establecido por la Office for National Statistics (ONS, 2006), según el cual se consideran publicables todas aquellas estimaciones cuyo coeficiente de variación (CV) se mantenga por debajo del umbral del 20 %. No obstante, cabe señalar que distintos institutos nacionales de estadística alrededor del mundo emplean metodologías y métricas alternativas para determinar la potencialidad de publicación de los datos. En este sentido, se destaca los estándares propuestos en el *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys* (Methodologies 2013), donde se detallan los requisitos de precisión y los criterios de variabilidad adoptados en encuestas de hogares dentro del marco del *European Statistical System* (ESS).

De esta manera, el presente trabajo no solo proporcionará un análisis teórico de los modelos FH y su adecuación al tratamiento de datos composicionales, sino que también explorará su implementación en la estadística pública, evaluando su potencial para mejorar y elevar la calidad de las estimaciones derivadas de encuestas oficiales.

2.3. Material empleado y Bases de Datos

El presente estudio se ha desarrollado empleando herramientas computacionales usuales dentro del análisis estadístico, destacando el uso del software `R` (R Development Core Team, 2024), el cual ha sido fundamental tanto para la implementación computacional de los modelos propuestos como para el tratamiento y la depuración de los datos utilizados. Adicionalmente, se ha recurrido a diversas librerías especializadas que permiten la manipulación eficiente de grandes volúmenes de información, el ajuste de modelos de inferencia estadística y la evaluación de la precisión de las estimaciones obtenidas, al igual que código propio para la estimación del MSE por *bootstrap* y para resumir todo el proceso de estimación del modelo en una sola función, el cual se expone en el Anexo 2.

Con respecto a los datos empleados, estos han sido obtenidos en el marco del trabajo desarrollado en la Línea 7 de investigación del CIDMEFEO por la Universidade da Coruña (UDC), en colaboración con el INE. En dicho contexto, el INE estableció una serie de objetivos en los cuales los modelos FH previamente descritos fueron aplicados. La base de datos utilizada en este estudio está constituida por las siguientes fuentes de información, de las cuales, a excepción de las dadas por el INE, todas tuvieron que ser obtenidas manualmente a partir de diversos códigos de `R`:

- Microdatos de la Encuesta de Población Activa (EPA), proporcionados por el INE.
- Datos de renta proporcionados por el INE.
- Registros de parados inscritos en las oficinas de empleo público, provenientes del Servicio Estatal Público de Empleo (SEPE).
- Registros de pensionistas que perciben pensiones contributivas, obtenidos del Instituto Nacional de la Seguridad Social (INSS).
- Datos de afiliaciones a la Seguridad Social, provenientes de la TGSS.
- Datos de población de 16 y más años desagregados por nacionalidad, extraídos del Marco de Personas del INE y proporcionados por el INE.

Con el objetivo de garantizar la coherencia y la integridad de la información, todos estos datos fueron integrados en una única base de datos, habiéndose llevado a cabo un proceso exhaustivo de limpieza, depuración y armonización de las distintas fuentes. Sin embargo, en el transcurso de este trabajo, se han identificado y abordado diversas dificultades inherentes a la integración y compatibilización de los datos, al igual que a su naturaleza en cuanto a su calidad se refiere, entre las que se destacan:

- **Definición del conjunto de municipios de interés:** Se estableció como criterio la selección de municipios con una población censada superior a 16.000 habitantes, para el Objetivo 1. Para garantizar la estabilidad en la estimación de los indicadores a lo largo del tiempo, se consideraron únicamente aquellos municipios que figuraban de manera consistente en los ocho trimestres analizados de la EPA (2021-2022).
- **Ausencia de identificadores municipales en los registros de pensiones contributivas:** Los datos publicados por la Seguridad Social carecían de información desagregada por municipio, lo que dificultó su integración con el resto de la base de datos. Para mitigar este problema, se implementaron técnicas de emparejamiento basadas en similitud textual entre los nombres de los municipios presentes en distintas fuentes oficiales.
- **Inconsistencias en la información sobre perceptores de pensiones en los microdatos de la EPA:** Se detectaron anomalías en la variable correspondiente a los perceptores de pensiones contributivas en algunos municipios, especialmente en los trimestres más recientes. En varios casos, se observaron valores reportados como nulos, lo cual resulta altamente improbable desde un punto de vista demográfico. Se implementaron métodos de imputación usuales en la literatura de la SAE para corregir estas inconsistencias.

- **Desajustes entre los datos de ocupados en la EPA y las afiliaciones a la Seguridad Social:** Se identificaron discrepancias significativas entre la información proporcionada por la EPA sobre la población ocupada residente y los registros administrativos de afiliaciones a la Seguridad Social. Esta divergencia se debe a diferencias metodológicas entre ambas fuentes: la EPA emplea un criterio de residencia, mientras que los datos de afiliaciones están basados en el lugar de trabajo. El desajuste es especialmente notorio en municipios de gran tamaño que actúan como polos de atracción laboral. En la modelización final, se optó por emplear la información sobre afiliaciones provenientes de los microdatos de la EPA construidos por el INE como variable auxiliar.

- **Dificultades en la integración de los datos de parados registrados en los microdatos de la EPA:** Se encontraron inconsistencias en los registros de parados, lo que impidió su uso directo en los modelos. Para abordar este problema, se llevó a cabo un proceso de validación y ajuste de los datos, evaluando posibles sesgos e implementando técnicas de estimación alternativa para mitigar los efectos de las anomalías detectadas.

Cabe señalar que, además de los desafíos asociados a la integración de fuentes heterogéneas, el presente estudio ha requerido el desarrollo de metodologías no implementadas en las librerías estándar de `R`. En particular, se han diseñado estrategias específicas para la aplicación del *bootstrap* paramétrico en la estimación de la variabilidad de las transformaciones aplicadas a las variables auxiliares, así como métodos para la implementación de dichas transformaciones en un entorno optimizado de cálculo paralelo. Estas estrategias han sido esenciales para mejorar la eficiencia computacional y garantizar la obtención de resultados de alta precisión en tiempos razonables.

2.4. Estructuración del trabajo

A continuación, se expone la estructura general sobre la cual se articulará el presente trabajo. Tras la introducción y el estado del arte previamente desarrollados, junto con la exposición detallada de los objetivos del estudio, así como la descripción del material empleado y las dificultades inherentes a su implementación, se procederá al desarrollo de los contenidos siguiendo una secuencia lógica, donde se tratarán los fundamentos de la inferencia en poblaciones finitas y el muestreo, todo ello aplicado al contexto que concierne a este trabajo, la SAE, y la explicación de los modelos mixtos para proseguir con los modelos FH.

En primer lugar, se presentará en el tercer Capítulo un análisis preliminar de las metodologías fundamentales en el ámbito de la estimación en áreas pequeñas, introduciendo los métodos de estimación directa y los modelos mixtos, y de la misma forma una exposición breve acerca del funcionamiento y construcción y *raison d'être* de la EPA y del *benchmarking*. Estos modelos mixtos constituyen la base conceptual a partir de la cual se construyen los modelos a nivel de área, facilitando así su posterior generalización. Se ofrecerá una revisión de estas técnicas, subrayando sus principales ventajas y limitaciones, justificando la necesidad del uso de modelos más sofisticados en escenarios donde la estimación directa resulta insuficiente debido a la escasez de datos en los dominios considerados.

A continuación, en el cuarto Capítulo, se abordará en profundidad el estudio de los modelos a nivel de área, proporcionando una exposición formal de los principios estadísticos en los que se fundamentan. Se iniciará con una descripción general de estos modelos, seguida de su especificación particular en el contexto de los modelos FH, los cuales constituyen la herramienta metodológica más utilizada en este campo. Dentro de este marco, se analizará en primer lugar el caso univariante, con atención a su versión estándar. Posteriormente, se procederá a la formulación del modelo FH en su versión bivariante dado su relevancia en la resolución de los problemas planteados en los objetivos 1 y 3 propuestos por el INE, mencionando su generalización teórica al caso multivariante. Análogamente, se tratará el aspecto relativo al uso de estos modelos para el caso de datos composicionales, datos en los que se basa este proyecto en los dos objetivos previamente mencionados, y que corresponden al interés de este trabajo.

En el Capítulo quinto se describirá la implementación empírica de los modelos aplicándolos a los datos proporcionados por el INE, es decir, a datos reales. Se detallará la metodología empleada para la puesta en producción de estos modelos, evaluando su desempeño y la calidad de las estimaciones obtenidas. Se analizará asimismo su aplicabilidad en la estimación de indicadores socioeconómicos y laborales, comparando los resultados con las estimaciones directas tradicionalmente utilizadas por el INE. A partir de este análisis comparativo, se identificarán las principales mejoras que los modelos FH pueden aportar en términos de precisión y estabilidad de las estimaciones, así como sus eventuales limitaciones. Se incluirá un apartado dedicado al estudio de la aplicabilidad de los modelos a datos reales en el quinto Capítulo, incluyendo la validación de los resultados obtenidos, mediante el uso de técnicas de diagnóstico que permitan evaluar la idoneidad de los supuestos adoptados y la fiabilidad de las estimaciones generadas a partir de comparación con los datos publicados por el INE y la distribución geográfica de los mismos.

Por último, el trabajo concluirá con el sexto Capítulo en el que se sintetizarán, a modo de cierre, los hallazgos más relevantes derivados de la investigación, destacando las contribuciones realizadas al conocimiento en esta área de la estadística. Además, se esbozarán posibles líneas futuras de investigación, explorando extensiones y mejoras metodológicas que permitan continuar avanzando en el desarrollo de modelos de estimación en áreas pequeñas, donde se destacará su aplicabilidad en contextos de estadística oficial y pública y su integración en sistemas de producción estadística de gran escala, ya que parte de los objetivos subyacentes al INE es la implementación de estos modelos propuestos como metodología base y alternativa.

En el primer Anexo se expondrán algunos resultados numéricos más detallados y mapas que, por sus dimensiones, se apartan al anexo para no perturbar la lectura del documento. Además, con el objetivo de facilitar la comprensión y replicación[1] del trabajo realizado, se incluirán en el anexo 2 una serie de tutoriales que ilustran el uso del software empleado en el análisis, mientras que en la sección 2 de dicho Anexo 2 se incluirán algunas funciones creadas para los objetivos del INE y otras usadas expuestas en Morales et al. (2021). Estos tutoriales están diseñados utilizando datos simulados o similares a los originales y permiten al lector familiarizarse con los métodos aplicados, así como adaptar las herramientas a contextos propios o investigaciones futuras.

[1] Se menciona que los datos exactos dados por el INE no pueden ser proporcionados de forma alguna por secreto estadístico, de forma que la replicación se entiende en el sentido de cómo usar las funciones empleadas y el esquema general del estudio.

Capítulo 3

Estudio de las bases fundacionales de la estadística en áreas pequeñas

Antes de abordar en detalle los modelos empleados en la SAE, resulta imprescindible establecer las medidas características que constituyen la base conceptual sobre la que se sustentan estas metodologías. Entre ellas, destacan las estimaciones directas, las cuales, como ya se ha mencionado, consisten en la obtención del parámetro de interés a partir de la información muestral disponible dentro de cada dominio, sin recurrir a ningún tipo de modelo estadístico, ya sea el modelo FH o los modelos basados en individuos. Estas estimaciones, aunque conceptualmente sencillas y directamente interpretables, suelen presentar alta variabilidad cuando los tamaños muestrales dentro de los dominios son reducidos, lo que justifica la necesidad de enfoques basados en modelos en áreas pequeñas que permitan mejorar la precisión de las estimaciones.

También se incluirá un apartado específico en el que se detallará el funcionamiento y la construcción de la EPA, dado que constituye la base fundamental de este trabajo y el núcleo duro de los datos disponibles. Asimismo, se abordará el concepto de *benchmarking*, ya que será esencial para el cumplimiento del Objetivo 2, en respuesta a la necesidad de garantizar la conformidad con los datos publicados por el INE.

Adicionalmente, se introducirá el marco teórico sobre el cual se construyen los modelos utilizados en la SAE, concretamente, los modelos de regresión mixtos. Estos modelos extienden la regresión clásica al considerar, además de los errores habituales del modelo de regresión, la presencia de efectos aleatorios que capturan la variabilidad atribuible a la existencia de agrupaciones naturales en los datos. En este contexto, las agrupaciones pueden corresponder a los dominios de interés, tales como provincias, municipios o determinadas segmentaciones sociodemográficas. La inclusión de estos efectos aleatorios permite capturar la heterogeneidad existente entre los distintos grupos, lo que contribuye a mejorar la calidad de las estimaciones al reducir la varianza y proporcionar predicciones más estables y robustas.

Este marco teórico servirá de base para la posterior introducción de los modelos a nivel de área, en los cuales se incorporan estructuras jerárquicas que permiten combinar información procedente de diversas fuentes y niveles de agregación. A partir de esta formulación, será posible motivar la adopción de modelos específicos, como el modelo FH en su versión univariante, bivariante y multivariante, todos ellos diseñados para abordar las limitaciones inherentes a la estimación directa en dominios con tamaños muestrales reducidos.

3.1. Estimadores directos basados en el diseño del muestreo

El objetivo fundamental del muestreo aplicado a las encuestas es la obtención de información estadísticamente representativa sobre una población a partir de una selección parcial de sus unidades. Esto permite inferir con rigor diversas características de interés, tales como la media del número de des-

empleados a nivel autonómico, el porcentaje de pensionistas en un conjunto de municipios dentro de una provincia, o cualquier otra magnitud relevante. En el caso de los objetivos planteados por el INE se solicitó, entre otros, el estudio de las estimaciones de ocupados, parados e inactivos y la tasa de desempleo a nivel municipal, autonómico cruzado con la nacionalidad principal del país, al igual que un estudio de los «ninis» a nivel provincial bajo conformidad al respecto de las publicaciones oficiales del INE. No obstante, la aplicabilidad del muestreo no se limita exclusivamente a variables de uso frecuente en la estadística pública, sino que se extiende a ámbitos más específicos. Por ejemplo, es posible diseñar encuestas dirigidas a evaluar la necesidad de centros especializados para infantes con trastornos del espectro autista en determinadas localidades o a estimar la intención de voto hacia una determinada formación política en periodos electorales.

La razón fundamental que justifica el uso de técnicas de muestreo radica en la inviabilidad, tanto económica como logística, de realizar un censo exhaustivo sobre el conjunto de la población, especialmente cuando el dominio de estudio es amplio. Por ello, tanto instituciones públicas como organismos privados confían en la teoría del muestreo para el diseño de encuestas que permitan obtener información precisa con un coste reducido.

En este contexto, se utilizan estimadores directos basados en el diseño muestral subyacente. Entre los más conocidos se encuentran el estimador de Horvitz y Thompson (1952), el de Hansen y Hurwitz (1943) y, en el ámbito específico de la estimación en áreas pequeñas, el estimador de Hájek (1971). Estos estimadores presentan ventajas notables en términos de simplicidad y facilidad de implementación, dado que no requieren procedimientos computacionales complejos ni supuestos adicionales sobre la estructura de los datos. Sin embargo, presentan una limitación significativa: su coeficiente de variación suele ser elevado cuando el tamaño muestral es reducido, lo cual es habitual en la estimación para áreas pequeñas.

A pesar de estas limitaciones, los estimadores directos constituyen una base fundamental dentro del análisis estadístico y, en muchos casos, son empleados como referencia o criterio de conformidad, es decir, como *benchmarking*, frente a las estimaciones derivadas de estimadores provenientes de modelos que incorporan información externa, como los modelos de áreas pequeñas. En particular, los modelos a nivel de área, que serán abordados y definidos en detalle en secciones posteriores, permiten mejorar la eficiencia de las estimaciones mediante el uso de información auxiliar.

Para formalizar el estudio posterior, resulta necesario establecer ciertas definiciones clave dentro de la teoría del muestreo. En términos generales, una población finita puede concebirse como un conjunto de unidades diferenciadas, tales como individuos, empresas u hogares, entre otras entidades de interés. La teoría del muestreo tiene por objeto el diseño y selección de muestras, la observación de características sobre las unidades seleccionadas y la inferencia estadística sobre la población en su conjunto a partir de la información obtenida en la muestra.

Con estas definiciones establecidas, se procede a introducir la notación requerida para el estudio riguroso de los estimadores considerados. Dicha notación sigue el esquema clásico en áreas pequeñas, destacando Rao y Molina (2015), Morales et al. (2021), Esteban et al (2020), Bugallo et al (2024a) o Diz-Rosales et al (2024), entre muchos otros. Se denotará por $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ al vector de las variables de interés correspondientes a todas las unidades poblacionales, donde N representa el tamaño total de la población.

En este contexto, se considera el marco de un diseño de muestreo probabilístico, en el cual cada muestra s tiene una probabilidad de selección $p(s)$, siendo s un subconjunto de la población U, ergo $s \in U$. Bajo este enfoque, un estimador de la población de T se denotará por $\widehat{T}$, y, análogamente, se introducen las nociones de sesgo y varianza asociadas a dicho estimador bajo el diseño de muestreo, definidas como

$$\mathbb{B}_\pi(\widehat{T}) = \mathbb{E}_\pi[\widehat{T} - T] = \sum_{s \in U} p(s)(\widehat{T}(s) - T),$$

$$\mathbb{V}\,\mathrm{ar}_\pi(\widehat{T}) = \sum_{s \in U} p(s)(\widehat{T}(s) - T)^2.$$

Se define, a su vez, el error cuadrático medio como

$$\mathrm{MSE}_\pi(\widehat{T}) = \mathbb{B}_\pi(\widehat{T})^2 + \mathbb{V}\,\mathrm{ar}_\pi(\widehat{T}).$$

Cabe destacar que los operadores esperanza y varianza se presentan con el subíndice π, lo cual es una convención estándar en la literatura de estimación en áreas pequeñas, que denota que las probabilidades se calculan bajo el diseño. Así, dicha notación enfatiza que los cálculos se realizan bajo el diseño muestral y no bajo un modelo estadístico específico.

En términos generales, dentro de la teoría del muestreo, existen múltiples estrategias para la selección de muestras. Entre ellas, se incluyen el muestreo aleatorio simple, con o sin reemplazamiento, el muestreo sistemático, el muestreo por conglomerados y el muestreo polietápico, entre otros. Sin embargo, en la práctica, los estimadores directos más comúnmente utilizados en este ámbito, tales como el estimador de Horvitz y Thompson (HT) y el estimador de Hájek, suelen aplicarse bajo un esquema de muestreo aleatorio simple sin reemplazamiento (Rao y Molina 2015; Morales et al. 2021). Por otro lado, el estimador de Hansen y Hurwitz opera en el marco de un muestreo aleatorio simple con reemplazamiento, aunque su uso en la estimación en áreas pequeñas es menos frecuente (Rao y Molina 2015). Por esta razón, el presente estudio se enfocará en el caso de muestreo aleatorio simple sin reemplazamiento, y concretamente en los anteriores dos estimadores empleados comúnmente en este ámbito.

Bajo dicho diseño, la probabilidad de selección de una muestra s de tamaño n, extraída de la población U, viene dada por:

$$p(s) = \frac{1}{\binom{N}{n}}.$$

Asimismo, se definen las probabilidades de inclusión de primer y segundo orden de la siguiente manera:

$$\pi_i = \mathbb{P}(i \in s) = \sum_{s \in s(i)} p(s), \quad \pi_{ij} = \mathbb{P}(i \in s, j \in s) = \sum_{s \in s(i,j)} p(s),$$

donde $s(i)$ y $s(i,j)$ representan los subconjuntos muestrales que contienen la unidad i y las unidades i y j, respectivamente.

Para el caso particular del muestreo aleatorio simple sin reemplazamiento, estas probabilidades adoptan la forma:

$$\pi_i = \frac{n}{N}, \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)}, \quad \forall i, j \in U,\ i \neq j.$$

En el contexto de este trabajo, se emplea un esquema de muestreo aleatorio simple sin reemplazamiento dentro de cada dominio de interés, lo que se denotará mediante el subíndice $d = 1, \ldots, D$, con D representando el número total de dominios. En consecuencia, el problema se enmarca en un muestreo aleatorio simple estratificado, donde los estratos corresponden a los dominios y los tamaños muestrales $n_1, \ldots, n_D$ dentro de cada dominio son fijados por el propio diseño y arquitectura del muestreo y, así, de la encuesta. En este caso, las probabilidades de inclusión se expresan como:

$$\pi_i = \frac{n_d}{N_d}, \quad \pi_{ij} = \frac{n_d(n_d-1)}{N_d(N_d-1)}, \quad i \neq j.$$

Finalmente, se impone la propiedad de absorción, la cual establece que $\pi_{ii} = \pi_i$ (Särndal et al. 1992; Rao y Molina 2015).

3.1.1. Estimadores directos clásicos en áreas pequeñas: Horvitz-Thompson y Hájek

Remarcada la base matemática sobre la cual se moverá este apartado de las estimaciones directas en el contexto de la SAE, se procederá a definir los dos estimadores más comunes en la literatura de áreas pequeñas: el de HT y el de Hájek. Nuevamente, es preciso establecer una notación para la aclaración del lector antes de la propia definición y de las propiedades de estos estimadores directos. En esta sección, nos basaremos en ejemplares de la literatura básicos dentro de la estimación en áreas pequeñas, como es Morales et al. (2021), con adaptaciones a este trabajo particular.

- **Índices**: s denota una muestra, mientras que $d = 1, \ldots, D$ y $j = 1, \ldots, N$ representan, respectivamente, los dominios o áreas pequeñas y las unidades, generalmente llamados individuos en el contexto de áreas pequeñas.
- **Población y muestra**: La población se representa como $U = \bigcup_{d=1}^{D} U_d$ y la muestra como $s = \bigcup_{d=1}^{D} s_d$, donde U_d y s_d corresponden a la población y la muestra dentro del dominio d, respectivamente.
- **Tamaños**: N representa el tamaño poblacional y n el tamaño muestral. Cuando N y n llevan subíndices, estos denotan el tamaño correspondiente del conjunto indexado. Por ejemplo, N_d indica el tamaño de la población en el dominio d, mientras que n_d lo será para la muestra en dicho dominio.
- **Totales y Medias**: Y y X representan los totales poblacionales de las variables de interés y y x, respectivamente. Si Y y X llevan subíndices, estos indican los totales correspondientes al conjunto indexado. $\overline{Y}$ y $\overline{X}$ denotan las medias poblacionales de las variables anteriores.
- **Pesos muestrales**: Los pesos teóricos del diseño muestral se denotan como w_j. Estos corresponden al inverso de las probabilidades de inclusión definidas previamente, es decir,

$$w_j = \frac{1}{\pi_j}.$$

  Esta elección genera estimadores insesgados (Rao 2003), ya que verifican la condición de insesgadez de la forma $\sum_{j \in s} p(s) w_j(s) = 1$, $j = 1, \ldots, N$.

Así, Horvitz y Thompson (1952) introdujeron una familia de estimadores directos para el total poblacional Y_d y la media $\overline{Y}_d$ dentro de un dominio d. En particular, estos estimadores toman la forma:

$$\widehat{Y}_d^{\mathrm{HT}} = \sum_{j \in s_d} w_j y_j = \sum_{j \in s_d} \frac{y_j}{\pi_j},$$

$$\widehat{\overline{Y}}_d^{\mathrm{HT}} = \frac{\widehat{Y}_d^{\mathrm{HT}}}{N_d},$$

donde se asume que el total poblacional N_d es conocido, siendo esto último lo especialmente remarcable acerca de esta familia de estimadores. A continuación, se presentan algunas de las propiedades fundamentales de estos estimadores.

Proposición 1 *Sea $\pi_j > 0$ para todo $j \in U_d$. Entonces se cumplen las siguientes dos propiedades importantes sobre su media y varianza como estimador:*

1. *El estimador $\widehat{Y}_d^{HT}$ es insesgado para Y_d, es decir,*

$$\mathbb{E}_\pi \left[\widehat{Y}_d^{HT} \right] = Y_d.$$

2. *La varianza del estimador* $\widehat{Y}_d^{HT}$ *bajo el diseño muestral está dada por*

$$\mathbb{V}\,\mathrm{ar}_\pi\left(\widehat{Y}_d^{HT}\right) = \sum_{i\in U_d}\sum_{j\in U_d}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

donde un estimador insesgado de dicha varianza está dado por

$$\widehat{\mathbb{V}\,\mathrm{ar}}_\pi\left(\widehat{Y}_d^{HT}\right) = \sum_{i\in s_d}\sum_{j\in s_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

Análogamente, se puede definir las mismas propiedades para el estimador de la media, $\widehat{\overline{Y}}_d^{\mathrm{HT}}$.

Proposición 2 *Sea* $\pi_j > 0$ *para todo* $j \in U_d$. *Entonces se cumplen análogamente a la Proposición 1 las siguientes propiedades:*

1. *El estimador* $\widehat{\overline{Y}}_d^{HT}$ *es insesgado para* $\overline{Y}_d$, *es decir,*

$$\mathbb{E}_\pi\left[\widehat{\overline{Y}}_d^{HT}\right] = \overline{Y}_d.$$

2. *La varianza del estimador* $\widehat{\overline{Y}}_d^{HT}$ *bajo el diseño muestral está dada por*

$$\mathbb{V}\,\mathrm{ar}_\pi\left(\widehat{\overline{Y}}_d^{HT}\right) = \frac{1}{N_d^2}\sum_{i\in U_d}\sum_{j\in U_d}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j},$$

con un estimador insesgado de la varianza dado por

$$\widehat{\mathbb{V}\,\mathrm{ar}}_\pi\left(\widehat{\overline{Y}}_d^{HT}\right) = \frac{1}{N_d^2}\sum_{i\in s_d}\sum_{j\in s_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

Dado que ya se han deducido las expresiones de las probabilidades de inclusión para un muestreo aleatorio simple sin reemplazamiento en los dominios, es posible formular expresiones más concisas para las propiedades de los estimadores previamente presentadas.

En el caso de un muestreo aleatorio simple con reemplazamiento, la probabilidad de inclusión de segundo orden satisface la independencia, es decir, $\pi_{ij} = \pi_i\pi_j$. Sin embargo, en el caso sin reemplazamiento, esta relación solo es una aproximación, lo que permite obtener cotas superiores para la estimación de la varianza, así como para la covarianza de los estimadores del total y de la media. Esta desigualdad se cumple si N_d es lo suficientemente grande y Y_d no es demasiado cercano a cero. Así, se define la varianza del estimador del total de HT como

$$\mathbb{V}\,\mathrm{ar}_\pi\left(\widehat{Y}_d^{\mathrm{HT}}\right) = \sum_{j\in U_d}\frac{1-\pi_j}{\pi_j}y_j^2 = \sum_{j\in U_d}(w_j-1)y_j^2.$$

Alternativamente, cuando se computa sobre la muestra del dominio:

$$\widehat{\mathbb{V}\,\mathrm{ar}}_\pi\left(\widehat{Y}_d^{\mathrm{HT}}\right) = \sum_{j\in s_d}\frac{1-\pi_j}{\pi_j^2}y_j^2 = \sum_{j\in s_d}w_j(w_j-1)y_j^2.$$

Para el estimador de la media del dominio, se obtiene:

$$\mathbb{V}\,\mathrm{ar}_\pi\left(\widehat{\overline{Y}}_d^{\mathrm{HT}}\right) = \frac{1}{N_d^2}\sum_{j\in U_d}(w_j-1)y_j^2,$$

$$\widehat{\mathbb{V}\,\mathrm{ar}}_\pi\left(\widehat{\overline{Y}}_d^{\mathrm{HT}}\right) = \frac{1}{N_d^2}\sum_{j\in s_d} w_j(w_j-1)y_j^2.$$

Särndal et al. (1992) presentan la siguiente fórmula para la covarianza entre dos estimadores directos:

$$\mathbb{C}\,\mathrm{ov}_\pi\left(\widehat{Y}_d^{\mathrm{HT}}, \widehat{Z}_d^{\mathrm{HT}}\right) = \sum_{i\in U_d}\sum_{j\in U_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}y_i z_j.$$

Un estimador insesgado de la covarianza anterior es:

$$\widehat{\mathbb{C}\,\mathrm{ov}}_\pi\left(\widehat{Y}_d^{\mathrm{HT}}, \widehat{Z}_d^{\mathrm{HT}}\right) = \sum_{i\in s_d}\sum_{j\in s_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}y_i z_j,$$

y sabiendo las expresiones de las probabilidades de inclusión y usando la aproximación de independencia de la probabilidad de segundo orden anterior se llega a que

$$\widehat{\mathbb{C}\,\mathrm{ov}}_\pi\left(\widehat{\overline{Y}}_d^{\mathrm{HT}}, \widehat{\overline{Z}}_d^{\mathrm{HT}}\right) = \frac{1}{N_d^2}\sum_{j\in s_d}\frac{1-\pi_j}{\pi_j^2}y_j z_j = \frac{1}{N_d^2}\sum_{j\in s_d} w_j(w_j-1)y_j z_j.$$

Además del estimador HT, existe otro estimador de interés y ampliamente utilizado en la SAE es el conocido como estimador de Hájek. Hájek (1971) propuso los siguientes estimadores directos para la media y el total del dominio:

$$\widehat{\overline{Y}}_d^{\mathrm{H}} = \frac{\widehat{Y}_d^{\mathrm{HT}}}{\widehat{N}_d} = \frac{\sum_{j\in s_d} w_j y_j}{\sum_{j\in s_d} w_j},$$

$$\widehat{Y}_d^{\mathrm{H}} = N_d\widehat{\overline{Y}}_d^{\mathrm{H}}.$$

Estos estimadores poseen las propiedades mencionadas en la siguiente proposición.

Proposición 3 *Si n_d es suficientemente grande y $\pi_j > 0\ \forall j \in U_d$, entonces se verifica que*

(a) $\mathbb{E}_\pi[\widehat{Y}_d^H] \approx Y_d$.

(b) $\mathbb{V}\,\mathrm{ar}_\pi\left(\widehat{Y}_d^H\right) \approx \sum_{i\in U_d}\sum_{j\in U_d}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(y_i-\overline{Y}_d)(y_j-\overline{Y}_d)$.

Análogamente, se define la covarianza del estimador del total Hájek como

$$\widehat{\mathbb{C}\,\mathrm{ov}}_\pi\left(\widehat{Y}_d^H, \widehat{Z}_d^H\right) = \frac{N_d^2}{\widehat{N}_d^2}\sum_{j\in s_d} w_j(w_j-1)(y_j-\widehat{\overline{Y}}_d^H)(z_j-\widehat{\overline{Z}}_d^H). \tag{3.1}$$

Para obtener el estimador de la covarianza de la media, se procede a dividir la covarianza de la ecuación (3.1) por el total poblacional al cuadrado, N_d^2, en el dominio d. Por lo general, se suele trabajar más con las medias que con los totales en la SAE, especialmente en modelos a nivel de área (Morales et al. 2021), y será, pues, el estimador que se tratará de ahora en adelante salvo que se indique lo contrario.

Como el estimador directo es aproximadamente insesgado por la elección de los pesos, se demuestra que, por su propia definición, el error cuadrático medio, y su estimación, que será denotado por mse, son (Rao y Molina 2015):

$$\mathrm{MSE}(\widehat{\overline{Y}}_d^{\mathrm{H}}) \approx \mathbb{V}\,\mathrm{ar}_\pi(\widehat{\overline{Y}}_d^{\mathrm{H}}), \quad \mathrm{mse}(\widehat{\overline{Y}}_d^{\mathrm{H}}) = \widehat{\mathbb{V}\,\mathrm{ar}}_\pi(\widehat{\overline{Y}}_d^{\mathrm{H}}).$$

Dada la presentación de ambos estimadores, resulta pertinente analizar cuál de ellos es más adecuado en distintos contextos. En principio, ambos estimadores pueden ser empleados, si bien el estimador de Hájek suele preferirse debido a ciertas propiedades que mejoran aspectos específicos del estimador clásico de HT (Särndal et al. 1992).

En particular, el estimador de Hájek exhibe un comportamiento más favorable en términos de varianza cuando la diferencia $y_j - \overline{Y}_d$ es pequeña, lo cual ocurre con frecuencia en el análisis de proporciones, un caso recurrente en el estudio de datos composicionales. Asimismo, en escenarios donde se presentan variaciones significativas en el tamaño muestral, una situación habitual en la SAE, el estimador de Hájek tiende a proporcionar mayor estabilidad. Esta ventaja se debe a la sustitución del valor poblacional N_d por su correspondiente estimador $\widehat{N}_d$, lo que permite mitigar las fluctuaciones derivadas de tamaños muestrales heterogéneos.

Para finalizar este apartado, mencionar que existen magnitudes derivadas de estas estimaciones directas que pueden ser de utilidad, como por ejemplo es, en el contexto de la estadística aplicada a variables laborales, la tasa de desempleo o tasa de paro.

Esta se define como el ratio entre el total de parados respecto al total de la población activa, es decir, parados y ocupados, por dominio, y que será denotada por R_d, definido como

$$R_d = \frac{Y_d}{Y_d + Z_d},$$

es decir, se trata de un estimador de tipo ratio de otras dos variables estimadas. Mencionar que se pueden emplear también las medias para el siguiente desarrollo y cálculo de la tasa de desempleo, pero aquí se expondrá la definición usual.

A fin de obtener una aproximación del error cuadrático medio de $\widehat{R}_d$, recurrimos a la técnica de linearización de Taylor aplicado a funciones diferenciables de variables aleatorias.

$$\widehat{R}_d \triangleq \widehat{R}_d^H = \frac{\widehat{Y}_d^H}{\widehat{Y}_d^H + \widehat{Z}_d^H},$$

con su error estimado

$$\mathrm{mse}_\pi(\widehat{R}_d^H) = \frac{(\widehat{Z}_d^H)^2}{(\widehat{Y}_d^H + \widehat{Z}_d^H)^4}\mathrm{mse}_\pi(\widehat{Y}_d^H) + \frac{(\widehat{Y}_d^H)^2}{(\widehat{Y}_d^H + \widehat{Z}_d^H)^4}\mathrm{mse}_\pi(\widehat{Z}_d^H) - 2\frac{\widehat{Z}_d^H\widehat{Y}_d^H}{(\widehat{Y}_d^H + \widehat{Z}_d^H)^4}\widehat{\mathbb{C}\mathrm{ov}}_\pi(\widehat{Z}_d^H, \widehat{Z}_d^H). \tag{3.2}$$

Para demostrar la obtención esta expresión, se emplea una linearización en serie de Taylor como se indicó. Así pues, sea $f : \mathbb{R}^2 \to \mathbb{R}$ la función definida como

$$f(y, z) = \frac{y}{y + z},$$

cuyos argumentos representan las estimaciones del total de desempleados (y) y de ocupados (z). Considerando una expansión de primer orden de $f(y, z)$ alrededor del punto $(y_0, z_0) = (Y_d, Z_d)$, donde cada coordenada representa el valor poblacional verdadero, tenemos:

$$f(y, z) \approx f(y_0, z_0) + \frac{\partial f}{\partial y}(y - y_0) + \frac{\partial f}{\partial z}(z - z_0).$$

Dado que $f(y, z)$ representa la tasa de desempleo, se obtiene la aproximación:

$$\widehat{R}_d \approx R_d + \frac{Z_d(\widehat{Y}_d - Y_d)}{(Y_d + Z_d)^2} - \frac{Y_d(\widehat{Z}_d - Z_d)}{(Y_d + Z_d)^2}.$$

Tomando esperanzas bajo la medida del diseño de muestreo π, resulta:

$$\mathbb{E}_\pi[\widehat{R}_d] \approx R_d + Z_d \frac{\mathbb{B}_\pi[\widehat{Y}_d]}{(Y_d+Z_d)^2} - Y_d \frac{\mathbb{B}_\pi[\widehat{Z}_d]}{(Y_d+Z_d)^2}.$$

Si los estimadores $\widehat{Y}_d$ y $\widehat{Z}_d$ son insesgados, como es al considerar estimadores de tipo Hájek, es decir, si $\mathbb{B}_\pi[\widehat{Y}_d] = 0$ y $\mathbb{B}_\pi[\widehat{Z}_d] = 0$, se concluye que, por propia construcción y de la esperanza anterior:

$$\mathbb{E}_\pi[\widehat{R}_d] \approx R_d.$$

Finalmente, para aproximar el error cuadrático medio de $\widehat{R}_d$, se eleva al cuadrado el término $\widehat{R}_d - R_d$ obtenido de la expansión de Taylor y se toma la esperanza, lo que lleva a la expresión:

$$\mathrm{MSE}_\pi(\widehat{R}_d) \approx \frac{Z_d^2\,\mathrm{MSE}_\pi(\widehat{Y}_d)}{(Y_d+Z_d)^4} + \frac{Y_d^2\,\mathrm{MSE}_\pi(\widehat{Z}_d)}{(Y_d+Z_d)^4} - \frac{2Y_dZ_d\mathbb{C}\,\mathrm{ov}_\pi(\widehat{Y}_d,\widehat{Z}_d)}{(Y_d+Z_d)^4}.$$

Dicha expresión puede ser estimada, usando estimadores de Hájek, mediante la expresión (3.2). Por lo tanto, se demuestra que la estimación del error cuadrático medio de la tasa de desempleo puede obtenerse por linearización de Taylor, validando así las estimaciones propuestas.

3.2. Diseño Muestral en la Encuesta de Población Activa

Se considera un país con una estructura geográfica jerárquica, en el cual el territorio se divide en estratos, dominios y conglomerados (López-Vizcaíno 2014; Morales et al. 2021). Se analiza la Encuesta de Población Activa (EPA) con un diseño muestral bietápico y estratificación en la primera etapa. En esta, las unidades de la primera etapa son los conglomerados (en este caso las secciones censales, siguiendo la nomenclatura del INE), los cuales se seleccionan sin reemplazo dentro de cada estrato, con probabilidades proporcionales al número de viviendas. En la segunda etapa, las unidades de selección son las viviendas. Dentro de cada conglomerado seleccionado en la primera etapa, se elige un número fijo de viviendas, que denotaremos por $\mathcal{V}$, mediante muestreo aleatorio simple sin reemplazamiento. Todas las personas que residen en las viviendas seleccionadas son entrevistadas.

Se introduce la siguiente notación en analogía a la anterior expuesta en la subsección previa:

- **Subíndices**: h para los estratos, a para los conglomerados, v para las viviendas y j para los individuos.
- **Tamaños poblacionales y tamaños muestrales**: V_{ha} y V_h representan el número de viviendas en el conglomerado a del estrato h y en el estrato h, respectivamente. De la misma forma, m_h indica el número de conglomerados muestreados en el estrato h, con $h = 1, \ldots, H$.
- **Muestra y submuestras**: La muestra total se denota por $s = \bigcup_{h=1}^{H} s_h$, donde cada s_h representa la submuestra del estrato h. A su vez, cada submuestra estratificada se descompone como $s_h = \bigcup_{a=1}^{m_h} s_{ha}$, donde s_{ha} denota la submuestra del conglomerado a en el estrato h.

La probabilidad de selección del conglomerado a en el estrato h en la primera etapa se aproxima suponiendo muestreo con reemplazo. Definiendo X_{ha} como el número de veces que el conglomerado a del estrato h aparece en la muestra, se obtiene:

$$\mathbb{P}(\mathrm{congl}_{ha}) = 1 - P(X_{ha} = 0) = 1 - \left(1 - \frac{V_{ha}}{V_h}\right)^{m_h} \approx m_h \frac{V_{ha}}{V_h},$$

siendo válida esta aproximación cuando m_h es pequeño en comparación con el número total de conglomerados en el estrato h. Bajo este supuesto, la probabilidad de que un conglomerado sea seleccionado más de una vez es despreciable.

Si un conglomerado a en el estrato h es seleccionado en la primera etapa, la probabilidad de selección de una vivienda v dentro de dicho conglomerado mediante muestreo aleatorio simple sin reemplazamiento es:

$$\mathbb{P}(\text{vivienda}_{hav} \mid \text{congl}_{ha}) = \frac{\mathcal{V}}{V_{ha}}.$$

Por lo tanto, la probabilidad de selección de una vivienda v en el conglomerado a del estrato h es:

$$\mathbb{P}(\text{vivienda}_{hav}) = \mathbb{P}(\text{congl}_{ha})\mathbb{P}(\text{vivienda}_{hav} \mid \text{congl}_{ha}) \approx \mathcal{V}m_h \frac{V_{ha}}{V_h}\frac{1}{V_{ha}} = \frac{\mathcal{V}m_h}{V_h}.$$

Dado que las probabilidades de inclusión de los individuos en una misma vivienda son equivalentes, la probabilidad de selección de un individuo j perteneciente a la vivienda v del conglomerado a en el estrato h es:

$$\pi_j = P(j \in s_h) = P(\text{vivienda}_{hav}) = \frac{\mathcal{V}m_h}{V_h}.$$

Esto implica que la muestra es autoponderada dentro de cada estrato. Recordar que todas las personas que residen en las viviendas seleccionadas son entrevistadas, de tal forma que $\pi_j \triangleq \pi_h$. Invirtiendo las probabilidades de inclusión π_h, se obtienen los pesos teóricos de la muestra:

$$w_j = \frac{1}{\pi_j} = \frac{V_h}{\mathcal{V}m_h} = w_h, \quad j \in s_h.$$

En este tipo de encuestas, estos pesos se corrigen debido a la no respuesta. En adelante, se denotará por w_j a los pesos corregidos por no respuesta. El estimador no calibrado del total de una variable y en la EPA es un estimador de tipo Hájek de la forma:

$$\widehat{Y}^{\text{EPA}^\dagger} = \sum_{h=1}^{H} N_h \frac{1}{\widehat{N}_h} \sum_{j \in s_h} w_j y_j,$$

donde:

- N_h representa el número total de individuos en el estrato h.
- $\widehat{N}_h = \sum_{j \in s_h} w_j = w_h n_h$, con n_h el número de individuos muestreados en el estrato h.

Este estimador es un estimador de razón postestratificado con pesos muestrales corregidos, donde los grupos de postestratificación corresponden a los estratos. Para un estudio más en detalle de la estructuración de la EPA, se recomienda la lectura del Capítulo 2 de la tesis de López-Vizcaíno (2014), donde además se detallan las definiciones rigurosas de estados de ocupación, desempleo e inactividad, y que serán usadas en la aplicación a datos reales.

3.2.1. Conformidad

Sea el estimador no calibrado:

$$\widehat{Y}^{\text{EPA}^\dagger} = \sum_{j \in s} w_j^b y_j,$$

y sean K variables auxiliares con totales poblacionales conocidos:

$$X_k = \sum_{j \in U} x_{jk}, \quad k = 1, \ldots, K.$$

El problema de calibración consiste en determinar pesos calibrados w_j^c minimizando, para una función ϕ que sea divergencia de Csiszár, donde ϕ es una función convexa definida en $[0,\infty)$ y que satisface $\phi(1)=\phi'(1)=0$ (Csiszár 1963):

$$\sum_{j\in s} w_j^b \phi\left(\frac{w_j^c}{w_j^b}\right),$$

sujeto a:

$$\sum_{j\in s} w_j^c x_{jk} = X_k, \quad k=1,\ldots,K.$$

Si se elige $\phi(x)=\frac{(x-1)^2}{2}$, la solución es:

$$w_j^c = w_j^b\left(1-\sum_{k=1}^{K}\lambda_k x_{jk}\right), \quad j\in s,$$

donde los coeficientes λ_k se obtienen resolviendo el sistema lineal correspondiente.

Sea $\widehat{Y}^{EPA}$ el estimador del total Y de la variable y en la EPA. Este estimador se expresa como:

$$\widehat{Y}^{EPA} = \sum_{j\in s} w_j^c y_j,$$

donde w_j^c son los pesos calibrados.

En el caso de dominios de estudio, el estimador EPA del total Y_d en el dominio d es:

$$\widehat{Y}_d^{EPA} = \sum_{j\in s_d} w_j^c y_j.$$

Dado que la población se puede descomponer en la unión disjunta de los dominios $U=\bigcup_{d=1}^{D} U_d$, se cumple la propiedad de conformidad (*benchmarking*):

$$\widehat{Y}^{EPA} = \sum_{d\in U} \widehat{Y}_d^{EPA}.$$

Tomando, pues, $\widehat{Y}^{EPA}$ como el estimador EPA del total Y, y sea $\widehat{Y}_1,\ldots,\widehat{Y}_D$ un conjunto de estimadores para los totales en los dominios $Y_1,\ldots,Y_D$, en general, estos no satisfacen la propiedad de *benchmarking*:

$$\widehat{Y}^{EPA} = \sum_{d=1}^{D} \widehat{Y}_d.$$

Para forzar el cumplimiento de esta propiedad, se definen los nuevos estimadores calibrados o conformes:

$$\widehat{Y}_d^c = \lambda_y \widehat{Y}_d, \quad \text{donde} \quad \lambda_y = \frac{\widehat{Y}^{EPA}}{\sum_{d=1}^{D}\widehat{Y}_d},$$

y, de este modo, se garantiza que:

$$\widehat{Y}^{EPA} = \sum_{d=1}^{D} \widehat{Y}_d^c.$$

Para la estimación de la varianza y covarianza de los estimadores ajustados, se suelen emplear las siguientes aproximaciones, siempre y cuando los pesos del *benchmarking*, los valores λ_x y λ_y, sean próximos a la unidad:

$$\mathbb{V}\,\mathrm{ar}_\pi(\widehat{Y}_d^c) \approx \lambda_y^2 \mathbb{V}\,\mathrm{ar}_\pi(\widehat{Y}_d),$$
$$\mathbb{C}\,\mathrm{ov}_\pi(\widehat{Y}_d^c, \widehat{Z}_d^c) \approx \lambda_y \lambda_z \mathbb{C}\,\mathrm{ov}_\pi(\widehat{Y}_d, \widehat{Z}_d).$$

Si bien es posible recurrir a aproximaciones para incorporar los pesos de *benchmarking*, cuando estos se desvían de manera significativa de la unidad, suele preferirse integrarlos directamente en el procedimiento de *bootstrap*. De este modo, las varianzas y covarianzas se estiman dentro del propio esquema de remuestreo, lo que permite prescindir de aproximaciones adicionales y obtener resultados más coherentes con la estructura del modelo ajustado.

3.3. Introducción a los Modelos Mixtos

El estudio de los modelos mixtos es esencial para comprender la SAE. En este campo, se emplean frecuentemente modelos lineales y sus extensiones generalizadas, ya que proporcionan un marco teórico robusto para abordar los problemas de estimación en áreas pequeñas (Jiang y Nguyen 2007).

Los modelos lineales están diseñados para analizar variables aleatorias independientes obtenidas de una misma población. En contraste, los modelos mixtos presentan una estructura jerárquica o multinivel, donde las observaciones son independientes entre distintos niveles o conglomerados, pero dependientes dentro de cada uno de ellos debido a características compartidas. Este tipo de datos introduce dos fuentes principales de variabilidad: la variabilidad entre conglomerados y la variabilidad dentro de cada conglomerado. Los modelos mixtos permiten modelar estas estructuras complejas, lo que los convierte en herramientas fundamentales para el análisis de datos reales (Demidenko 2013).

Los modelos lineales mixtos (LMMs) amplían el enfoque de los modelos lineales tradicionales al incorporar la posibilidad de modelar correlaciones dentro de los datos. A diferencia de los modelos lineales, los LMMs pueden manejar errores correlacionados, lo que los hace especialmente útiles en escenarios donde las observaciones no son completamente independientes. Además, estos modelos permiten la inclusión de efectos aleatorios y factores jerárquicos, así como la consideración de estructuras de correlación espacial y temporal (Morales et al. 2021).

Una de las aplicaciones más relevantes de los LMMs es en la SAE, donde su capacidad para integrar múltiples fuentes de información y modelar distintos componentes de error resulta crucial. Estos modelos no solo permiten mejorar la precisión de las estimaciones al vincular todas las observaciones de la muestra, sino que también pueden capturar la heterogeneidad entre áreas. Dentro del ámbito de la SAE, se distinguen principalmente dos tipos de modelos: los modelos a nivel de área y los modelos a nivel de unidad. Fay y Herriot (1979) introdujeron un modelo a nivel de área en los Estados Unidos para estimar el ingreso per cápita en regiones pequeñas, mientras que Battese et al. (1988) aplicaron un modelo a nivel de unidad en la estimación de superficies de cultivo por condado.

Antes de introducirse en los modelos de la SAE se realiza una breve introducción a los modelos lineales mixtos, ya que muchos de los conceptos que se manejan provienen de este campo de la regresión estadística.

3.3.1. Estimación en los modelos lineales mixtos

Sea el siguiente modelo lineal mixto:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{3.3}$$

donde $\boldsymbol{y} \in \mathbb{R}^n$ representa el vector de observaciones, $\boldsymbol{\beta} \in \mathbb{R}^p$ es el vector de efectos fijos, y $\boldsymbol{u} \in \mathbb{R}^q$ es el vector de efectos aleatorios. Las matrices de diseño $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ y $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$ asocian los efectos fijos y aleatorios, respectivamente, con las observaciones. Finalmente, $\boldsymbol{e} \in \mathbb{R}^n$ es el vector de errores aleatorios del modelo.

Se asume que los efectos aleatorios y los errores de muestreo son estadísticamente independientes y siguen distribuciones normales con media cero (Demidenko 2013), es decir:

$$\begin{aligned} \mathbb{E}[\boldsymbol{u}] &= \mathbf{0}, & \mathbb{E}[\boldsymbol{e}] &= \mathbf{0}, \\ \mathbb{V}\,\mathrm{ar}(\boldsymbol{u}) &= \boldsymbol{V}_u, & \mathbb{V}\,\mathrm{ar}(\boldsymbol{e}) &= \boldsymbol{V}_e, \end{aligned}$$

donde $\boldsymbol{V}_u$ y $\boldsymbol{V}_e$ son matrices de varianza-covarianza que dependen de un conjunto de parámetros, los cuales representan los componentes de varianza del modelo.

A partir de la ecuación (3.3), se obtiene la expresión de la varianza del vector de observaciones:

$$\boldsymbol{V} = \mathbb{V}\,\mathrm{ar}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{V}_u\boldsymbol{Z}^\top + \boldsymbol{V}_e,$$

donde se supone que la matriz $\boldsymbol{V}$ es no singular, garantizando la existencia de la inversa y permitiendo la estimación de los parámetros del modelo.

Se asume, al inicio, que las componentes de varianza del modelo (3.3) son conocidas. El término aleatorio del modelo es $\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$, con varianza dada por:

$$\mathbb{V}\,\mathrm{ar}(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}) = \boldsymbol{Z}\boldsymbol{V}_u\boldsymbol{Z}^\top + \boldsymbol{V}_e = \boldsymbol{V}.$$

Para transformar el modelo y obtener términos aleatorios incorrelacionados con varianza unitaria, multiplicamos por $\boldsymbol{V}^{-1/2}$:

$$\boldsymbol{V}^{-1/2}\boldsymbol{y} = \boldsymbol{V}^{-1/2}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}^{-1/2}(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}).$$

Definiendo $\boldsymbol{y}^* = \boldsymbol{V}^{-1/2}\boldsymbol{y}$, $\boldsymbol{e}^* = \boldsymbol{V}^{-1/2}(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e})$ y $\boldsymbol{X}^* = \boldsymbol{V}^{-1/2}\boldsymbol{X}$, el modelo transformado queda expresado como:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{e}^*,$$

donde la varianza de $\boldsymbol{e}^*$ se reduce a la identidad:

$$\mathbb{V}\,\mathrm{ar}(\boldsymbol{e}^*) = \boldsymbol{V}^{-1/2}\boldsymbol{V}\boldsymbol{V}^{-1/2} = \boldsymbol{I}_n.$$

En este contexto, podemos aplicar el método de mínimos cuadrados ordinarios para estimar $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{e}^{*\top}\boldsymbol{e}^*.$$

Bajo la suposición de normalidad, el estimador $\widehat{\boldsymbol{\beta}}$ también coincide con el estimador de máxima verosimilitud (MLE, por sus siglas en inglés) de $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}.$$

Por lo general, en la SAE se le denomina a este estimador como el mejor estimador lineal insesgado, o BLUE por sus siglas en inglés.

Corolario 1 *Bajo el modelo (3.3), el mejor predictor lineal insesgado (BLUP) de* $\boldsymbol{u}$ *está dado por:*

$$\widehat{\boldsymbol{u}} = \boldsymbol{V}_u\boldsymbol{Z}^\top\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \tag{3.4}$$

Además, se cumple que:

$$\widehat{\boldsymbol{u}} = \mathbb{E}_{\widehat{\boldsymbol{\beta}}}[\boldsymbol{u} \mid \boldsymbol{y}].$$

El estimador (3.4) tiene las siguientes propiedades fundamentales:

1. Es óptimo en el sentido de minimizar la esperanza condicional:

$$\mathbb{E}[(\widehat{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{A}(\widehat{\boldsymbol{u}} - \boldsymbol{u})],$$

   para cualquier matriz $\boldsymbol{A}$ definida positiva.

2. Es lineal respecto a $\boldsymbol{y}$ e insesgado.

Dado que $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_e)$ y $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_u)$, y bajo $\boldsymbol{e} \perp \boldsymbol{u}$, la función de densidad conjunta de $\boldsymbol{y}$ y $\boldsymbol{u}$ se expresa como:

$$f(\boldsymbol{y}, \boldsymbol{u}) = f(\boldsymbol{y} \mid \boldsymbol{u}) f(\boldsymbol{u}),$$

donde:

$$f(\boldsymbol{y} \mid \boldsymbol{u}) = c \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^\top \boldsymbol{V}_e^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\right),$$

$$f(\boldsymbol{u}) = \exp\left(-\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{V}_u^{-1}\boldsymbol{u}\right),$$

y c es una constante positiva.

Proposición 4 *Se cumple que* $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}$ *y* $\widehat{\boldsymbol{u}} = \widetilde{\boldsymbol{u}}$*, donde:*

$$(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{u}}) = \arg\max_{\boldsymbol{\beta}, \boldsymbol{u}} f(\boldsymbol{y}, \boldsymbol{u}),$$

cuyo resultado se deriva de los sistemas de ecuaciones presentados por Henderson (1953, 1975).

Ahora bien, al trasladar estos resultados a la práctica surge inmediatamente la dificultad de que las componentes de varianza-covarianza de los efectos aleatorios no están disponibles de antemano, de modo que no existe una expresión cerrada para su estimación conjunta con los efectos fijos. Para salvar esta carencia, se adopta un procedimiento iterativo cuyo objetivo es maximizar numéricamente la función de verosimilitud del modelo. En cada paso se parte de un valor inicial para los componentes de varianza (por ejemplo, obtenido por momentos o a partir de estimaciones de modelos más sencillos), se evalúa la verosimilitud y sus derivadas respecto a dichos componentes y, mediante el algoritmo de Fisher-Scoring, se combina la información del gradiente (score) con la matriz de información esperada o de Fisher para obtener un desplazamiento óptimo en el espacio de parámetros. Este desplazamiento corrige simultáneamente las estimaciones de todas las varianzas de los efectos aleatorios, de manera que cada iteración acerca la solución al máximo de la verosimilitud.

El mismo esquema puede implementarse en dos variantes principales, según se utilice la función de verosimilitud completa o la restringida. En el método ML se trabaja directamente con la probabilidad conjunta de todas las observaciones, mientras que en REML se construye la verosimilitud sobre contrastes ortogonales a los efectos fijos para reducir el sesgo en los componentes de varianza. Una vez satisfecho el criterio de convergencia, las últimas actualizaciones se toman como estimaciones óptimas de las varianzas y se procede a recomputar los efectos fijos y aleatorios condicionados a ellas, cerrando así el ciclo de estimación. Esto, de todas formas, se tratará en la sección 3.3.2 con más detalle.
Así, consideremos el modelo lineal mixto:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \cdots + \boldsymbol{Z}_m\boldsymbol{u}_m + \boldsymbol{e}, \tag{3.5}$$

donde:

- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ es el vector de observaciones muestrales.
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ es el vector de efectos fijos.
- $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iq_i})^\top$ representa los efectos aleatorios asociados al i-ésimo factor aleatorio, con $i = 1, \ldots, m$.

- $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$ es el vector de errores muestrales.
- $\boldsymbol{X}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ son matrices de diseño de dimensiones $n \times p$, $n \times q_1$, $\ldots$, $n \times q_m$, respectivamente.

Para reformular el modelo (3.5) en la forma general de un modelo mixto lineal de la ecuación (3.3), definimos:

$$\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m), \quad \boldsymbol{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_m^\top)^\top, \quad q = \sum_{i=1}^{m} q_i.$$

De esta manera, el modelo se reescribe como:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

manteniendo la estructura estándar de un modelo lineal mixto.

A este modelo se le ha de presuponer una serie de condiciones para que los parámetros sean estimables. Para ver dichas condiciones en detalle, se recomienda la lectura de Morales et al. (2015). Cuando sea necesario, se enfatizará la dependencia de $\boldsymbol{V}$ con respecto a $\boldsymbol{\sigma} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2)^\top$ escribiendo explícitamente $\boldsymbol{V}(\boldsymbol{\sigma})$. Se define $M = p + m + 1$ y considera el vector de parámetros desconocidos como

$$\boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{\sigma}^\top),$$

donde $\boldsymbol{\beta} \in \mathbb{R}^p$ y $\boldsymbol{\sigma} \in \mathbb{R}^{m+1}$. Así, el espacio paramétrico se define como

$$\Theta = \left\{ \boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{\sigma}^\top) \in \mathbb{R}^M \,:\, \boldsymbol{\beta} \in \mathbb{R}^p,\, \sigma_0^2 > 0,\, \sigma_i^2 \geq 0 \text{ para } i = 1, \ldots, m \right\}.$$

Este conjunto define las restricciones naturales sobre los componentes de $\boldsymbol{\sigma}$, asegurando la positividad de la varianza residual y la no negatividad de las varianzas de los efectos aleatorios. Estos parámetros del modelo se estimarán con los métodos que se explicarán en la Sección 3.3.2.

3.3.2. Aplicación del algoritmo de Fisher-Scoring para las estimaciones por máxima verosimilitud y máxima verosimilitud restringida.

En el modelo lineal mixto (3.5), la función de densidad conjunta de $\boldsymbol{y}$, dado θ, sigue una distribución normal multivariante:

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) = (2\pi)^{-n/2} |\boldsymbol{V}|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right).$$

De forma resumida, el estimador de máxima verosimilitud (ML) $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ se obtiene maximizando la función de log-verosimilitud:

$$l(\boldsymbol{\theta}) = \log f_{\boldsymbol{\theta}}(\boldsymbol{y}).$$

Las ecuaciones del *scoring*, obtenidas al igualar a cero el gradiente de $l(\boldsymbol{\theta})$, son:

$$S_{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0,$$

$$S_{\sigma_i^2} = -\frac{1}{2} \operatorname{tr}\left(\boldsymbol{V}^{-1} \frac{\partial \boldsymbol{V}}{\partial \sigma_i^2} \right) + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1} \frac{\partial \boldsymbol{V}}{\partial \sigma_i^2} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0, \quad i = 0, \ldots, m.$$

Dado que estas ecuaciones no tienen solución analítica cerrada, se utilizan métodos iterativos como Newton-Raphson o el mencionado Fisher-Scoring:

$$\widehat{\boldsymbol{\theta}}^{(r+1)} = \widehat{\boldsymbol{\theta}}^{(r)} - \boldsymbol{H}(\widehat{\boldsymbol{\theta}}^{(r)})^{-1} \boldsymbol{S}(\widehat{\boldsymbol{\theta}}^{(r)}),$$

donde $H(\theta)$ es la matriz hessiana de $l(\theta)$, o en el caso del Fisher-Scoring, tomando esperanzas de la expresión anterior y cambiando el signo:

$$\widehat{\boldsymbol{\theta}}^{(\boldsymbol{r+1})} = \widehat{\boldsymbol{\theta}}^{(\boldsymbol{r})} + \boldsymbol{F}(\widehat{\boldsymbol{\theta}}^{(\boldsymbol{r})})^{-\mathbf{1}}\boldsymbol{S}(\widehat{\boldsymbol{\theta}}^{(\boldsymbol{r})}),$$

donde $F(\theta)$ es la matriz de información de Fisher.

Bajo condiciones de regularidad, el estimador de máxima verosimilitud $\widehat{\boldsymbol{\theta}}$ es asintóticamente normal y consistente (Jiang 2007):

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \boldsymbol{F}^{-\mathbf{1}}_{\boldsymbol{\beta\beta}}(\boldsymbol{\sigma})\right),$$

$$\widehat{\boldsymbol{\sigma}} \sim \mathcal{N}_m\left(\boldsymbol{\sigma}, \boldsymbol{F}^{-\mathbf{1}}_{\boldsymbol{\sigma\sigma}}(\boldsymbol{\sigma})\right),$$

donde

$$\boldsymbol{F}_{\boldsymbol{\beta\beta}}(\boldsymbol{\sigma}) = (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}),$$

y

$$\boldsymbol{F}_{\sigma_i^2\sigma_j^2}(\boldsymbol{\sigma}) = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \sigma_i^2}\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \sigma_j^2}\right)$$

es la matriz por bloques y las entradas de la relacionada con los errores aleatorios de la matriz de información de Fisher para cada parámetro de $\boldsymbol{\theta}$. Recordar que, por definición, estos elementos de la matriz de Fisher se calculan, en general, como

$$F_{ab} = -\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_a \partial\theta_b}\right],$$

y las expresiones previas provienen de la aplicación de esta definición para el caso particular que atañe a este apartado.

Siempre que es posible, en la SAE se tiende a emplear la estimación de máxima verosimilitud residual (REML), propuesto por Patterson y Thompson (1971), la cual se introduce para reducir el sesgo de los estimadores de máxima verosimilitud de los componentes de varianza, ya que en estadística oficial la insesgadez prima sobre la varianza. Para ello, se transforma el vector $\boldsymbol{y}$ en dos vectores independientes $\boldsymbol{y_1} = \boldsymbol{K_1 y}$ y $\boldsymbol{y_2} = \boldsymbol{K_2 y}$, con la condición de que la distribución de $\boldsymbol{y_1}$ no dependa del parámetro de regresión fija $\boldsymbol{\beta}$. Sea $\boldsymbol{K_1}$ una matriz tal que $\boldsymbol{K_1 X} = \mathbf{0}$. Por lo tanto,

$$\mathbb{E}[\boldsymbol{y_1}] = \mathbb{E}[\boldsymbol{K_1 y}] = \mathbb{E}[\boldsymbol{K_1}(\boldsymbol{X\beta} + \boldsymbol{Z_1 u_1} + \cdots + \boldsymbol{Z_m u_m} + \boldsymbol{e})] = 0.$$

El vector $\boldsymbol{y_2}$ se selecciona de manera que sea independiente de $\boldsymbol{y_1}$, por lo que satisface

$$\mathbb{E}[\boldsymbol{y_1 y_2}^\top] = \boldsymbol{K_1}\mathbb{E}[\boldsymbol{y y}^\top]\boldsymbol{K_2}^\top = \boldsymbol{K_1 V K_2}^\top = 0.$$

Las filas $\boldsymbol{k}^\top$ de la matriz $\boldsymbol{K_1}$ se denominan contrastes, ya que cumplen con $\boldsymbol{k}^\top\boldsymbol{X} = 0$. El número máximo de contrastes linealmente independientes es $n - \mathrm{r}(\boldsymbol{X})$. Suponemos que $\boldsymbol{X}$ tiene rango completo p, por lo que la matriz $\boldsymbol{K_1}$ puede seleccionarse de manera que su rango sea $n - p$. La matriz $\boldsymbol{K_2}$ se selecciona con rango p.

Para introducir la matriz $\boldsymbol{K_1}$, consideramos el modelo sin efectos aleatorios

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \text{con } \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma_\varepsilon}),$$

siendo $\boldsymbol{\Sigma_\varepsilon}$ la matriz de varianza-covarianza de los errores conocida. El MLE de $\boldsymbol{\beta}$ es

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Sigma_\varepsilon}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\Sigma_\varepsilon}^{-1}\boldsymbol{y}.$$

Se define, pues, el vector transformado como un residuo estandarizado de la forma

$$\boldsymbol{y_1} = \boldsymbol{\Sigma_\varepsilon}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) = \boldsymbol{\Sigma_\varepsilon}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Sigma_\varepsilon}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma_\varepsilon}^{-1}\boldsymbol{y}\right) = \boldsymbol{K_1}\boldsymbol{y},$$

donde

$$\boldsymbol{K_1} = \boldsymbol{\Sigma_\varepsilon}^{-1} - \boldsymbol{\Sigma_\varepsilon}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Sigma_\varepsilon}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma_\varepsilon}^{-1}.$$

Además, se selecciona $\boldsymbol{K_2} = \boldsymbol{X}^\top\boldsymbol{V}^{-1}$, y como $\boldsymbol{K_1} = \boldsymbol{K_1}^\top$, se cumple que

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y_1}] &= \mathbb{E}[\boldsymbol{K_1}\boldsymbol{y}] = \boldsymbol{K_1}\boldsymbol{X}\boldsymbol{\beta} = 0,\\
\mathbb{E}[\boldsymbol{y_2}] &= \mathbb{E}[\boldsymbol{K_2}\boldsymbol{y}] = \boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta},\\
\mathrm{Var}(\boldsymbol{y_1}) &= \boldsymbol{K_1}\boldsymbol{V}\boldsymbol{K_1}^\top,\\
\mathrm{Var}(\boldsymbol{y_2}) &= \boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X},\\
\mathbb{E}[\boldsymbol{y_1}\boldsymbol{y_2}^\top] &= \boldsymbol{K_1}\boldsymbol{V}\boldsymbol{K_2}^\top = 0.
\end{aligned}$$

Dado que el número máximo de columnas linealmente independientes en $\boldsymbol{K_1}$ es $n - \mathrm{r}(\boldsymbol{X})$, podemos seleccionar $n - \mathrm{r}(\boldsymbol{X})$ de estas columnas para construir una submatriz $\boldsymbol{K}$ de dimensión $n \times (n - \mathrm{r}(\boldsymbol{X}))$ que satisfaga $\boldsymbol{K}^\top\boldsymbol{X} = 0$. Se redefine el vector y se toma el anterior vector $\boldsymbol{y_1} = \boldsymbol{K}^\top\boldsymbol{y}$ y $\boldsymbol{y_2}$. Como $\mathrm{r}(\boldsymbol{X}) = p$, se tiene que

$$\begin{aligned}
\boldsymbol{y_1} &\sim \mathcal{N}_{n-p}(0, \boldsymbol{K}^\top\boldsymbol{V}\boldsymbol{K}),\\
\boldsymbol{y_2} &\sim \mathcal{N}_p(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}),
\end{aligned}$$

y son independientes. Sea $\boldsymbol{\sigma} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2)^\top$ y $\boldsymbol{P} = \boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{K}^\top$. La función de verosimilitud de $\boldsymbol{y_1}$ es

$$l(\boldsymbol{\sigma}) = -\frac{n-p}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{K}^\top\boldsymbol{V}\boldsymbol{K}| - \frac{1}{2}\boldsymbol{y_1}^\top(\boldsymbol{K}^\top\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{y_1}.$$

La matriz de información de Fisher $\boldsymbol{F(\sigma)}$ se obtiene como

$$\boldsymbol{F}_{\sigma_i^2\sigma_j^2} = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\sigma_j^2}\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\sigma_i^2}\right). \tag{3.6}$$

mientras que el gradiente se escribe como

$$S_{\sigma_i^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\sigma_i^2}\right) + \frac{1}{2}\left(\boldsymbol{y}^\top\boldsymbol{P}\frac{\partial\boldsymbol{V}}{\partial\sigma_i^2}\boldsymbol{P}\boldsymbol{y}\right)$$

Para calcular los estimadores REML, el método de Fisher-Scoring usa la iteración

$$\boldsymbol{\sigma}^{(r+1)} = \boldsymbol{\sigma}^{(r)} + \boldsymbol{F}^{-1}(\boldsymbol{\sigma}^{(r)})\boldsymbol{S}(\boldsymbol{\sigma}^{(r)}),$$

donde $\boldsymbol{F}(\boldsymbol{\sigma}^{(r)})$ es la matriz de información de Fisher evaluada en $\boldsymbol{\sigma}^{(r)}$. Finalmente, el estimador de máxima verosimilitud residual de $\boldsymbol{\beta}$ es

$$\widehat{\boldsymbol{\beta}}_{REML} = (\boldsymbol{X}^\top\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y}.$$

Nuevamente, se demuestra que, bajo condiciones de regularidad, el estimador de máxima verosimilitud $\widehat{\boldsymbol{\theta}}$ es asintóticamente normal y consistente (Jiang 2007):

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, \boldsymbol{F}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-\mathbf{1}}(\boldsymbol{\sigma})\right),$$

$$\widehat{\boldsymbol{\sigma}} \sim \mathcal{N}_{m+1}\left(\boldsymbol{\sigma}, \boldsymbol{F}^{-\mathbf{1}}_{\boldsymbol{\sigma}\boldsymbol{\sigma}}(\boldsymbol{\sigma})\right),$$

donde la matriz de varianza-covarianza de $\widehat{\boldsymbol{\sigma}}$ se define bajo las componentes de (3.6) y $F_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\sigma})$ es la misma que en método de máxima verosimilitud. Se observa que $F(\boldsymbol{\sigma})$ es una matriz de tamaño $(m+1)\times(m+1)$; sin embargo, la matriz de información de Fisher necesaria para calcular los estimadores de máxima verosimilitud, $F(\boldsymbol{\theta})$, es de tamaño $(p+m+1)\times(p+m+1)$, resultando así también en una ventaja computacional respecto al basado en la ML.

Así, con todos los elementos descritos y necesarios para el estudio, se procederá, pues, en el siguiente capítulo a introducir los modelos a nivel de área, y más concretamente los modelos de tipo Fay-Herriot, objetivo del trabajo.

Capítulo 4

Modelos Fay-Herriot

El modelo de Fay-Herriot es una herramienta clave en la SAE, especialmente cuando los modelos a nivel unidad presentan restricciones en la disponibilidad de datos auxiliares. Se recuerda que uno de los problemas fundamentales de los modelos a nivel de individuo, en particular, los EBLUPs de parámetros lineales, es que estos requieren datos auxiliares agregados, lo que limita la cantidad de variables disponibles y puede generar discrepancias entre los registros administrativos y los datos muestrales. Para la estimación de parámetros no lineales, se necesitaría un censo con las variables auxiliares, lo que supone una barrera práctica para muchas oficinas estadísticas.

A pesar de la pérdida de información inherente a la agregación de datos, el modelo de Fay-Herriot ofrece ventajas sustanciales. En primer lugar, permite incorporar un mayor número de variables auxiliares en la estimación. En segundo lugar, evita la restricción de los estimadores a nivel unidad, que exigen la coincidencia exacta de las variables auxiliares en la muestra de encuesta y en los registros administrativos externos. Finalmente, el modelo ha mostrado un buen desempeño en escenarios donde el número de áreas pequeñas es grande y los tamaños muestrales son reducidos, lo que lo convierte en una alternativa robusta a los modelos a nivel unidad en estas circunstancias (Jiang y Lahiri 2006). No solo eso, sino que es más sencilla la obtención de datos agregados que a nivel de individuo, desagregados.

El modelo de Fay-Herriot, como se hubo mencionado en la parte relativa al estado del arte, propuesto originalmente para estimar ingresos per cápita en pequeñas áreas de EE.UU. (Fay y Herriot 1979), plantea un modelo mixto a nivel de área con efectos aleatorios. Su principal ventaja es que permite el uso de un mayor número de variables auxiliares sin la restricción de coincidencia exacta con los datos administrativos. Sin embargo, la agregación de datos conlleva pérdida de información respecto a los modelos a nivel unidad.

El modelo ha sido ampliamente estudiado y extendido en diversas direcciones. En particular, la estimación del MSE ha sido abordada por autores como Prasad y Rao (1990), Datta y Lahiri (2000) y González-Manteiga et al. (2010), entre otros. Asimismo, se han desarrollado metodologías para la construcción de intervalos de confianza y pruebas de hipótesis (Molina et al. 2015; Marhuenda et al. 2016) y se ha analizado el impacto de la estimación de parámetros en la eficiencia del EBLUP (Jiang y Tang, 2011).

Entre las extensiones del modelo, destacan los enfoques bayesianos paramétricos y semiparamétricos (Polettini, 2017), así como las metodologías para abordar errores en las variables auxiliares (Ybarra y Lohr, 2008; Datta et al. 2018). Además, se han desarrollado criterios de selección de modelos basados en variantes del criterio de información de Akaike (Marhuenda et al. 2014; Lombardía et al. 2017).

4.1. Modelo Univariante Fay-Herriot

Para introducir el enfoque de modelos a nivel de área, consideramos el siguiente modelo lineal mixto general:

$$\boldsymbol{y}_{D\times 1} = \boldsymbol{X}_{D\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{Z}_{D\times q}\boldsymbol{u}_{q\times 1} + \boldsymbol{e}_{D\times 1},$$

donde D denota el número total de dominios, $\boldsymbol{X}$ y $\boldsymbol{Z}$ son matrices de diseño conocidas, $\boldsymbol{\beta}$ es el vector de efectos fijos, y $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_u)$ representa el vector de efectos aleatorios. Los errores del modelo están dados por $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_e)$, y se supone que $\boldsymbol{u}$ y $\boldsymbol{e}$ son independientes.

Las matrices de varianza $\boldsymbol{V}_u$ y $\boldsymbol{V}_e$ se definen en función de parámetros de varianza desconocidos, típicamente σ_u^2 y σ_d^2, que deben ser estimados a partir de los datos. Para el desarrollo teórico, se consideran conocidas ambas matrices. No obstante, en la práctica, $\boldsymbol{V}_u$ es desconocida y, por ende, es preciso estimar sus componentes. En el caso de $\boldsymbol{V}_e$, se considera, por hipótesis de los modelos de área, conocida, pero en la práctica se estima a partir de la expresión de la evaluación de covarianzas de Hájek (ec. 3.1).

Condicionalmente a los efectos aleatorios $\boldsymbol{u}$, la variable respuesta tiene esperanza:

$$\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{u}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},$$

y varianza:

$$\mathbb{V}\mathrm{ar}(\boldsymbol{y} \mid \boldsymbol{u}) = \boldsymbol{V}_e.$$

Por tanto, marginalmente (es decir, integrando sobre la distribución de $\boldsymbol{u}$), se obtiene:

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}, \qquad \mathbb{V}\mathrm{ar}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{V}_u\boldsymbol{Z}^\top + \boldsymbol{V}_e.$$

Además, la matriz de varianza de los efectos aleatorios es:

$$\mathbb{V}\mathrm{ar}(\boldsymbol{u}) = \boldsymbol{V}_u,$$

y la covarianza cruzada entre $\boldsymbol{y}$ y $\boldsymbol{u}$ es:

$$\mathbb{C}\mathrm{ov}(\boldsymbol{y}, \boldsymbol{u}) = \mathbb{E}\left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\boldsymbol{u}^\top\right] = \boldsymbol{Z}\boldsymbol{V}_u.$$

Este planteamiento constituye la base del modelo a nivel de área en el contexto de estimación en áreas pequeñas, que será introducido a continuación, en el que se busca mejorar la precisión de los estimadores directos incorporando información auxiliar a través de los efectos aleatorios.

4.1.1. Teorema de Predicción y Modelo de Fay-Herriot

Dada la formulación previamente establecida, es conveniente emplear un teorema ampliamente utilizado en el contexto de la inferencia estadística en poblaciones finitas, conocido como Teorema de Predicción de Henderson.

El Teorema de Predicción de Henderson se formula con el propósito de proporcionar el BLUP de cualquier combinación lineal de efectos fijos y aleatorios bajo un modelo mixto. En esencia, lo que persigue es encontrar, para un parámetro de interés que dependa a la vez de los coeficientes de regresión (efectos fijos) y de los componentes aleatorios, la forma óptima de combinarlos de modo que el error cuadrático medio de la predicción quede minimizado dentro de la familia de predictores lineales e insesgados.

En el ámbito de la SAE, este teorema adquiere un valor central: cualquier estimador de un parámetro de dominio (una media, un total, una razón, etc.) puede expresarse como una combinación de una parte "global" explicada por covariables agregadas y un término "local" que recoge la variabilidad no explicada a nivel de área. Gracias al Teorema de Henderson, basta con identificar los vectores que definen esa combinación para obtener de forma sistemática el predictor óptimo y su error de predicción asociado. Este teorema es, pues, ampliamente empleado en la SAE por su versatilidad, resumiendo varios de los artículos de Henderson, entre los que destaca Henderson (1959).

Teorema 1 *Sea el parámetro de interés de la forma*

$$\tau = \boldsymbol{l}^\top \boldsymbol{\beta} + \boldsymbol{m}^\top \boldsymbol{u}.$$

donde $\boldsymbol{l}$*, y* $\boldsymbol{m}$ *son vectores que acompañan a* $\boldsymbol{\beta}$ *y* $\boldsymbol{u}$ *relacionados con la información auxiliar en el caso de* $\boldsymbol{l}$ *y con* $\boldsymbol{Z}$ *en el caso de* $\boldsymbol{m}$*. El BLUP*[1]*, de* τ*, está dado por*

$$\widetilde{\tau} = \boldsymbol{l}^\top \widetilde{\boldsymbol{\beta}} + \boldsymbol{m}^\top \widetilde{\boldsymbol{u}},$$

donde

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{y},$$
$$\widetilde{\boldsymbol{u}} = \boldsymbol{V}_u \boldsymbol{Z}^\top \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}).$$

La varianza del error del predictor $\widetilde{\tau}$ *está dada por*

$$\mathbb{V}\,\mathrm{ar}(\widetilde{\tau} - \tau) = \boldsymbol{m}^\top \boldsymbol{T} \boldsymbol{m} + (\boldsymbol{l}^\top - \boldsymbol{m}^\top \boldsymbol{T} \boldsymbol{Z}^\top \boldsymbol{V}_e^{-1} \boldsymbol{X}) \boldsymbol{Q} (\boldsymbol{l}^\top - \boldsymbol{m}^\top \boldsymbol{T} \boldsymbol{Z}^\top \boldsymbol{V}_e^{-1} \boldsymbol{X})^\top,$$

donde

$$\boldsymbol{Q} = (\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}, \quad \boldsymbol{T} = \boldsymbol{V}_u - \boldsymbol{V}_u \boldsymbol{Z}^\top \boldsymbol{V}^{-1} \boldsymbol{Z} \boldsymbol{V}_u.$$

Este resultado (Henderson 1959) establece la estructura del predictor lineal óptimo, en particular, dentro del modelo de Fay-Herriot, y será el empleado por defecto durante el trabajo. De todas formas, las conclusiones de la obtención de estos resultados se pueden conseguir a partir de los métodos previamente descritos de la máxima verosimilitud, pero aquí se expondrá este conveniente teorema ya que es la perspectiva más empleada dentro de la rama teórica descriptiva de la SAE.

El modelo FH univariante combina estimaciones directas con variables auxiliares para mejorar la precisión de las estimaciones, siendo un modelo bietápico que considera tanto el nivel de muestreo como el nivel de área. Así, se establece la siguiente notación, análoga a la ya presentada en apartados anteriores.

Sea $\mu_d = \overline{Y}_d$ la característica de interés, en este caso la media de una variable aleatoria, en el área d, donde $d = 1, \ldots, D$. El estimador directo de μ_d se denota por $y_d = \widehat{\overline{Y}}^{\mathrm{H}}_d$, el cual tiende a ser la media muestral estimada por Hájek por las razones explicadas en el Capítulo 3. Sean N_d y n_d el tamaño poblacional y muestral, respectivamente. Se recuerda que la media del dominio se define como:

$$\overline{Y}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} y_{di},$$

donde y_{di} es la variable objetivo medida a nivel de unidad. El estimador de $\overline{Y}_d$, donde se empleará el estimador de Hájek, se expresa como:

$$y_d = \widehat{\overline{Y}}^{\mathrm{H}}_d = \frac{1}{\widehat{N}_d} \sum_{i=1}^{n_d} w_{di} y_{di}\,.$$

En la mayoría de las referencias, el estimador se denota simplemente por y_d en vez de emplear $\overline{y}_d$ por simplicidad de notación, y será, pues, la que se emplee en este trabajo.

El modelo de Fay-Herriot, en general, se estructura en dos niveles. En el primer nivel, denominado modelo muestral, se asume que los estimadores directos y_d son insesgados y pueden expresarse como:

$$y_d = \mu_d + e_d, \quad d = 1, \ldots, D,$$

[1]Emplearemos, por simplicidad y conveniencia en la notación, la tilde ($\widetilde{\ }$) para los BLUP y BLUE y el acento circunflejo ($\widehat{\ }$) para sus versiones empíricas, en consonancia con lo que se suele escribir en la literatura en la SAE.

donde e_d son errores de muestreo independientes y distribuidos normalmente:

$$e_d \sim \mathcal{N}(0, \sigma_d^2).$$

La varianza σ_d^2 se asume conocida. En el segundo nivel, denominado modelo de conexión o de enlace, la verdadera característica del área μ_d se relaciona con un vector de p covariables $\boldsymbol{x}_d = (x_{d1}, \ldots, x_{dp})$ mediante:

$$\mu_d = \boldsymbol{x}_d\boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D,$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes de regresión y u_d son errores del modelo, independientes y distribuidos normalmente:

$$u_d \sim \mathcal{N}(0, \sigma_u^2).$$

Los parámetros $\boldsymbol{\beta}$ y σ_u^2 son desconocidos y se estiman a partir de los datos disponibles. Así, con lo anterior expuesto, el modelo Fay-Herriot puede expresarse en forma matricial como:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

donde $\boldsymbol{y} = (y_1, \ldots, y_D)^\top$ es el vector de observaciones, $\boldsymbol{X} = \text{col}(\boldsymbol{x}_d)$ es la matriz de diseño, donde «col» es un operador que apila por columnas la matriz en cuestión, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ es el vector de coeficientes, $\boldsymbol{Z} = \boldsymbol{I}_D$ es la matriz identidad de orden D, $\boldsymbol{u} = (u_1, \ldots, u_D)^\top$ es el vector de efectos aleatorios, y $\boldsymbol{e} = (e_1, \ldots, e_D)^\top$ es el vector de errores de muestreo. Las matrices de varianzas se definen como:

$$\boldsymbol{V_u} = \mathbb{V}\,\text{ar}(\boldsymbol{u}) = \sigma_u^2\boldsymbol{I_D}, \quad \boldsymbol{V_e} = \mathbb{V}\,\text{ar}(\boldsymbol{e}) = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2), \quad \boldsymbol{V} = \mathbb{V}\,\text{ar}(\boldsymbol{y}) = \boldsymbol{V_u} + \boldsymbol{V_e}.$$

4.1.2. Estimación de los parámetros del modelo de Fay-Herriot

El BLUP de μ_d se obtiene mediante:

$$\widetilde{\mu}_d = \boldsymbol{x}_d\widetilde{\boldsymbol{\beta}} + \widetilde{u}_d,$$

donde:

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{y}, \quad \widetilde{\boldsymbol{u}} = \boldsymbol{V_u}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}).$$

El BLUP de u_d, para el caso univariante, suele ser expresado como:

$$\widetilde{u}_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2}(y_d - \boldsymbol{x}_d\widetilde{\boldsymbol{\beta}}).$$

que no es más que una media ponderada entre el estimador directo y el estimador del modelo (Rao y Molina 2015; Morales et al. 2021).

Cuando la varianza de los efectos aleatorios σ_u^2 es desconocida, resulta imprescindible estimarla antes de construir el predictor empírico óptimo. Tradicionalmente, esta estimación se aborda mediante métodos de momentos, entre los que destacan los procedimientos basados en la descomposición de los residuos propuestos por Prasad y Rao (1990), aunque puede dar estimaciones negativas, y el estimador III de Henderson (1975), que extrae las componentes de varianza de la descomposición total del modelo, garantizando insesgadez aunque con ligera tendencia a sobreestimar sus valores (Dogan et al. 2014).

En un marco más contemporáneo, bajo la hipótesis de normalidad conjunta de datos y efectos aleatorios, se recurre a los métodos ML y REML, los cuales estiman de forma simultánea los parámetros de efectos fijos y componentes de varianza mediante la maximización de la función de verosimilitud o de la verosimilitud restringida, respectivamente. La implementación habitual de estos métodos se basa en el algoritmo de Fisher-Scoring, cuya convergencia y propiedades asintóticas han sido detalladas en

la Sección 3.3.2. En particular, la modalidad REML se prefiere en el contexto de estimación para áreas pequeñas por su menor sesgo en muestras reducidas, como muestran Rao y Molina (2015) y Thompson y Patterson (1971).

El estimador empírico del mejor predictor lineal insesgado (EBLUE) de $\boldsymbol{\beta}$ se define como:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{y},$$

donde se escribe por simplicidad $\widehat{\boldsymbol{V}} = \text{diag}(\widehat{\sigma}_u^2 + \sigma_d^2)$.

El objeto de inferencia en este contexto es la estimación del valor medio μ_d en cada dominio d, definida como la esperanza del valor observado y_d, es decir, $\mu_d = \mathbb{E}[y_d]$. Bajo el modelo lineal mixto a nivel de área, específicamente, bajo el modelo Fay-Herriot, el estimador más comúnmente utilizado es el EBLUP, que se obtiene sustituyendo los parámetros desconocidos por sus estimaciones en el BLUP, como se demostrará a continuación.

Así, el EBLUP de μ_d se expresa como:

$$\widehat{\mu}_d = \boldsymbol{x}_d \widehat{\boldsymbol{\beta}} + \widehat{u}_d \tag{4.1}$$

Este estimador combina de forma óptima la información directa proveniente de la muestra (y_d) con la información auxiliar ($\boldsymbol{x}_d$) a través del modelo, ponderando según la precisión relativa de ambas fuentes. La forma empírica de este predictor lo convierte en la herramienta principal para la estimación en dominios pequeños, especialmente cuando los tamaños muestrales son reducidos.

Para llegar a la ecuación (4.1), partimos de la definición del BLUP teórico

$$\widetilde{\mu}_d \;=\; \boldsymbol{x}_d \widetilde{\boldsymbol{\beta}} \;+\; \widetilde{u}_d,$$

y del hecho de que

$$\widetilde{u}_d \;=\; \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2} \left(y_d - \boldsymbol{x}_d \widetilde{\boldsymbol{\beta}}\right).$$

Sustituyendo en $\widetilde{\mu}_d$ obtenemos

$$\widetilde{\mu}_d = \boldsymbol{x}_d \widetilde{\boldsymbol{\beta}} \;+\; \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2} \left(y_d - \boldsymbol{x}_d \widetilde{\boldsymbol{\beta}}\right) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2}\, y_d \;+\; \left(1 - \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2}\right) \boldsymbol{x}_d \widetilde{\boldsymbol{\beta}},$$

y, así, llegamos a la forma convexa

$$\widetilde{\mu}_d = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_d^2}\, y_d \;+\; \frac{\sigma_d^2}{\sigma_u^2 + \sigma_d^2}\, \boldsymbol{x}_d \widetilde{\boldsymbol{\beta}}.$$

En la práctica, basta reemplazar σ_u^2 por su estimador $\widehat{\sigma}_u^2$ y $\widetilde{\boldsymbol{\beta}}$ por $\widehat{\boldsymbol{\beta}}$, obteniendo

$$\widehat{\mu}_d = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \sigma_d^2}\, y_d \;+\; \frac{\sigma_d^2}{\widehat{\sigma}_u^2 + \sigma_d^2}\, \boldsymbol{x}_d \widehat{\boldsymbol{\beta}}.$$

El EBLUP de μ_d se utiliza como estimador de la media del dominio $\overline{Y}_d$. Como ya se dijo, al tratarse de una media ponderada, se pueden ver dos casos en particular. Si el estimador directo y_d es preciso (n_d es grande), entonces $\sigma_d^2 \approx 0$ y $\sigma_d^2 \ll \sigma_u^2$, por lo que $\widehat{\mu}_d \approx y_d = \widehat{\overline{Y}}_d^{\,\text{H}}$. Por el contrario, si el estimador directo y_d no es preciso (n_d es pequeño), entonces $\sigma_d^2 \gg 0$ y $\sigma_d^2 \gg \sigma_u^2$, por lo que $\widehat{\mu}_d \approx \boldsymbol{x}_d \widehat{\boldsymbol{\beta}}$. El concepto de grande o pequeño es un debate dentro de la SAE en cuanto al tamaño muestral, pero se suele considerar pequeño cuando el tamaño muestral, n_d, está por debajo de 20 o 30 (Morales et al 2021, Rao 2003).

La estimación conjunta de los parámetros $\boldsymbol{\beta}$ y σ_u^2 en el modelo de Fay-Herriot se realiza bajo la hipótesis de normalidad (Fay y Herriot 1979), utilizando métodos ML y REML. Ambos métodos permiten obtener estimadores de los parámetros de manera simultánea, aunque presentan características y propiedades distintas.

Como ya se trató en el Capítulo 3, el método de máxima verosimilitud se basa en la maximización de la función de verosimilitud, la cual, para el caso del modelo FH univariante, se expresa como:

$$l_{\mathrm{ML}}(\sigma_u^2, \boldsymbol{\beta}) = -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}),$$

Los estimadores obtenidos mediante este método son consistentes y asintóticamente normales, con distribuciones dadas por:

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}, (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right), \quad \widehat{\sigma}_u^2 \sim \mathcal{N}\left(\sigma_u^2, \boldsymbol{F}^{-1}_{\sigma_u^2\sigma_u^2}\right),$$

donde $\boldsymbol{F}$ es la matriz de información de Fisher. Sin embargo, los estimadores ML pueden estar sesgados en muestras pequeñas, lo que motiva el uso del método REML.

El método REML, para el modelo FH univariante, se basa en la función de verosimilitud de la forma:

$$l_{\mathrm{REML}}(\sigma_u^2, \boldsymbol{\beta}) = -\frac{D-p}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{X}^\top\boldsymbol{X}\right| - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log\left|\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\right| - \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{P}\boldsymbol{y},$$

donde $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}$. Los estimadores REML conservan las propiedades de consistencia y normalidad asintótica de los estimadores ML, pero además son insesgados en muestras pequeñas, lo que los convierte en una opción preferible en la práctica.

A pesar de las ventajas de los métodos ML y REML, la hipótesis de normalidad en los errores del modelo puede ser difícil de sostener en aplicaciones con datos reales, especialmente en áreas con tamaños muestrales pequeños. Para abordar este problema, se ha demostrado que los estimadores REML mantienen su consistencia y normalidad asintótica incluso en situaciones donde no se cumple la normalidad. Además, trabajos como los de Jiang (1996, 1997, 1998) han establecido que el MSE derivado por Prasad y Rao (1990) es robusto bajo la no normalidad de μ_d.

Con ello, los métodos ML y REML proporcionan herramientas robustas para la estimación de los parámetros en el modelo de Fay-Herriot, incluso en situaciones donde la hipótesis de normalidad no se cumple estrictamente. Estas propiedades hacen que los estimadores REML sean ampliamente utilizados en la práctica, por encima de los estimadores ML por la garantía de insesgadez, como se mencionó previamente.

4.1.3. Inferencia en el modelo Fay-Herriot

Gracias a la normalidad asintótica de los estimadores REML, es posible realizar la construcción de intervalos de confianza y la realización de tests de hipótesis (Morales et al. 2021). Los intervalos de confianza asintóticos para los componentes de $\boldsymbol{\beta}$ y σ_u^2 se construyen como:

$$\widehat{\sigma}_u^2 \pm z_{\alpha/2}\sqrt{v}, \quad \widehat{\beta}_i \pm z_{\alpha/2}\sqrt{q},$$

donde v y q son las varianzas estimadas de $\widehat{\sigma}_u^2$ y $\widehat{\boldsymbol{\beta}}_i$, respectivamente, de la forma $F^{-1}(\widehat{\sigma}_u^2) = v$ y $\left(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} = q$, y $z_{\alpha/2}$ es el cuantil que deja a la derecha de la cola una probabilidad de $\alpha/2$ en la distribución normal estándar.

Para contrastar hipótesis sobre los coeficientes $\boldsymbol{\beta}$, se calcula el p-valor asintótico como:

$$p\text{-valor} = 2\mathbb{P}_{H_0}\left(\widehat{\beta}_i > |\beta_0|\right) = 2\mathbb{P}\left(\mathcal{N}(0,1) > \frac{|\beta_0|}{\sqrt{q}}\right),$$

donde la hipótesis nula consiste en la suposición de la nulidad de $\widehat{\beta}_i$.

4.1.4. Evaluación del MSE del estimador $\widehat{\mu}_d$ basado en el modelo Fay-Herriot

La evaluación del desempeño del modelo de Fay-Herriot se basa principalmente en dos medidas de precisión: el MSE y el CV. Aunque el MSE es una medida estándar en la teoría estadística, en el ámbito de la estadística oficial el CV resulta especialmente relevante.

El coeficiente de variación se define como:

$$\mathrm{CV}(\mu_d) \approx \mathrm{RRMSE}(\mu_d) = \frac{\mathrm{RMSE}(\mu_d)}{|\mu_d|} = \frac{\sqrt{\mathrm{MSE}(\mu_d)}}{|\mu_d|}. \tag{4.2}$$

En la práctica, dado que ni μ_d ni su MSE son conocidos, se utiliza la versión estimada del CV para el estimador de interés, denotada por $\widehat{\mathrm{CV}}(\widehat{\mu}_d)$, dada por:

$$\widehat{\mathrm{CV}}(\widehat{\mu}_d) = \frac{\sqrt{\widehat{\mathrm{MSE}}(\widehat{\mu}_d)}}{|\widehat{\mu}_d|}, \tag{4.3}$$

donde $\widehat{\mu}_d$ es el estimador del valor medio del dominio (en este caso, aunque puede ser del total o de la proporción también), y $\widehat{\mathrm{MSE}}(\widehat{\mu}_d)$ es una estimación del MSE correspondiente.

Este indicador permite comparar la precisión de las estimaciones no sólo entre dominios, sino también entre distintos tipos de estimadores (por ejemplo, estimadores directos versus estimadores modelados), independientemente de la escala de la variable. Por ello, el CV es particularmente útil para evaluar la eficiencia relativa de las estimaciones en áreas pequeñas.

4.1.4.1. MSE analítico con σ_u^2 conocida

Bajo la suposición de que la varianza de los efectos aleatorios σ_u^2 es conocida, el MSE del predictor BLUP teórico $\widetilde{\mu}_d = \boldsymbol{x}_d\widetilde{\boldsymbol{\beta}} + \widetilde{u}_d$ se descompone, siguiendo Prasad y Rao (1990), en dos componentes:

$$\mathrm{MSE}\big(\widetilde{\mu}_d\big) \;=\; g_{1d}(\sigma_u^2) \;+\; g_{2d}(\sigma_u^2)\,,$$

donde

$$g_{1d}(\sigma_u^2) = \frac{\sigma_u^2\,\sigma_d^2}{\sigma_u^2+\sigma_d^2}, \qquad g_{2d}(\sigma_u^2) = \Big(\tfrac{\sigma_d^2}{\sigma_u^2+\sigma_d^2}\Big)^2\, \boldsymbol{x}_d\big(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{x}_d^\top.$$

Aquí, g_{1d} mide la contribución al error de predicción debida al componente aleatorio u_d y al error de muestreo e_d, reflejando la varianza residual de la combinación lineal. En cambio, g_{2d} recoge la incertidumbre en la estimación de los parámetros fijos $\boldsymbol{\beta}$, ya que depende de la matriz de diseño $\boldsymbol{X}$ y de la matriz generalizada de varianza $\boldsymbol{V} = \mathrm{diag}(\sigma_u^2+\sigma_d^2)$.

En particular, si σ_d^2 es pequeño frente a σ_u^2, el peso de la parte muestral decrece y g_{2d} domina la varianza; inversamente, para $\sigma_d^2 \gg \sigma_u^2$, $g_{1d} \approx \sigma_u^2$ se convierte en la principal fuente de error (Prasad y Rao, 1990).

4.1.4.2. MSE del EBLUP con σ_u^2 estimada

En la práctica, σ_u^2 se desconoce y debe estimarse, por ejemplo, vía REML o ML, lo que añade una nueva fuente de variabilidad. Datta y Lahiri (2000) demuestran que el MSE del predictor empírico $\widehat{\mu}_d = \boldsymbol{x}_d\widehat{\boldsymbol{\beta}} + \widehat{u}_d$ incorpora un término g_{3d}:

$$\mathrm{MSE}\big(\widehat{\mu}_d\big) = g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2) + g_{3d}(\sigma_u^2) + o(D^{-1}),$$

siendo

$$g_{3d}(\sigma_u^2) = \frac{\sigma_d^4}{(\sigma_u^2+\sigma_d^2)^3}\,\mathrm{avar}\big(\widehat{\sigma}_u^2\big), \qquad \mathrm{avar}\big(\widehat{\sigma}_u^2\big) = 2\left(\sum_{j=1}^{D}\frac{1}{(\sigma_u^2+\sigma_j^2)^2}\right)^{-1}.$$

Este término corrige la variabilidad introducida al reemplazar σ_u^2 por su estimador $\widehat{\sigma}_u^2$, y resulta especialmente relevante cuando el número de áreas D es moderado (Datta y Lahiri 2000). Es importante tener en cuenta que este nuevo término depende del método de estimación (Datta y Lahiri 2000)

4.1.4.3. Estimador plug-in

El estimador clásico del MSE conocido como "Prasad-Rao corregido" se obtiene sustituyendo σ_u^2 por $\widehat{\sigma}_u^2$ y duplicando la contribución de g_{3d} para compensar el sesgo de orden $o(D^{-1})$:

$$\widehat{\mathrm{MSE}}(\widehat{\mu}_d) = g_{1d}(\widehat{\sigma}_u^2) + g_{2d}(\widehat{\sigma}_u^2) + 2\, g_{3d}(\widehat{\sigma}_u^2).$$

Prasad y Rao (1990) prueban que $\mathbb{E}[\, g_{1d}(\widehat{\sigma}_u^2)\,] \approx g_{1d}(\sigma_u^2) - g_{3d}(\sigma_u^2)$, por lo que el factor 2 corrige con precisión el sesgo introducido.

Cuando el tamaño muestral n_d crece y $\sigma_d^2 \ll \sigma_u^2$:

$$g_{1d}(\sigma_u^2) \approx \sigma_d^2, \quad g_{2d}(\sigma_u^2) \approx 0, \quad g_{3d}(\sigma_u^2) \approx 0,$$

y por tanto $\widehat{\mathrm{MSE}}(\widehat{\mu}_d) \approx \sigma_d^2$, recuperando la varianza del estimador directo. Esta propiedad garantiza coherencia con la inferencia clásica en dominios con muestra abundante (Rao y Molina 2015).

4.1.4.4. Método de remuestreo *bootstrap* para estimar el MSE

Cuando el MSE analítico no está disponible en algunos casos, por ejemplo, en presencia de transformaciones no lineales o en modelos multivariantes, se recurre al *bootstrap* paramétrico (González-Manteiga et al. 2008, 2010). El procedimiento para estimar el MSE de los EBLUPs $\widehat{\mu}_d$ en el modelo de Fay-Herriot es el siguiente:

Paso 1. Ajuste inicial. Ajustar el modelo de Fay-Herriot a los datos observados $\{(y_d, \boldsymbol{x}_d)\}_{d=1}^{D}$ y obtener los estimadores $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_u^2$.

Paso 2. Generación de efectos aleatorios *bootstrap*. Para cada réplica $b = 1, \ldots, B$:

$$u_d^{*(b)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, \widehat{\sigma}_u^2\big), \quad d = 1, \ldots, D,$$

y definir la "verdadera" media en el mundo bootstrap

$$\mu_d^{*(b)} = \boldsymbol{x}_d\, \widehat{\boldsymbol{\beta}} + u_d^{*(b)}.$$

Paso 3. Generación del error de muestreo *bootstrap*. Para cada d, simular

$$e_d^{*(b)} \sim \mathcal{N}\big(0, \sigma_d^2\big),$$

donde σ_d^2 es la varianza de diseño conocida de y_d, y construir

$$y_d^{*(b)} = \mu_d^{*(b)} + e_d^{*(b)}.$$

Paso 4. Reajuste en cada réplica. Ajustar de nuevo el modelo de Fay-Herriot sobre $\{y_d^{*(b)}, \boldsymbol{x}_d\}$ para obtener $\widehat{\boldsymbol{\beta}}^{*(b)}$, y $\widehat{\sigma}_u^{2\,(b)}$ y calcular el EBLUP *bootstrap*

$$\widehat{\mu}_d^{*(b)} = \frac{\widehat{\sigma}_u^{2\,(b)}}{\widehat{\sigma}_u^{2\,(b)} + \sigma_d^2}\, y_d^{*(b)} \;+\; \frac{\sigma_d^2}{\widehat{\sigma}_u^{2\,(b)} + \sigma_d^2}\, \boldsymbol{x}_d\, \widehat{\boldsymbol{\beta}}^{*(b)}.$$

Paso 5. Estimación del MSE[2]. El estimador *bootstrap* directo del MSE es

$$\mathrm{mse}^*(\widehat{\mu}_d) \triangleq \widehat{\mathrm{MSE}}^*(\widehat{\mu}_d) \;=\; \frac{1}{B}\sum_{b=1}^{B}\big(\widehat{\mu}_d^{*(b)} - \mu_d^{*(b)}\big)^2.$$

Para corregir el sesgo de primer orden en mse*, puede emplearse un doble *bootstrap*, lo cual mejora la precisión (Hall y Maiti, 2006; Esciulescu y Fuller, 2015; Diz-Rosales et al. 2023).

Sin embargo, estudios recientes, como el realizado por Diz-Rosales et al. (2023), demuestran que el *bootstrap* paramétrico propuesto por González-Manteiga et al. (2008, 2010) es comparable en términos de sesgo. Este enfoque, aunque no incluye una corrección explícita del sesgo, presenta ventajas en términos de eficiencia computacional y precisión en la estimación del MSE, lo que lo convierte en una opción preferible en muchas aplicaciones prácticas.

Así, aunque existen múltiples métodos para la estimación del MSE en el modelo de Fay-Herriot, el *bootstrap* paramétrico de González-Manteiga et al. (2008, 2010) se ha consolidado como una herramienta robusta y eficiente, incluso en comparación con métodos más complejos que incluyen correcciones de sesgo.

Con esto, se tiene el cálculo del MSE y el CV, junto con la construcción de intervalos de confianza y tests de hipótesis, los cuales proporcionan herramientas para la evaluación y validación del modelo de Fay-Herriot en aplicaciones de la SAE.

4.2. Modelo Bivariante Fay-Herriot (BFH)

El análisis conjunto de variables relacionadas es fundamental en numerosos problemas estadísticos y aplicaciones prácticas. En muchos contextos, las variables de interés presentan una fuerte correlación que no puede ser ignorada sin comprometer la calidad de las estimaciones. Un ejemplo claro se encuentra en el estudio del consumo de los hogares, los ingresos y los niveles de pobreza dentro de una determinada región, donde estas magnitudes no solo están correlacionadas, sino que pueden compartir restricciones estructurales o composicionales. Situaciones análogas surgen en el análisis del mercado laboral, donde el número total de ocupados, parados e inactivos se encuentra sujeto a restricciones inherentes a la población activa y a la dinámica del empleo.

El uso de modelos univariantes, como el modelo Fay-Herriot aplicado de manera separada a cada variable, resulta problemático en estos escenarios. Al tratar cada componente de manera independiente, se pierde la posibilidad de capturar relaciones fundamentales entre las variables, lo que puede conducir a estimaciones inadecuadas o inconsistentes. La ignorancia de posibles restricciones comunes introduce además el riesgo de generar resultados incongruentes, dificultando la interpretación conjunta de las estimaciones.

Por el contrario, los modelos multivariantes permiten una formulación más general y flexible, donde la información compartida entre las variables se incorpora explícitamente en la estimación. Ignorar esta estructura común puede traducirse en una pérdida significativa de información y, en consecuencia, en una reducción de la eficiencia de los estimadores. Además, cuando se pretende realizar un análisis conjunto, la aplicación de modelos univariantes por separado complica la inferencia y la interpretación global de los resultados. Por estas razones, es deseable contar con una metodología que no solo capture la dependencia entre las variables, sino que también preserve restricciones o condiciones que puedan ser relevantes en el contexto del estudio.

El modelo Fay-Herriot multivariante se presenta como una herramienta adecuada para abordar estas problemáticas (Benavent y Morales 2016). Su formulación permite modelar de manera conjunta varias variables, incorporando explícitamente la estructura de dependencia subyacente. Gracias a ello, se logra una mayor precisión en las estimaciones, aprovechando la información compartida para mejorar la eficiencia de los procedimientos inferenciales. Además, su flexibilidad lo hace aplicable en contextos

[2]Se comenta que es común que los estimadores del MSE se denoten por minúsculas.

diversos, incluyendo aquellos en los que la dependencia entre las variables es particularmente relevante. En comparación con alternativas como el modelo Fay-Herriot multinomial (López-Vizcaíno 2014, 2015), su formulación ofrece ventajas en términos de adaptabilidad y generalidad, aunque éste último sigue siendo una opción atractiva cuando se trabaja con respuestas composicionales.

En este apartado, se pondrá especial énfasis en el caso bivariante del modelo Fay-Herriot, que se denotará por BFH por sus siglas en inglés, y su aplicación a problemas en los que las respuestas presentan una estructura composicional con tres categorías. Esta formulación permite capturar la relación entre las variables de manera más efectiva, asegurando estimaciones consistentes y facilitando una interpretación coherente de los resultados. De todas formas, la totalidad de los resultados aquí presentados pueden ser extrapolados al caso general de más variables.

4.2.1. Modelo de Fay-Herriot Bivariante

Así, se considera la siguiente notación. Sea $\boldsymbol{y}_d = (y_{d1}, y_{d2})^\top$ el vector de estimadores directos de proporciones $\overline{\boldsymbol{Y}}_d = (\overline{Y}_{d1}, \overline{Y}_{d2})^\top$ de alguna variable de clasificación. Sea $\boldsymbol{V}_{ed} = \widehat{\mathbb{V}\mathrm{ar}}_\pi(\boldsymbol{y}_d)$ la matriz de estimadores de covarianza basados en el diseño, que puede contener covarianzas positivas y negativas.

Sea $\boldsymbol{\mu}_d = \mathbb{E}_\pi(\boldsymbol{y}_d) = (\mu_{d1}, \mu_{d2})^\top$ el vector de esperanzas basadas en diseño de $\boldsymbol{y}_d$. El modelo BFH se define en dos etapas. La primera etapa es

$$\boldsymbol{y}_d = \boldsymbol{\mu}_d + \boldsymbol{e}_d, \quad d = 1, \ldots, D,$$

donde los vectores $\boldsymbol{e}_d \overset{i.i.d.}{\sim} \mathcal{N}_2(0, \boldsymbol{V}_{ed})$ son independientes y las matrices de covarianza 2×2 $\boldsymbol{V}_{ed} = \left(\sigma_{dij}\right)_{i,j=1,2}$ son conocidas.

Además, se supone que las μ_{dk} están relacionadas linealmente con las variables explicativas p_k asociadas a la k-ésima categoría en el dominio d. Para $k = 1, 2$, sea $\boldsymbol{x}_{dk} = (x_{dk1}, \ldots, x_{dkp_k})$ un vector de fila que contiene las variables explicativas p_k para μ_{dk}, y $\boldsymbol{X}_d = \mathrm{diag}(\boldsymbol{x}_{d1}, \boldsymbol{x}_{d2})_{2\times p}$ la matriz correspondiente, con $p = \sum_{k=1}^{q-1} p_k$. Sea $\boldsymbol{\beta}_k$ un vector columna de tamaño p_k que contiene los parámetros de regresión para μ_{dk} y sea $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top\right)^\top_{p\times 1}$. Entonces el modelo BFH es

$$\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta} + \boldsymbol{u}_d, \quad \boldsymbol{u}_d \overset{i.i.d.}{\sim} \mathcal{N}_2(0, \boldsymbol{V}_{ud}), \quad d = 1, \ldots, D,$$

donde los vectores $\boldsymbol{u}_d$ son independientes e independientes de los vectores $\boldsymbol{e}_d$. Las matrices de covarianza 2×2 $\boldsymbol{V}_{ud}$ no están estructuradas y dependen de 3 parámetros desconocidos, $\theta_1 = \sigma^2_{u1}$, $\theta_2 = \sigma^2_{u2}$, $\theta_3 = \rho$. La matriz $\boldsymbol{V}_{ud}$ explica la estructura de covarianza de los estimadores directos $\boldsymbol{y}_d$ que no es tenida en cuenta por los errores de muestreo $\boldsymbol{e}_d$ ni por las variables auxiliares $\boldsymbol{X}_d$.

En forma matricial, el modelo BFH es

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \cdots + \boldsymbol{Z}_D\boldsymbol{u}_D + \boldsymbol{e},$$

donde $\boldsymbol{e}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_D$ son independientes con distribuciones

$$\boldsymbol{e} \sim \mathcal{N}_{2D}\left(0, \boldsymbol{V}_e\right), \quad \boldsymbol{u} \sim \mathcal{N}_{2D}(0, \boldsymbol{V}_u) \quad \text{y} \quad \boldsymbol{u}_d \sim \mathcal{N}_2(0, \boldsymbol{V}_{ud}), \quad d = 1, \ldots, D,$$

y $\boldsymbol{X} = \underset{1\le d\le D}{\mathrm{col}}(\boldsymbol{X_d})$, $\boldsymbol{Z}_d = \underset{1\le \ell\le D}{\mathrm{col}}\left(\delta_{\ell d}\boldsymbol{I}_2\right)$, $\boldsymbol{Z} = \mathrm{col}^\top_{1\le d\le D}\left(\boldsymbol{Z}_d\right) = \boldsymbol{I}_{2D}$. Bajo este modelo se cumple que

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$$

$$\boldsymbol{V} = \mathbb{V}\,\mathrm{ar}(\boldsymbol{y}) = \boldsymbol{Z}^\top\boldsymbol{V}_u\boldsymbol{Z} + \boldsymbol{V}_e = \boldsymbol{V}_u + \boldsymbol{V}_e = \underset{1\le d\le D}{\mathrm{diag}}\left(\boldsymbol{V}_d\right)$$

donde $\boldsymbol{V}_d = \boldsymbol{V}_{ud} + \boldsymbol{V}_{ed}$, $d = 1, \ldots, D$.

En Esteban et al. (2020) se puede consultar las expresiones de las estimaciones de los parámetros y de los predictores composicionales propuestos para el caso TFH (caso trivariante Fay-Herriot), mientras que para el caso BFH podemos ir a Morales et al. (2021).

4.2.2. Estimación e inferencia de los parámetros del modelo BFH

El enfoque clásico para la predicción en este modelo se basa en el Teorema de la Predicción de Henderson (Teorema 1, Sección 4.1.1), que justifica la búsqueda de los mejores predictores lineales insesgados y los mejores estimadores lineales insesgados.

Este teorema será aplicado sobre $\boldsymbol{\mu}$, que se obtiene sustituyendo las matrices de covarianza estimadas en las expresiones de los BLUP y BLUE. El predictor de $\boldsymbol{\mu}$ es:

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{u}},$$

donde $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{u}}$ son los estimadores empíricos de $\boldsymbol{\beta}$ y $\boldsymbol{u}$, respectivamente. Estos estimadores se utilizan para predecir las respuestas en cada área, teniendo en cuenta tanto las variables explicativas como los efectos aleatorios.

En cuanto a los métodos de estimación, tanto el método ML como REML son aplicables al modelo BFH, con ligeras adaptaciones en la función de log-verosimilitud. Ambos métodos conservan sus ventajas y desventajas habituales, como la capacidad de REML para proporcionar estimaciones menos sesgadas en comparación con ML, especialmente en muestras pequeñas.

El método REML es el más común para estimar los parámetros desconocidos del modelo BFH, denotados como $\boldsymbol{\theta}$, los cuales están contenidos en la matriz de varianza-covarianza $\boldsymbol{V}$. La función de log-verosimilitud restringida para el modelo BFH se define como:

$$\ell_{\text{REML}}(\boldsymbol{\theta}) = -\frac{2D-r}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{X}^\top\boldsymbol{X}| - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{P}\boldsymbol{y},$$

donde $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}$. Esta función se maximiza para obtener los estimadores REML de $\boldsymbol{\theta}$. Los estimadores obtenidos mediante este método poseen propiedades asintóticas bien definidas. En particular, se distribuyen aproximadamente como una normal multivariante:

$$\widehat{\boldsymbol{V}}_u \sim \mathcal{N}_3(\boldsymbol{V_u}, \boldsymbol{F}^{-1}(\boldsymbol{V_u})), \quad \widehat{\boldsymbol{\beta}} \sim \mathcal{N}_r(\boldsymbol{\beta}, (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}),$$

donde $\boldsymbol{F}$ es la matriz de información de Fisher. Recordar que en la matriz $\widehat{\boldsymbol{V}}_u$ es donde se hallan los parámetros $\widehat{\boldsymbol{\theta}}$.

Gracias a las características de normalidad de los métodos REML y su distribución asintótica, se pueden usar para construir intervalos de confianza asintóticos de nivel $(1-\alpha)$ para las componentes θ_ℓ de $\boldsymbol{\theta}$ y β_i de $\boldsymbol{\beta}$, es decir,

$$\widehat{\theta}_\ell \pm z_{\alpha/2}v_{\ell\ell}^{1/2}, \ell = 1,2,3, \quad \widehat{\beta}_i \pm z_{\alpha/2}q_{ii}^{1/2}, i = 1,\ldots,r,$$

donde $F^{-1}(\widehat{\boldsymbol{\theta}}) = (v_{ab})_{a,b=1,2,3}$, $\left(X'V^{-1}(\widehat{\theta})X\right)^{-1} = (q_{ij})_{i,j=1,\ldots,r}$ y z_α es el cuantil α de la distribución $\mathcal{N}(0,1)$. Para $\widehat{\beta}_i = \beta_0$, el valor p asintótico para el contraste de hipótesis $H_0 : \beta_i = 0$ es

$$p\text{-valor } = 2\mathbb{P}_{H_0}\left(\widehat{\beta}_i > |\beta_0|\right) = 2\mathbb{P}\left(\mathcal{N}(0,1) > |\beta_0|/\sqrt{q_{ii}}\right).$$

Para inicializar los algoritmos basados en la verosimilitud (Sección 3.3.2), pueden emplearse métodos basados en momentos, como el propuesto por Prasad y Rao (1990), que sirven como puntos de arranque para cada categoría por separado. Alternativamente, se pueden utilizar los valores de $\widehat{\sigma}_u^2$ estimados a partir de modelos FH univariante ajustados por REML para cada categoría, es decir, marginales.

Una estrategia alternativa para la determinación de los parámetros iniciales del método iterativo por Fisher-Scoring para mejorar la estimación de los componentes de varianza aleatoria consiste en ampliar el ámbito de análisis a un supradominio, de forma que se aproveche una población más numerosa para obtener parámetros de varianza más estables y luego se apliquen esos valores al dominio pequeño de interés. Concretamente, primero se ajusta un modelo de Fay-Herriot por REML sobre el conjunto ampliado, por ejemplo, considerando a todos los individuos mayores de 16 años en el municipio, lo

que permite estimar con menor variabilidad la varianza de los efectos aleatorios σ_u^2 y las varianzas de muestreo σ_d^2. A continuación, dichos estimadores globales se "trasladan" al subgrupo específico, por ejemplo, solamente a la franja de edad 30-40 años, sin volver a recalcular las varianzas a partir de la muestra limitada. De este modo, se consigue un compromiso entre la precisión numérica, derivada del mayor tamaño efectivo de muestra del supradominio, y la pertinencia local de los EBLUPs en el dominio reducido, manteniendo la coherencia del modelo y evitando inestabilidades en la estimación de los componentes de varianza.

Finalmente, para la inferencia sobre el estimador $\widehat{\boldsymbol{\mu}}_d$, se recomienda el uso del bootstrap paramétrico, siguiendo el enfoque propuesto por González-Manteiga et al. (2008). Este método es particularmente útil cuando se realizan transformaciones en las variables respuesta, ya que permite estimar varianzas, intervalos de confianza y errores cuadráticos medios. El argumentario al respecto de la estructura del método *bootstrap* es similar al explicado en el apartado del modelo FH para el caso univariante.

4.2.3. Modelo BFH para el caso composicional: transformación logística aditiva (alr)

El modelo BFH se extiende, no solo para casos generales como el anterior, sino también para abordar situaciones en las que las respuestas son proporciones o composiciones, es decir, cuando las variables de interés representan partes de un todo que suman uno o una cantidad poblacional fija como un total. Este enfoque es particularmente útil en contextos donde se requiere modelizar datos bivariantes o multivariantes que están restringidos a un espacio simplicial, como en el caso de proporciones de categorías dentro de un área geográfica o demográfica. A continuación, se desarrolla el marco teórico y metodológico para la aplicación del modelo BFH a respuestas transformadas y composicionales, junto con las técnicas de estimación y predicción asociadas. Esta es la base del Objetivo 1, que consiste en la obtención de los totales de parados, ocupados e inactivos y las tasas de desempleo por municipios de más de 16.000 habitantes.

4.2.3.1. Notación, introducción y problema de los ceros

En primer lugar, se introduce una notación auxiliar para trabajar con proporciones y con sus transformaciones. Así, cada dominio estará dividido en subconjuntos U_{dk}, $k = 1, \ldots, q$, definidos por la variable de clasificación k, que cataloga las unidades en un número finito q de categorías. Dichas categorías, en el modelo BFH, son $q = 3$, que como ejemplo se pueden considerar como ocupados ($k = 1$), parados ($k = 2$), e inactivos ($k = 3$).

Establecida esta notación y clasificación del espacio de conjuntos, se redefinen las proporciones muestrales y los recuentos, para casos donde hay varias categorías

$$\overline{z}_{dk} = \frac{\widetilde{z}_{dk}}{n_d}, \quad \widetilde{z}_{dk} = \sum_{j \in s_d} z_{dj}, \quad d = 1, \ldots, D.$$

Los estimadores directos de Hájek de $\overline{Z}_{dk}$ y Z_{dk} son

$$\widehat{\overline{Z}}_{dk}^{\mathrm{H}} = \frac{\widehat{Z}_{dk}^{\mathrm{H}}}{\widehat{N}_d}, \quad \text{donde} \quad \widehat{Z}_{dk}^{\mathrm{H}} = \sum_{j \in s_d} w_{dj} z_{dj} = \widehat{\overline{Z}}_{dk}^{\mathrm{H}} \widehat{N}_d, \quad \widehat{N}_d = \sum_{j \in s_d} w_{dj}.$$

A continuación, se empleará, por simplicidad, la notación $z_{dk} \triangleq \widehat{\overline{Z}}_{dk}^{H}$, $k = 1, \ldots, q$, y el vector de estimaciones directas como $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dq-1})^\top$. Estas proporciones se asumen positivas, es decir, $z_{dk} > 0$, y en el caso de respuestas composicionales, se supone que suman uno:

$$z_{d1} + z_{d2} + \cdots + z_{dq} = 1.$$

En este contexto, es crucial hacer un matiz acerca de la posible nulidad de las proporciones, ya que este es un problema significativo que puede generar dificultades en la estimación de los modelos en el

contexto de la SAE, y que fue especialmente relevante en el estudio del Objetivo 3 explicados en la sección 2.2.

En la práctica, pueden existir composiciones de dominio $(z_{d1}, \ldots, z_{dq})$ en las que algunos componentes sean cero, es decir, $z_{dk} = 0$. Dado que el logaritmo de cero es $-\infty$, no es posible aplicar transformaciones logarítmicas para analizar datos composicionales. En el caso particular del Objetivo 1, se cuenta con pocos de estos ceros. Una solución práctica para este problema consiste en reemplazar los ceros por un valor numérico pequeño $\varepsilon > 0$ y realizar un proceso de ajuste por redondeo.

Por ejemplo, si $(z_{d1}, \ldots, z_{dq})$ tiene m ceros, se puede aplicar la siguiente aproximación: se toma como δ el error de redondeo máximo de los valores z_{dk} positivos. Se denota como u a la décima parte del mínimo de los valores z_{dk}, excluyendo aquellos que estén por debajo de δ. De esta forma, se suma a todos los valores inferiores a δ un valor u, y se resta un valor $u/(m-\kappa)$ a aquellos valores superiores, siendo κ el número total de valores por debajo de δ. Este ajuste permite resolver los problemas relacionados con los ceros en los valores z_{dk}, garantizando que en aquellos cruces sin muestra de una categoría, como en el caso de la categoría de parados, donde suele ser común la no observación de los mismos en áreas muy pequeñas, se asigna un valor u que evita la nulidad. Este es el procedimiento adoptado para las categorías en cuestión para el tratamiento de ceros en los casos prácticos. Otra alternativa consiste en utilizar la aproximación de Aitchison (1982), donde a cada variable se le suma una pequeña cantidad relacionada con la cantidad de ceros que tenga.

Es importante señalar que el Objetivo 3 presentó una mayor complejidad en comparación con los objetivos previos, ya que, en numerosos casos, se observaron configuraciones de datos en las que una única categoría laboral domina por completo la medición. Estas configuraciones, representadas por ternas del tipo $(1, 0, 0)$ en el caso bivariante, o, en general, formas del tipo $(0, \ldots, 1, \ldots, 0)$, indican que una de las categorías de la estimación, como por ejemplo de la fuerza laboral, como el porcentaje de ocupados, parados o inactivos, concentra toda la estimación, mientras que las otras dos categorías tienen valores nulos. Esta particularidad incrementa la dificultad del análisis, pues introduce varianza y patrones específicos que deben ser tratados adecuadamente en el modelado.

Para abordar estos casos, se ajustan las varianzas y covarianzas del modelo de manera que reflejen fielmente la estructura de los datos. Este ajuste es esencial para garantizar la fiabilidad de los estimadores y para representar adecuadamente los valores atípicos. Al aplicar esta metodología a datos reales, un elevado número de ajustes puede generar en los gráficos de estimadores directos y transformados la aparición de líneas rectas, verticales u horizontales, que ilustran la presencia de categorías dominantes en la muestra. Este fenómeno se observó con claridad en el desarrollo del Objetivo 3; sin embargo, no se profundiza en él aquí, ya que utiliza la misma estructura modelística que el Objetivo 1. El objetivo principal de este trabajo es la aplicación práctica de los modelos FH en la SAE sobre datos reales. No obstante, conviene destacar que la coexistencia de categorías dominantes, y, por ende, de ceros en las demás, es habitual en este campo y requiere un tratamiento específico durante el análisis y la modelización.

4.2.3.2. Transformación alr de los datos composicionales

Retornando al aspecto teórico, para modelizar estas proporciones, se aplica una transformación $h(\cdot)$ a las respuestas, obteniendo un vector transformado $\boldsymbol{y}_d = h(\boldsymbol{z}_d)$, que se utiliza como variable dependiente en el modelo BFH. La transformación debe ser adecuada para garantizar que el modelo BFH sea válido y que las propiedades estadísticas de las respuestas transformadas permitan una inferencia adecuada. En concreto, la transformación que se presentará a continuación es especialmente común al tratar con este tipo de datos (Esteban et al. 2020).

Una de las transformaciones recomendadas es la transformación logística aditiva (alr), y se aplica a los datos composicionales con el objetivo de superar las limitaciones inherentes al espacio del símplice, donde estos datos residen originalmente. En contextos composicionales, como es el caso de las proporciones entre categorías laborales (ocupados, parados e inactivos), cada observación está compuesta por un vector de proporciones que suman uno, lo cual induce una dependencia perfecta entre los componentes y restringe el análisis estadístico convencional.

La transformación alr permite proyectar los datos desde el símplice a un espacio euclídeo de dimensión $q-1$, donde q es el número de componentes composicionales. Se recuerda al lector que aquí se trabajará con tres categorías. No obstante, se expone la generalización a q categorías. Esta proyección se realiza mediante la transformación logarítmica de cada componente con respecto a una categoría de referencia, generalmente elegida por criterios de interpretabilidad o estabilidad numérica.

El principal beneficio de esta transformación es que convierte los datos a un dominio no restringido, lo que permite aplicar modelos de Fay-Herriot, cuya formulación asume normalidad y no colinealidad entre las variables explicativas. Además, la transformación alr conserva la estructura relativa de los componentes, respetando la naturaleza composicional de los datos, y facilita la interpretación de los efectos relativos entre categorías.

En el contexto de la estimación en áreas pequeñas, el uso de la transformación alr es especialmente relevante, ya que mejora la estabilidad numérica de los estimadores, evita problemas derivados de proporciones cercanas a cero (que, sin transformación, generan distorsiones significativas en el ajuste del modelo) y permite capturar adecuadamente la variabilidad relativa entre categorías. Por tanto, la alr se configura como una herramienta esencial para modelar datos composicionales dentro del marco de modelos lineales mixtos, asegurando una representación adecuada de la estructura y dependencia subyacente entre componentes.

Cabe señalar, además, que su uso no solo es conveniente desde un punto de vista técnico, sino también metodológicamente coherente con el enfoque de análisis composicional propuesto en la literatura clásica (Aitchison 1982), siendo preferido frente a transformaciones alternativas como la *clr* (*centered log-ratio*) cuando se busca simplicidad en la estructura del modelo y mayor facilidad de interpretación en contextos aplicados. Para más detalle, se recomienda la lectura del material suplementario del artículo de Esteban et al. (2020), donde también se tratan otras transformaciones de interés.

La transformación se define, componente a componente, como:

$$y_{dk} = \operatorname{alr}(z_{dk}) = \log\left(\frac{z_{dk}}{z_{dq}}\right), \quad k = 1, \ldots, q-1,$$

donde z_{dq} es la categoría de referencia. Esta transformación es particularmente útil cuando las proporciones no son cercanas a cero, ya que evita problemas numéricos asociados con valores extremos.

La inversa de esta transformación permite recuperar las proporciones originales a partir de las respuestas transformadas:

$$z_{dk} = \operatorname{alr}^{-1}(y_{dk}) = \frac{\exp\{y_{dk}\}}{1 + \exp\{y_{d1}\} + \cdots + \exp\{y_{dq-1}\}}, \quad k = 1, \ldots, q-1,$$

y para la categoría de referencia:

$$z_{dq} = 1 - z_{d1} - \cdots - z_{dq-1} = \frac{1}{1 + \exp\{y_{d1}\} + \cdots + \exp\{y_{dq-1}\}}.$$

Para obtener una aproximación de la varianza de las respuestas transformadas, se emplea una expansión de Taylor de primer orden de la transformación alr alrededor de un punto fijo $\boldsymbol{z}_0$, habitualmente tomado como $\boldsymbol{z}_0 = \mathbf{1}_{q-1}/q$, siendo $\mathbf{1}_g$ un vector columna de unos de dimensión g. Esta linealización permite expresar la varianza de la variable transformada $\boldsymbol{y}_d$ como una función lineal de la varianza original de las composiciones $\boldsymbol{z}_d$.

$$\widehat{\mathbb{V}\mathrm{ar}}_\pi(\boldsymbol{y}_d) \approx \boldsymbol{H}_0 \, \widehat{\mathbb{V}\mathrm{ar}}_\pi(\boldsymbol{z}_d) \, \boldsymbol{H}_0^\top,$$

donde $\boldsymbol{H}_0$ es una matriz que depende de las proporciones $\boldsymbol{z}_d$. Los elementos de esta matriz se definen como:

$$H_{kk}(\boldsymbol{z}_d) = \frac{1}{z_{dk}} + \frac{1}{z_{dq}}, \quad H_{k_1 k_2}(\boldsymbol{z}_d) = \frac{1}{z_{dq}} \quad \text{si} \quad k_1 \neq k_2.$$

Esta aproximación es esencial para incorporar la incertidumbre asociada con las respuestas transformadas en el modelo BFH.

4.2.3.3. Estimación en el modelo composicional

El modelo BFH se aplica al vector de respuestas transformadas $\boldsymbol{y}_d$, con el objetivo de predecir las proporciones originales $\boldsymbol{z}_d$. En este contexto, las esperanzas condicionales $\boldsymbol{\mu}_d = \mathbb{E}[\boldsymbol{y}_d|\boldsymbol{u}_d]$ se modelizan como:

$$\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta} + \boldsymbol{u}_d,$$

donde $\boldsymbol{X}_d$ es una matriz de variables explicativas, $\boldsymbol{\beta}$ es un vector de parámetros de regresión, y $\boldsymbol{u}_d$ es un vector de efectos aleatorios con distribución normal bidimensional. Los parámetros del modelo se estiman utilizando métodos como REML, que maximizan una función de log-verosimilitud restringida para obtener estimadores insesgados de los parámetros de varianza.

Una vez estimados los parámetros del modelo, se procede a predecir las proporciones composicionales. Sin embargo, debido a la no linealidad de la transformación, surgen problemas relacionados con la desigualdad de Jensen, que implica que (Esteban et al. 2020)

$$\mathbb{E}_\pi[y_{dk}] \neq \log\left(\frac{\overline{Z}_{dk}}{\overline{Z}_{dq}}\right).$$

En efecto, y ya incluso para el caso sin transformar, el estimador directo z_{dk} no se calcula utilizando el inverso de las probabilidades de inclusión, sino con factores de expansión que no están calibrados a los totales del dominio. Por lo tanto, z_{dk} suele estar sesgado con respecto a la distribución del diseño muestral. En consecuencia, los predictores basados en los modelos FH enfrentan dificultades para cumplir la suposición anterior cuando se trabaja con datos reales.

Por otro lado, para cumplir con la suposición de que la esperanza basada en el diseño es la realización de la esperanza condicional del modelo, que denotamos por $\mathbb{E}_{\text{BFH}}$, es decir,

$$\mathbb{E}_{\text{BFH}}\left[z_{dk} \mid u_{\text{BFH},dk}\right] = \mathbb{E}_\pi\left[z_{dk}\right],$$

las funciones de los efectos fijos y aleatorios del modelo que se deben predecir bajo el modelo BFH, al mismo estilo que con el TFH (Esteban et al. 2020) deben ser, denotando $\mathbb{E}_C$ como la esperanza bajo el modelo composicional

$$\mathbb{E}_C\left[z_{dk} \mid \boldsymbol{u}_d\right] = \int_{\mathbb{R}^2} \frac{\exp\{y_{dk}\}}{1+\exp\{y_{d1}\}+\exp\{y_{d2}\}} f_{\mathcal{N}_2(\boldsymbol{X}_d\boldsymbol{\beta}+\boldsymbol{u}_d,\boldsymbol{V}_{ed})}(\boldsymbol{y}_d)\, d\boldsymbol{y}_d\,, \quad k=1,2,$$

y $\mathbb{E}_C\left[z_{d3} \mid \boldsymbol{u}_d\right] = 1 - \mathbb{E}_C\left[z_{d1} \mid \boldsymbol{u}_d\right] - \mathbb{E}_C\left[z_{d2} \mid \boldsymbol{u}_d\right], d = 1,\ldots,D$, que son funciones no lineales de $\boldsymbol{u}_d$ basadas en integrales que no se pueden resolver analíticamente. Además, los mejores predictores empíricos de $\mathbb{E}_C\left[z_{dk} \mid \boldsymbol{u}_d\right]$ son integrales no analíticamente tratables de $\mathbb{E}_C\left[z_{dk} \mid \boldsymbol{u}_d\right]$ con respecto a la densidad de $\boldsymbol{u}_d$ condicionada a $\boldsymbol{y}_d$.

Para abordar este problema, se propone un predictor práctico, basado en $\mu_{dk} = \mathbb{E}_C\left[y_{dk} \mid \boldsymbol{u}_d\right]$, en un enfoque de *plug-in*:

$$\widehat{p}^C_{dk} = \frac{\exp\{\widehat{\mu}_{dk}\}}{1+\sum_{l=1}^q \exp\{\widehat{\mu}_{dl}\}},$$

donde $\widehat{\mu}_{dk}$ es la predicción de la esperanza condicional para la categoría k en el área d, y dada por $\widehat{\mu}_{dk} = \boldsymbol{x}_{dk}\widehat{\beta}_k + \widehat{u}_{dk}$. Este predictor permite estimar las proporciones composicionales de manera eficiente, aunque introduce cierto sesgo debido a la no linealidad de la transformación. En la práctica, la condición de que la esperanza condicional bajo el modelo sea igual a bajo el diseño se soluciona adecuadamente bajo un ajuste particularmente bueno con variables bastante correlacionadas con la variable de interés a predecir.

Además, a partir de la definición anterior se tomará como predictores composicionales de las proporciones del dominio $\overline{Z}_{dk}$ a $\widehat{p}_{dk}$, y los predictores composicionales de los totales $Z_{dk} = N_d\overline{Z}_{dk}$ a $N_d\widehat{\overline{Z}}_{dk}$, $k = 1,2,3$.

Otra forma de estimar estas proporciones puede ser a través de los EBPs. Así, los predictores composicionales óptimos, también llamados *best predictors*, de las proporciones p_{dk}^C están dados por (Esteban et al. 2020)

$$p_{dk}^B = p_{dk}^B(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\theta}}[p_{dk}^C | \boldsymbol{y}_d], \quad k = 1, 2,$$

y

$$p_{d3}^B = 1 - p_{d1}^B - p_{d2}^B.$$

Se cumple que

$$p_{dk}^{\mathrm{B}} = \frac{\displaystyle\int_{\mathbb{R}^2} \frac{\exp\left(\mu_{dk}^C\right)}{1 + \exp\left(\mu_{d1}^C\right) + \exp\left(\mu_{d2}^C\right)} f(\boldsymbol{y}_d \mid \boldsymbol{u}_d)\, f_{\boldsymbol{\theta}}(\boldsymbol{u}_d)\, d\boldsymbol{u}_d}{\displaystyle\int_{\mathbb{R}^2} f(\boldsymbol{y}_d \mid \boldsymbol{u}_d)\, f_{\boldsymbol{\theta}}(\boldsymbol{u}_d)\, d\boldsymbol{u}_d}$$

que se puede aproximar como

$$p_{dk}^B = \frac{A_{dk}(\boldsymbol{y}_d, \boldsymbol{\beta}, \boldsymbol{\theta})}{B_d(\boldsymbol{y}_d, \boldsymbol{\beta}, \boldsymbol{\theta})},$$

donde

$$A_{dk}(\boldsymbol{y}_d, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^2} \frac{\exp\{\mu_{dk}^C\}}{1 + \exp\{\mu_{d1}^C\} + \exp\{\mu_{d2}^C\}} \times$$

$$\times \exp\left(-\frac{1}{2}(\boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta} - \boldsymbol{u}_d)^\top \boldsymbol{V}_{ed}^{-1}(\boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta} - \boldsymbol{u}_d)\right) f_{\boldsymbol{\theta}}(\boldsymbol{u}_d)\, d\boldsymbol{u}_d,$$

$$B_d(\boldsymbol{y}_d, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^2} \exp\left(-\frac{1}{2}(\boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta} - \boldsymbol{u}_d)^\top \boldsymbol{V}_{ed}^{-1}(\boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta} - \boldsymbol{u}_d)\right) f_{\boldsymbol{\theta}}(\boldsymbol{u}_d)\, d\boldsymbol{u}_d,$$

y

$$\mu_{dk}^C = \boldsymbol{x}_{dk}\beta_k + u_{dk}, \quad \boldsymbol{y}_d | \boldsymbol{u}_d \sim \mathcal{N}_2(\boldsymbol{X}_d\boldsymbol{\beta} + \boldsymbol{u}_d, \boldsymbol{V}_{ed}), \quad \boldsymbol{u}_d \sim \mathcal{N}_2(\boldsymbol{0}, \boldsymbol{V}_{ud}),$$

y, con ello, el EBP de p_{dk}^C se define como

$$\widehat{p}_{dk}^E = \widehat{p}_{dk}^B(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}), \quad k = 1, 2,$$

el cual puede aproximarse mediante algoritmos de Monte Carlo, método usual en el contexto de la SAE, con la aproximación anterior. Estos métodos suelen ser algo lentos computacionalmente, y dan pie a explorar otras formas que podrían ser de interés por su velocidad serían por análisis numérico y cuadraturas basadas en *sparse grids*, como se puede estudiar en Zhong y Feng (2023).

4.2.3.4. MSE para predictores composicionales mediante *bootstrap* paramétrico

Para estimar la incertidumbre asociada con las predicciones, se utiliza un enfoque de *bootstrap* paramétrico. Este método implica generar muestras *bootstrap* a partir de los parámetros estimados del modelo y calcular las predicciones correspondientes. El algoritmo de *bootstrap* paramétrico, basado en los trabajos de González-Manteiga (2008, 2010) adaptado para respuestas composicionales consta de los siguientes pasos:

Para la estimación del MSE de los predictores *plug-in* composicionales o EBPs transformados, se emplea el método de remuestreo *bootstrap* paramétrico propuesto nuevamente por González-Manteiga et al. (2008a). Este desarrollo se puede ver también reflejado en Esteban et al. (2020). Este procedimiento es especialmente útil cuando las estimaciones se obtienen a partir de transformaciones no lineales como la alr. Los pasos del algoritmo son los siguientes:

Paso 1. Ajuste inicial. Ajustar el modelo transformado de Fay-Herriot bivariante a los datos observados $\{(\boldsymbol{y}_d, \boldsymbol{X}_d)\}_{d=1}^D$, y estimar los parámetros $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\theta}}$, que incluyen tanto los efectos fijos como los componentes de varianza y covarianza.

Paso 2. Generación de datos *bootstrap*. Para cada réplica $b = 1, \ldots, B$, y para cada área $d = 1, \ldots, D$, realizar lo siguiente:

$$\boldsymbol{u}_d^{*(b)} \sim \mathcal{N}_2\big(\boldsymbol{0}, \boldsymbol{V}_{ud}(\widehat{\boldsymbol{\theta}})\big), \quad \boldsymbol{e}_d^{*(b)} \sim \mathcal{N}_2\big(\boldsymbol{0}, \boldsymbol{V}_{ed}\big), \quad \boldsymbol{y}_d^{*(b)} = \boldsymbol{X}_d\widehat{\boldsymbol{\beta}} + \boldsymbol{u}_d^{*(b)} + \boldsymbol{e}_d^{*(b)},$$
$$\boldsymbol{\mu}_d^{*(b)} = \boldsymbol{X}_d\widehat{\boldsymbol{\beta}} + \boldsymbol{u}_d^{*(b)}, \quad \boldsymbol{p}_d^{*(b)} = \mathrm{alr}^{-1}(\boldsymbol{\mu}_d^{*(b)}).$$

Paso 3. Cálculo del predictor *bootstrap*. A partir de los datos simulados $(\boldsymbol{y}_d^{*(b)}, \boldsymbol{X}_d)$, ajustar nuevamente el modelo BFH y calcular los predictores estimados de las proporciones, $\widehat{\boldsymbol{p}}_d^{*(b)}$.

Paso 4. Estimación del MSE. Repetir los pasos 2 y 3 para $b = 1, \ldots, B$, y estimar el MSE *bootstrap* para cada área $d = 1, \ldots, D$ y componente composicional $k = 1, 2, 3$, mediante:

$$\mathrm{mse}^*(\widehat{p}_{dk}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{p}_{dk}^{*(b)} - p_{dk}^{*(b)}\right)^2.$$

Este enfoque proporciona estimaciones de la incertidumbre asociada con las predicciones, incluso en presencia de transformaciones no lineales y respuestas composicionales. Sea como fuere, existen otras transformaciones, como la clr y la ilr, ya mencionadas brevemente, también usadas en el ámbito de los datos composicionales, y que se deja al lector la consulta de Esteban et al. (2020) y, más concretamente, de su material complementario para más información.

Capítulo 5

Aplicación del modelo Fay-Herriot a los problemas de interés para el INE

5.1. Uso del modelo FH univariante para el estudio del total de «ninis» a nivel provincial

En esta sección se aplica el modelo mixto composicional basado en un FH univariante a los datos de la EPA en las provincias de España, abarcando desde el primer trimestre de 2021 hasta el cuarto trimestre de 2022. El objetivo del análisis, relacionado con el Objetivo 2 expuesto en la sección 2.2, es la estimación de las proporciones de individuos clasificados como ninis con respecto a la población de 16 a 24 años en cada dominio. Dicha categoría se considerará como la principal a lo largo del informe y del desarrollo metodológico, mientras que su complementaria, correspondiente a la proporción de no ninis, se tomará como variable de referencia en la estimación, centrándose el estudio en la estimación de las proporciones de ninis dentro de la población de referencia.

Cada dominio está compuesto por un subconjunto U_d, definido según la variable indicadora de clasificación que distingue entre individuos ninis y no ninis, asignando cada unidad a una de estas dos categorías de manera exclusiva. Se consideran $D = 52$ dominios, correspondientes a las provincias y ciudades autónomas de España, observadas a lo largo de ocho trimestres durante el período 2021-2022. En lo sucesivo, se readoptará la notación $z_d \triangleq \widehat{\overline{Z}}_d^{\mathrm{H}}$ para denotar el estimador directo de la proporción objetivo en el dominio d, mientras que $y_d \triangleq \mathrm{alr}(z_d)$ representará la variable transformada para el modelo FH univariante, la cual se reduce, en una dimensión, a la transformación logística para evitar la presunción de linealidad de la probabilidad.

El modelo FH univariante se ajusta a los datos objetivo y a un conjunto de variables agregadas provenientes de la propia EPA, alineándose con los intereses del INE. Las covariables consideradas en el modelo a nivel de dominio, seleccionadas a partir del análisis descriptivo previo, son las siguientes:

1. Proporción de individuos con estudios primarios (*Est. Prim.*) y estudios secundarios (*Est. Secund.*), expresada en tanto por uno con respecto a la población de 16 y más años que ha completado dichos niveles educativos dentro del dominio correspondiente.

2. Proporción de ocupados (*Ocupados*) e inactivos (*Inactivos*), calculada con respecto a la población de 16 y más años.

3. Proporción de población extranjera (*Extranjeros*), expresada en tanto por uno con respecto a la población de 16 y más años.

El análisis se ha realizado empleando tanto los estimadores directos previamente definidos como el modelo FH univariante con la transformación alr descrita en la subsección 4.2.3. En todos los casos, los

trimestres han sido tratados de manera independiente y, cuando sea necesario, se presentarán ambos enfoques con el propósito de evaluar comparativamente su desempeño y determinar si el modelo FH univariante proporciona una mejora estadísticamente significativa en la estimación de las proporciones de ninis en cada dominio.

Conviene, en este punto, recordar brevemente los tres enfoques de estimación que se están empleando en este estudio para la proporción de ninis. En primer lugar, se encuentra el método directo, basado en el estimador de Hájek descrito en el Apartado 3.1.1, cuya expresión se denota por $\widehat{\overline{Z}}_d^{\mathrm{H}}$.

En segundo lugar, se utiliza el modelo FH univariante con transformación composicional alr, explicado en el Apartado 4.2.3.2. Este modelo opera inicialmente sobre la transformación log-ratio aditiva de la variable de interés, ajustando el modelo sobre la variable transformada. Posteriormente, para obtener estimaciones comparables con el método directo, se aplica la inversa de la transformación alr, obteniéndose así el estimador modelizado:

$$\widehat{\overline{Z}}_d^{\mathrm{FH}} = \mathrm{alr}^{-1}(\widehat{y}_d).$$

Por último, se incorpora una tercera metodología: el modelo FH univariante bajo *benchmarking*, desarrollado en el Apartado 3.2.1. Este modelo impone restricciones de coherencia para asegurar que las estimaciones a nivel desagregado se ajusten a los totales conocidos o publicados a niveles superiores. Denotamos sus predicciones como $\widehat{\overline{Z}}_d^{\mathrm{ben}}$, las cuales son resultado de una modificación del estimador modelizado que garantiza la consistencia con cifras oficiales agregadas, manteniendo a su vez una ganancia sustancial en la precisión respecto al enfoque directo.

5.1.1. Estudio descriptivo de los datos y análisis del modelo

El cuadro 5.1 representa las covariables y las correlaciones respecto a la variable respuesta: la proporción de ninis en relación con la población de 16 a 24 años. Se enfoca específicamente en el cuarto trimestre de 2022 (T8), seleccionado a modo ilustrativo por ser el último trimestre en los datos. En el mismo se destaca que la correlación más fuerte y positiva con la proporción de ninis ocurre con la proporción de personas con estudios primarios y la concentración de población de origen extranjero, mientras que la ocupación muestra una correlación inversa menos significativa.

Conforme aumenta el nivel educativo, la correlación disminuye, indicando que existe una menor probabilidad de entrar en la condición de nini a medida que aumenta el nivel educativo, lo cual es un patrón esperado.

Es crucial observar la conservación de signos de las correlaciones al aplicar la transformación alr, explicada en el Apartado 4.2.3.2, al estimador directo, de forma que las conclusiones son análogas. Esta transformación proporciona una mayor robustez tanto contextual como conclusiva al uso de dicha variable a lo largo del informe, asegurando que las inferencias realizadas sean más fiables y coherentes en diversos contextos analíticos.

	Est. Prim.	Est. Secund.	Ocupados	Inactivos	Extranjeros
z_d	0.28	0.26	-0.14	-0.22	0.26
y_d	0.30	0.22	-0.15	-0.20	0.24

Cuadro 5.1: Correlaciones de la variable respuesta z_d y de su transformada y_d, con las covariables para el cuarto trimestre del año 2022.

Para explorar en detalle estas relaciones, las cuales son significativas a un 99 % de confianza por el método de Spearman (Best y Roberts 1975), se analizarán gráficamente mediante gráficos de dispersión (Fig. 5.1).

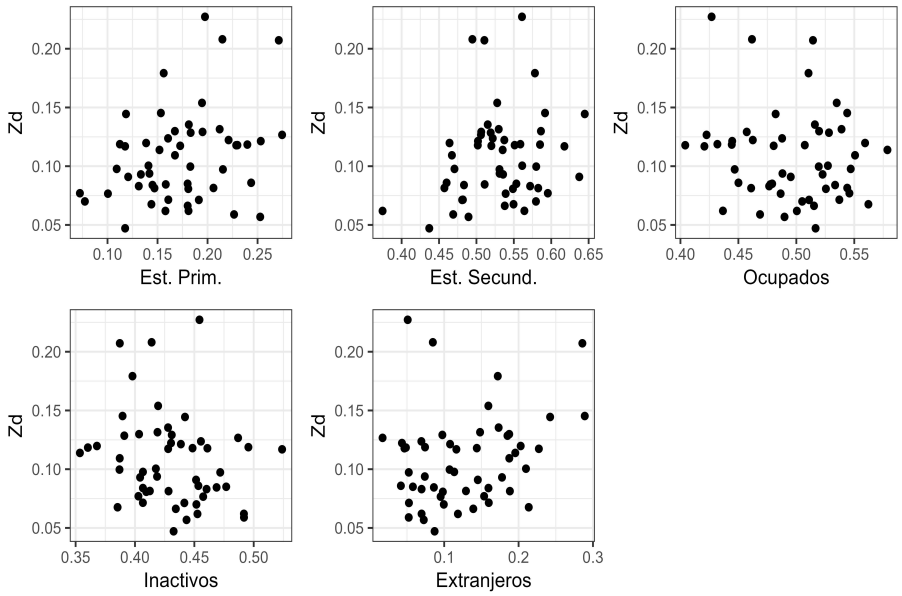


Figura 5.1: Gráficos de dispersión entre la variable objetivo z_d, proporción de ninis respecto a la población de 16 a 24 años y las covariables de interés para el cuarto trimestre del año 2022.

A continuación, se presenta el ajuste del modelo FH univariante con transformación alr para el cuarto trimestre de 2022. El Cuadro 5.2 recoge los coeficientes estimados $\widehat{\beta}_l$ para cada una de las covariables utilizadas, junto con sus errores estándar, intervalos de confianza al 95 % y valores p asociados a la prueba de hipótesis nula $H_0 : \beta_l = 0$. El propósito de este análisis es valorar el grado de influencia de cada covariable sobre la proporción transformada de jóvenes clasificados como ninis y verificar la significación estadística de dichos efectos.

Los resultados obtenidos indican que todas las covariables incluidas en el modelo resultan estadísticamente significativas al nivel del 5 %, y en la mayoría de los casos al nivel del 1 %. Esta fuerte significación otorga valor al modelo ajustado y respalda su validez empírica para la predicción de la proporción de ninis en dominios con tamaño muestral reducido. Así, se desgrana cada estimación de $\widehat{\beta}_l$:

- **Estudios primarios:** el coeficiente estimado es de 2.27, con un intervalo de confianza al 95 % entre 0.65 y 3.89, y un valor $p < 0{,}01$. Esta magnitud refleja una relación positiva y significativa entre la proporción de población con estudios primarios y la presencia de jóvenes ninis. En términos prácticos, este resultado sugiere que cuanto mayor es la proporción de personas con bajo nivel educativo en una provincia, mayor es la propensión de los jóvenes de esa área a no participar ni en el sistema educativo ni en el mercado laboral.
- **Estudios secundarios:** aunque el coeficiente es algo menor (1.78), sigue siendo positivo y significativo ($p \approx 0{,}01$). Esto indica que incluso la finalización de estudios secundarios no garantiza por sí sola la inserción laboral o educativa de los jóvenes, aunque su efecto sea menos pronunciado que el de los estudios primarios.
- **Ocupación:** la variable con mayor magnitud en términos absolutos es la proporción de personas ocupadas, con un coeficiente negativo de -4.57 y un intervalo de confianza entre -6.09 y -3.05 ($p < 0{,}01$). Este resultado confirma la expectativa teórica de que en provincias con mayores niveles de ocupación, la proporción de jóvenes ninis es significativamente menor. La relación inversa y fuerte sugiere que un entorno laboral activo tiene efectos positivos indirectos sobre la inclusión juvenil.

- **Inactividad:** similar al caso anterior, el coeficiente asociado a la inactividad es negativo (-3.69), aunque de menor magnitud que el de la ocupación. También es altamente significativo. Esto puede interpretarse como una señal de que en territorios con mayor proporción de población inactiva, los jóvenes enfrentan un entorno con menos incentivos o redes activas que faciliten su participación.
- **Población extranjera:** esta variable presenta un coeficiente positivo de 2.44, significativo al 1 %. Aunque la asociación no implica causalidad, el resultado apunta a una relación entre la concentración de población extranjera y una mayor proporción de jóvenes en situación de vulnerabilidad educativa o laboral. Esto podría reflejar dinámicas estructurales ligadas a barreras de integración, diferencias en niveles educativos previos o dificultades de acceso a oportunidades formales.

La consistencia de los signos obtenidos con los valores de correlación presentados previamente (ver Cuadro 5.1) refuerza la validez del modelo. Además, el hecho de que todos los intervalos de confianza excluyan el valor cero proporciona evidencia estadística robusta de que las covariables incluidas en el modelo tienen un impacto significativo sobre la variable respuesta. Este patrón también revela la adecuación del uso del modelo FH transformado, capaz de capturar las relaciones latentes entre factores estructurales y el fenómeno nini, mejorando la precisión en contextos donde las estimaciones directas resultan inestables o inviables.

	Lím. Inf.	Beta	Lím. Sup.	Error Estándar	p-valor
Est. Prim.	0.65	2.27	3.89	0.81	0.00
Est. Secund.	0.32	1.78	3.24	0.73	0.01
Ocupados	-6.09	-4.57	-3.05	0.76	0.00
Inactivos	-5.67	-3.69	-1.70	0.99	0.00
Extranjeros	0.72	2.44	4.16	0.86	0.00

Cuadro 5.2: Estimadores e intervalos de confianza al 95 % con sus valores p para la variable objetivo transformada en el cuarto trimestre del año 2022.

5.1.2. Diagnosis del modelo

En cuanto al análisis de residuos, la Figura 5.2 muestra los residuos estandarizados correspondientes al modelo ajustado para el cuarto trimestre de 2022. Se aprecia un comportamiento adecuado del modelo en términos de ajuste: la mediana de los residuos se sitúa en torno a cero, lo que indica ausencia de sesgo sistemático, y la mayoría de los valores se encuentran dentro del rango comprendido entre -3 y 3, límite convencionalmente adoptado como criterio de diagnóstico. Las líneas rojas horizontales rayadas señalan estos umbrales, que son útiles para identificar observaciones atípicas. En este caso, la presencia de valores extremos es mínima, lo cual refuerza la adecuación del modelo a los datos y sugiere que no existen áreas particularmente mal ajustadas.

Por otro lado, la Figura 5.3 ofrece una comparación directa entre las estimaciones obtenidas mediante el modelo FH transformado y las estimaciones directas de las proporciones de ninis respecto a la población joven de 16 a 24 años. En este gráfico se observa una distribución bastante simétrica de los valores predichos en torno a los valores observados, lo que apunta a una correspondencia razonable entre ambas metodologías. Dado que los estimadores directos de proporciones son, por construcción,

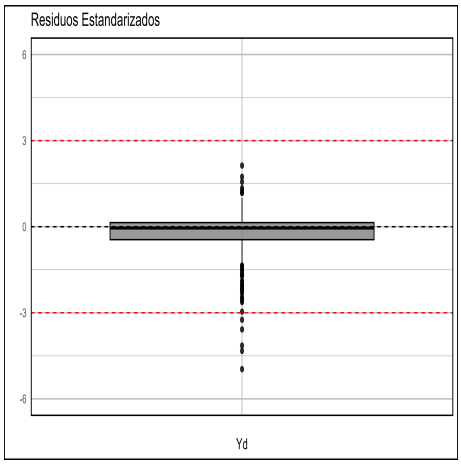


Figura 5.2: Residuos estandarizados del modelo FH univariante para la proporción de ninis respecto a la población de 16 a 24 años.

prácticamente insesgados bajo el diseño muestral, esta simetría sugiere que los estimadores composicionales heredan parcialmente esa propiedad. Es decir, el modelo no introduce distorsiones sistemáticas relevantes, y sus predicciones mantienen una relación estrecha con las observaciones empíricas. Este comportamiento es coherente con lo reportado en la literatura especializada (véase López-Vizcaíno et al. 2014, 2015; Morales et al. 2021).

A continuación, en la Figura 5.4, se analiza la evolución del error de predicción en función del tamaño muestral mediante el uso del RMSE, tal como se define en la ecuación (4.2). En este análisis comparativo se incluyen tres enfoques: el estimador directo (color verde), el modelo FH univariante (color roja) y el modelo FH univariante bajo *benchmarking* (color negra), descrito con detalle en el Apartado 3.2.1. La tendencia general muestra que el modelo FH univariante presenta una clara mejora en términos de error medio cuadrático: no solo logra valores de RMSE más bajos que el estimador directo, sino que también exhibe menor variabilidad. Esta ventaja se mantiene de forma consistente a lo largo de todo el rango de tamaños muestrales analizados.

Además, el modelo FH con *benchmarking* presenta un rendimiento intermedio: aunque introduce una ligera penalización en términos de variabilidad, derivada de la restricción de conformidad con totales agregados oficiales, sus errores se mantienen en niveles aceptables. Este resultado es particularmente relevante en contextos institucionales donde es necesario garantizar la coherencia entre diferentes niveles de agregación estadística. Así, el modelo FH bajo *benchmarking* ofrece una alternativa sólida y compatible con las exigencias de la estadística pública oficial, donde el *benchmarking* suele ser un requerimiento necesario.

El uso de una escala semilogarítmica permite visualizar mejor cómo pequeños incrementos en el tamaño muestral, especialmente en valores bajos, tienen un impacto significativo en la reducción del RMSE. No obstante, a partir de cierto umbral ($log_{10}(n_d) > 1{,}5$, es decir, $n_d > 30$), el RMSE del modelo FH univariante tiende a estabilizarse y tener menos fluctuaciones erráticas que el método directo, lo que indica que aumentos adicionales en el tamaño muestral generan mejoras marginales en la precisión.

En conjunto, estos resultados destacan la eficacia del enfoque modelizado para mejorar la precisión de las estimaciones en dominios con tamaños muestrales reducidos, proporcionando predicciones más estables sin sacrificar la fidelidad respecto a los datos observados.

En conclusión, el modelo FH univariante demuestra ser superior al método directo tanto en precisión, evidenciada por un menor RMSE, como en estabilidad, reflejada en una menor variabilidad del error. Su capacidad para mantener un bajo RMSE con tamaños muestrales relativamente pequeños lo hace una opción idónea en situaciones con limitaciones de datos o recursos. En contraste, el método directo, con su alta variabilidad y menor precisión, se presenta como una alternativa menos confiable,

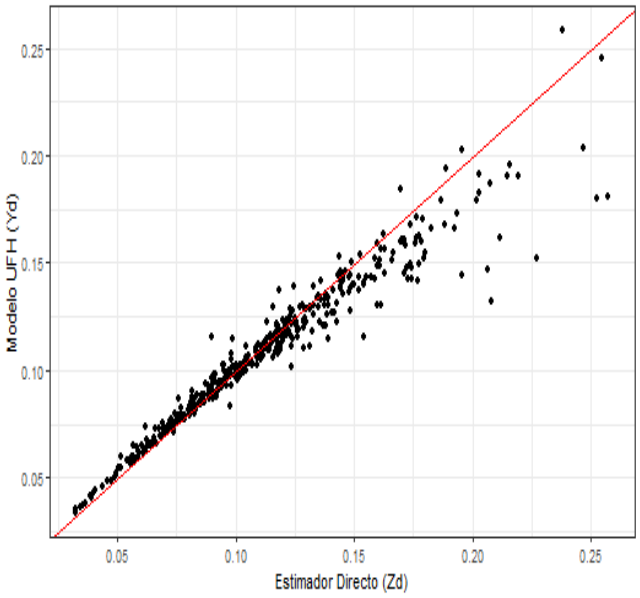


Figura 5.3: Estimadores directos frente a los composicionales del modelo FH univariante bajo alr para la proporción de ninis respecto a la población de 16 a 24 años.

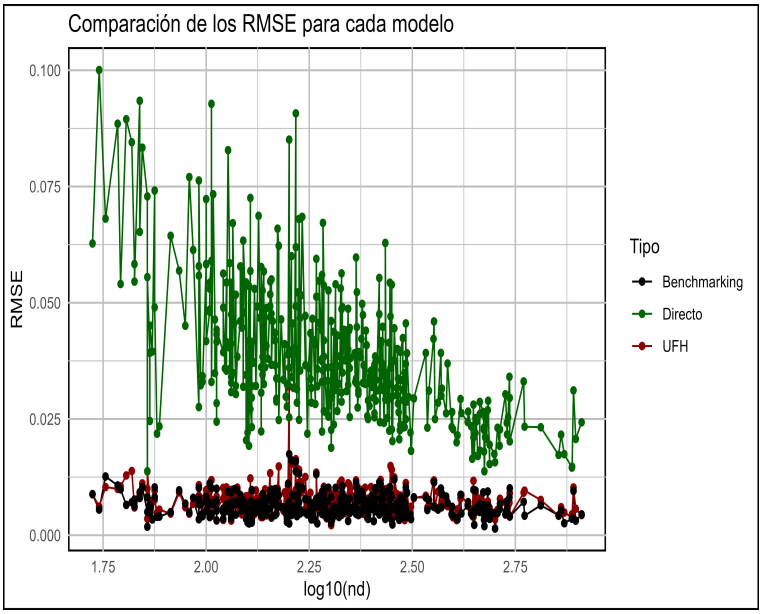


Figura 5.4: RMSE de la estimación de la proporción de ninis respecto a la población de 16 a 24 años conforme aumenta el tamaño muestral (escala semilogarítmica).

especialmente en contextos donde no es viable garantizar un tamaño muestral grande.

A continuación, analizamos los errores relativos (RRMSE), definido en la ecuación (4.2), tanto de las estimaciones directas como del modelo FH univariante propuesto y su versión bajo *benchmarking*.

La comparación de los métodos a través de este gráfico (ver Fig. 5.5), en el que se presentan tanto el modelo FH univariante como su versión con *benchmarking* frente a las estimaciones directas, evidencia

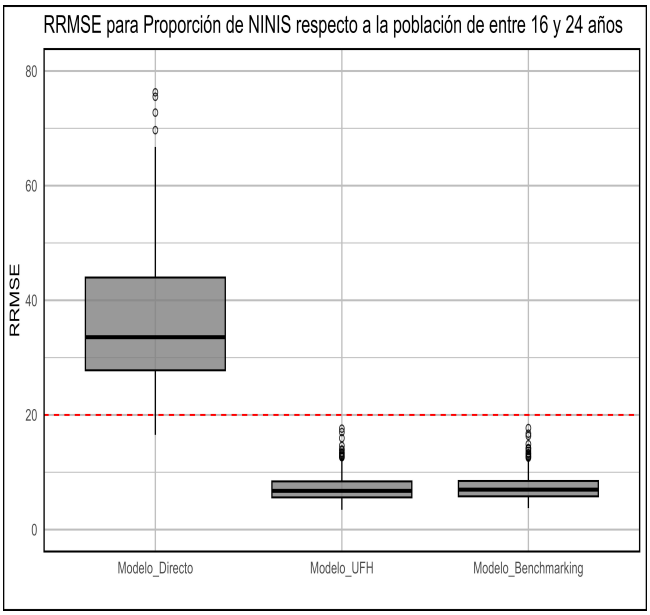


Figura 5.5: RRMSE de la proporción de ninis respecto a la población de 16 a 24 años.

la superioridad del modelo FH univariante en la estimación de la proporción de ninis. La menor altura de la caja, junto con la posición inferior de la mediana y la menor dispersión en los valores de RRMSE, indican que el modelo FH univariante proporciona estimaciones más precisas y estables. Además, este método no solo mejora la exactitud en promedio, sino que también reduce la variabilidad, lo que resulta fundamental para la toma de decisiones basada en estos datos. Cabe destacar que todas las provincias presentan un CV inferior al umbral del 20 % establecido por ONS (2006). Las conclusiones para el modelo bajo *benchmarking* son similares.

Finalmente, evaluamos la estabilidad temporal del modelo FH univariante a lo largo de los trimestres. Para ello, en el Cuadro 5.3 se presentan un estudio de estabilidad temporal que versa como sigue: para cada dominio d, $d = 1, \ldots, D$, con $D = 52$, se calcula la desviación típica de los estimadores a lo largo de los ocho trimestres, y de la distribución obtenida, se calculan los deciles de dicha desviación típica.

Así, se presentan en el Cuadro 5.3 para estos deciles de las desviaciones típicas de las predicciones realizadas con el modelo FH univariante (en la parte superior) y las estimaciones directas (en la parte inferior) para los ocho períodos de estudio, desde el primer trimestre de 2021 hasta el cuarto trimestre de 2022, calculadas a nivel provincial.

Los resultados muestran que la estabilidad temporal mejora significativamente al emplear el modelo FH univariante. En particular, se observa una reducción en la desviación típica superior al 22.5 % a medida que aumenta el tamaño muestral, lo que refuerza la ventaja del modelo propuesto sobre el estimador directo en términos de estabilidad y precisión. Este efecto se intensifica con el incremento del tamaño muestral, consolidando la superioridad del modelo FH univariante. Por su parte, el modelo bajo *benchmarking* exhibe una evolución temporal análoga, manteniendo una tendencia similar al modelo FH univariante.

Al final de este trabajo se exponen las provincias ordenados por orden de tamaño muestral de la población de 16 a 24 años para el segundo trimestre de 2022 y el valor de sus predicciones para los ninis (ver Cuadro 7.1), e idénticamente para el cuarto trimestre de 2022 en el Cuadro 7.2, por poner dos ejemplos.

	q_0	$q_{0,1}$	$q_{0,2}$	$q_{0,3}$	$q_{0,4}$	$q_{0,5}$	$q_{0,6}$	$q_{0,7}$	$q_{0,8}$	$q_{0,9}$	q_1
$\widehat{\overline{Z}}_d^{\text{FH}}$	0.010	0.015	0.019	0.020	0.022	0.024	0.028	0.029	0.032	0.034	0.052
$\widehat{\overline{Z}}_d^{\text{ben}}$	0.012	0.016	0.019	0.022	0.023	0.025	0.031	0.033	0.036	0.039	0.053
$\widehat{\overline{Z}}_d^{\text{H}}$	0.010	0.017	0.021	0.024	0.026	0.030	0.031	0.035	0.039	0.044	0.065

Cuadro 5.3: Deciles de la desviación típica temporal para los ocho períodos de estudio. Arriba, el modelo propuesto (FH univariante), $\widehat{\overline{Z}}_d^{\text{FH}}$, seguido del modelo bajo *benchmarking*, $\widehat{\overline{Z}}_d^{\text{ben}}$, y finalmente el modelo de estimaciones directas de Hájek, $\widehat{\overline{Z}}_d^{\text{H}}$.

5.1.3. Mapas de los resultados y conclusiones del Objetivo 2

Finalmente la Figura 5.6, que ilustra la distribución de la proporción de ninis en relación con la población de 16 a 24 años en las distintas provincias españolas para el cuarto trimestre de 2022. Este mapa se complementa con un segundo gráfico que muestra el CV asociado a dichas proporciones, proporcionando una visión integral tanto de la magnitud de la proporción de ninis como de la estabilidad de las estimaciones obtenidas.

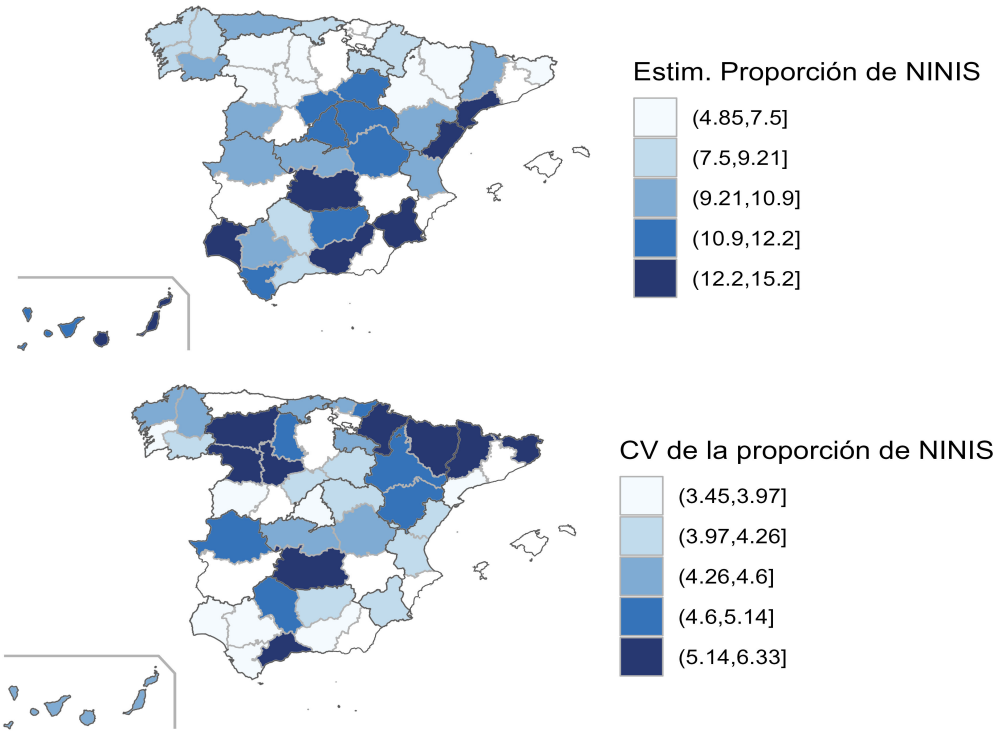


Figura 5.6: Distribución de la proporción porcentual de ninis respecto a la población de 16 a 24 años (superior) y su CV asociado (inferior) en España por provincias para el modelo FH univariante, para el T8.

En el mapa de la Figura 5.6, bajo el nombre de leyenda «Estimación de la proporción de ninis», las

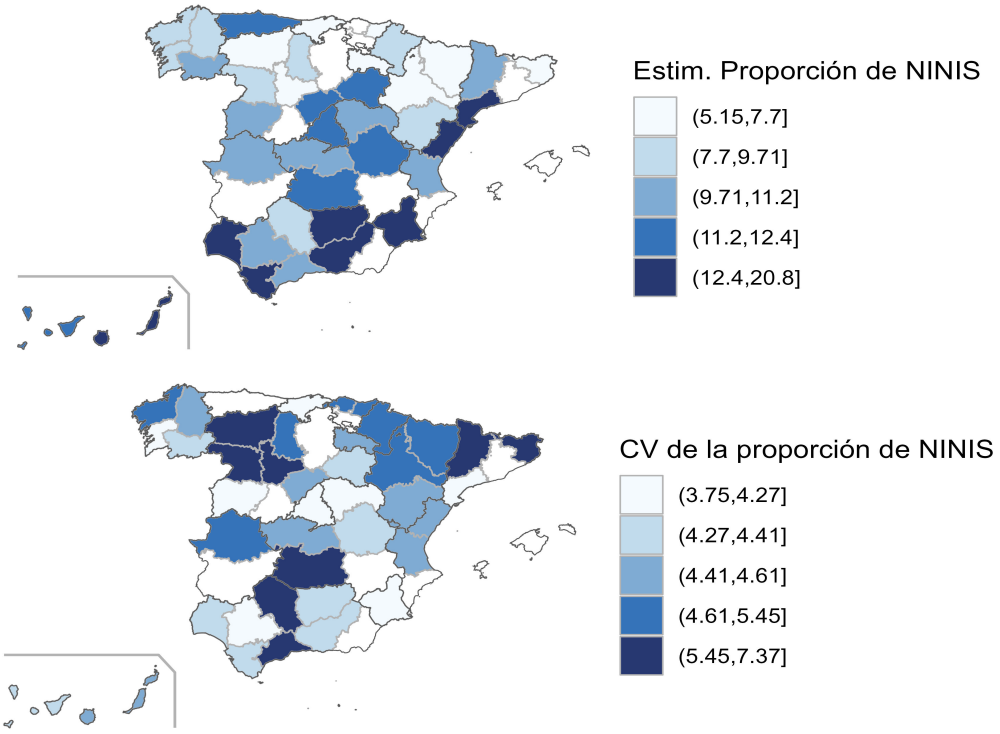


Figura 5.7: Distribución de la proporción porcentual de ninis respecto a la población de 16 a 24 años (superior) y su CV asociado (inferior) en España por provincias para el modelo bajo *benchmarking*, para el T8.

proporciones de ninis respecto a la población de 16 a 24 años oscilan entre 4.85 % y 15.2 %. El mapa de la Figura 5.7, situado en la parte inferior, bajo el nombre de leyenda «Coeficiente de variación de la proporción de ninis», muestra el coeficiente de variación de las estimaciones de la proporción de ninis, cuyas valoraciones varían desde 3.34 % hasta 6.5 %. Es relevante señalar que las provincias con mayor proporción de ninis no siempre coinciden con aquellas que presentan mayor variabilidad, lo que sugiere que algunas regiones, a pesar de tener altas proporciones de ninis, también muestran estimaciones más estables, y viceversa.

Asimismo, es destacable que las áreas con mayor concentración de ninis se localizan principalmente en el sur y levante de España, así como en la denominada España Vacía, siendo Soria, Ciudad Real y Guadalajara ejemplos representativos de esta situación, con estimaciones de ninis respecto a la población de 16 a 24 años considerablemente elevadas. Se pueden obtener conclusiones similares en el caso del modelo bajo *benchmarking*, aunque, a diferencia del modelo FH univariante, este presenta una mayor variabilidad en cuanto a las estimaciones de la proporción de ninis debido a los parámetros de calibración empleados.

Con todo y como conclusiones de esta aplicación, a lo largo de este apartado se ha evidenciado cómo, para un mismo problema, como es la estimación de la proporción de ninis respecto a la población de 16 a 24 años, existen enfoques metodológicos que mejoran considerablemente las varianzas y los errores relativos. El estimador directo, siendo el más sencillo de aplicar, suele generar problemas significativos cuando los tamaños muestrales son pequeños. Esta limitación se manifiesta en la alta variabilidad y los elevados errores relativos, lo que hace que las estimaciones directas sean menos confiables en contextos con muestras reducidas.

Para abordar estas limitaciones se emplea modelos basados en la SAE, que ha demostrado ser eficaz

en la reducción de errores relativos elevados y en la mejora de la precisión de las estimaciones cuando los tamaños muestrales son pequeños (generalmente cuando $n_d < 30$) (Morales et al. 2021; Prasad y Rao 1990; López-Vizcaíno 2014, López-Vizcaíno et al. 2015). En este estudio, se optó por un modelo FH univariante bajo una transformación alr para tratar el problema binomial, resultando en mejores predicciones para muestras pequeñas y mayor estabilidad en las estimaciones. Los errores relativos obtenidos con este modelo se mantuvieron por debajo del valor máximo admisible de publicación, es decir, se consiguió que $CV \leq 20\,\%$, donde el límite del 20 % es el considerado por la ONS (ONS, 2006).

Adicionalmente, estudios previos han mostrado que la transformación alr produce resultados superiores en comparación con modelos que no emplean dicha transformación. Si bien es cierto que el modelo sin la transformación es más simple, asume una probabilidad lineal, lo cual no es adecuado, ya que esta podría escapar del intervalo $[0, 1]$. Sin embargo, los resultados son aún satisfactorios, con los CVs generalmente por debajo del 20 %. No obstante, se optó por el modelo bajo transformación alr por coherencia y mejores resultados, dado que la transformación alr permite manejar adecuadamente la naturaleza composicional de los datos, lo que mejora la precisión y la estabilidad de las estimaciones (López-Vizcaíno 2014, López-Vizcaíno et al. 2015).

En el caso específico de este estudio, no fue necesario emplear información auxiliar, ya que el modelo con mejores resultados fue aquel cuyas covariables provinieron de la EPA. No obstante, para estudios más detallados en la búsqueda de variables auxiliares y covariables, se recomienda consultar los trabajos de Lombardía et al. (2017, 2018), que ofrecen una guía exhaustiva sobre cómo seleccionar y utilizar variables auxiliares en modelos de estimación en áreas pequeñas.

Además de la mejora en la precisión y estabilidad, se observó que los resultados obtenidos con el modelo FH univariante bajo la transformación alr son consistentemente estables a lo largo del tiempo y, en algunos casos, superiores a los obtenidos mediante estimadores directos. Esto refuerza la conclusión de que el modelo FH univariante con la transformación alr es adecuado y cumple con los objetivos planteados, proporcionando coeficientes de variación significativamente bajos, al igual que dicho modelo bajo *benchmarking*.

Por otro lado, los resultados también se ven reforzados mediante el uso del modelo propuesto bajo *benchmarking*, el cual mejora la estimación predicha, aunque a costa de un ligero incremento en la variabilidad temporal y el CV. Sin embargo, a nivel autonómico, los totales coinciden con los publicados por el INE, lo que aporta valor añadido al modelo.

En resumen, el empleo de metodologías de estimación en áreas pequeñas, y específicamente el modelo FH univariante bajo la transformación alr, ha demostrado ser una solución eficaz para la estimación de la proporción de ninis en contextos de tamaños muestrales reducidos y la importancia de la conformidad en estos contextos. Este enfoque no solo mejora la precisión de las estimaciones, sino que también asegura una mayor estabilidad temporal, cumpliendo con los estándares de calidad y precisión requeridos para la toma de decisiones informadas y la formulación de políticas públicas basadas en datos robustos y confiables.

5.2. Uso del modelo BFH para el estudio de variables laborales a nivel municipal

En esta sección se aplica el modelo mixto composicional basado en el modelo BFH explicado en la subsección 4.2.3 a los datos de la EPA en los municipios de más de 16.000 habitantes de España, desde el primer trimestre de 2021 al cuarto trimestre de 2022. Este objetivo se corresponde con el Objetivo 1 de los propuestos por el INE a la línea investigadora, explicados en el Apartado 2.2.

De acuerdo a este Objetivo 2, el objetivo son las proporciones de las tres categorías de la variable de situación laboral, es decir, ocupados, parados e inactivos en los municipios de más de 16.000 habitantes de España. El caso de los inactivos se trata como variable de referencia en el ajuste del modelo, centrándose ahora en las proporciones de parados y de ocupados. Recordemos que cada dominio está dividido en subconjuntos U_{dk}, $k = 1, \ldots, q$, definidos por la variable de clasificación de la fuerza de

trabajo, que clasifica las unidades en un número finito de categorías. A su vez, se trata de $D = 394$ dominios y ocho trimestres.

El modelo BFH se ajusta a la variable respuesta y a un conjunto de variables agregadas auxiliares significativas extraídas de registros administrativos. Las variables auxiliares a nivel de dominio obtenidas del estudio descriptivo anterior son las siguientes:

1. Parados registrados en el SEPE (*ParoR*): porcentaje de los parados registrados en el Servicio Público de Empleo (SEPE).

2. Renta Neta Media por Persona (*RNMP*): donde se toma la división con el máximo para pasarlo a tanto por uno con el objetivo de que tenga la misma escala que el resto de covariables.

3. Contratos (*Contratos*): número de contratos registrados.

4. Afiliados (*AfilSS*): porcentaje del número de afiliados a la SS bajo los datos de la EPA[1].

	$k=1$	$k=2$	$k=3$
$k=1$	1.00	-0.41	-0.86
$k=2$	-0.41	1.00	-0.11
$k=3$	-0.86	-0.11	1.00

Cuadro 5.4: Correlaciones entre los estimadores directos $\widehat{\overline{Z}}^{\mathrm{H}}_{dk}$.

El estudio se ha realizado empleando no solo los estimadores directos expuestos previamente, sino también el modelo FH univariante y BFH. En todos los casos, se han tratado los trimestres separadamente. De la misma forma, se han tratado dos casos del modelo BFH: uno bajo la transformación alr previamente mencionada, y otro sin la misma. En todos los escenarios donde sea necesario, se expondrán ambos casos, comparándolos y viendo si se produce una mejora significativa de los resultados. Como nota aclaratoria, la EPA proporciona datos con cuatro categorías, añadiendo la de menores de 16 años como una categoría adicional a las de parados, ocupados e inactivos. Así, se mantienen las tres categorías de interés y se busca que sumen unidad.

[1]Es importante señalar que, en el contexto de este análisis, se ha optado por emplear como variable auxiliar la estimación de la proporción de ocupados proveniente directamente de los microdatos extendidos de la EPA, en lugar de utilizar los registros administrativos de afiliaciones a la SS contenidos en las bases de datos auxiliares. Esta decisión no es trivial y responde a consideraciones tanto metodológicas como empíricas.

En efecto, aunque los datos administrativos de afiliación pueden parecer en principio una fuente más exhaustiva y precisa, al estar basados en registros completos y no en encuestas, presentan una limitación crítica: su definición no es plenamente compatible con la lógica del muestreo por residencia utilizada en la EPA. En concreto, las estadísticas de afiliación hacen referencia al lugar del centro de trabajo, mientras que la EPA recoge la situación laboral de las personas en su lugar de residencia habitual. Esta discrepancia genera problemas de consistencia territorial, especialmente en áreas urbanas que actúan como polos laborales para municipios colindantes, dando lugar a distorsiones en la interpretación de las tasas de ocupación a nivel provincial o municipal.

Por estas razones, se ha considerado más oportuno utilizar la proporción de ocupados estimada por la EPA, aun siendo una variable muestral y no un dato censal. Esta elección, lejos de debilitar el modelo, permite asegurar una mayor coherencia conceptual con el resto de variables utilizadas y garantiza que la información auxiliar se encuentre referida al mismo marco poblacional y territorial. Si bien esta estimación introduce algo de variabilidad adicional, la ganancia en precisión obtenida por la homogeneidad en la definición supera claramente este inconveniente. Esta decisión también está en línea con recomendaciones metodológicas presentes en la literatura sobre estimación en áreas pequeñas, donde se prioriza la coherencia definicional y la compatibilidad semántica entre fuentes sobre la exhaustividad formal (Rao y Molina 2015).

*Caso **sin** alr*		Lím. Inf.	Beta	Lím. Sup.	Error Estándar	p-valor
z_1	RNMP	0.74	0.79	0.84	0.03	0.00
	Contratos	-0.00	-0.00	0.00	0.00	0.36
	AfilSS	0.25	0.30	0.35	0.02	0.00
z_2	Intercept	0.04	0.07	0.09	0.01	0.00
	ParoR	0.50	0.67	0.84	0.09	0.00
	RNMP	-0.12	-0.08	-0.04	0.02	0.00
*Caso **con** alr*		Lím. Inf.	Beta	Lím. Sup.	Error Estándar	p-valor
y_1	RNMP	-1.85	-1.60	-1.34	0.13	0.00
	Contratos	-1.68	-1.13	-0.59	0.28	0.00
	AfilSS	2.35	2.65	2.96	0.16	0.00
y_2	Intercept	-1.86	-1.37	-0.88	0.25	0.00
	ParoR	5.81	8.78	11.75	1.51	0.00
	RNMP	-2.75	-2.00	-1.24	0.38	0.00

Cuadro 5.5: Estimadores de los parámetros del modelo e intervalos de confianza al 95 % con sus p-valores para las variables objetivo y_1 (tabla superior) e y_2 (tabla inferior).

En primer lugar, se encuentra el método directo, basado en el estimador de Hájek descrito en el Apartado 3.1.1, cuya expresión se denota por $\widehat{\overline{Z}}_d^{\mathrm{H}}$. En segundo lugar, se utiliza el modelo FH bivariante con transformación alr, explicado en el Apartado 4.2.3.2. Este modelo opera inicialmente sobre la transformación log-ratio aditiva de la variable de interés, ajustando el modelo sobre la variable transformada. Posteriormente, para obtener estimaciones comparables con el método directo, se aplica la inversa de la transformación alr, obteniéndose así el estimador modelizado:

$$\widehat{\overline{Z}}_d^{\mathrm{BFH}} = \mathrm{alr}^{-1}(\widehat{y}_d).$$

5.2.1. Estudio descriptivo de los datos y análisis del modelo

Primero de todo, se exponen las correlaciones, véase el Cuadro 5.4, entre las categorías del estimador directo $\widehat{\overline{Z}}_{dk}^{\mathrm{H}}$, es decir, para las proporciones de ocupados ($k = 1$), parados ($k = 2$) e inactivos ($k = 3$). Se puede observar el comportamiento esperable: correlaciones negativas entre las variables.

Como se observa, recordando que y_1 se refiere a los ocupados e y_2 a los parados, no es extraño ver una correlación positiva entre los parados registrados en el SEPE para el caso de y_1 y negativa para y_2. En el caso de los afiliados a la SS, el razonamiento es análogo para los ocupados y negativo para los parados. Además, es esperable las relaciones negativa y positiva de la renta neta media por

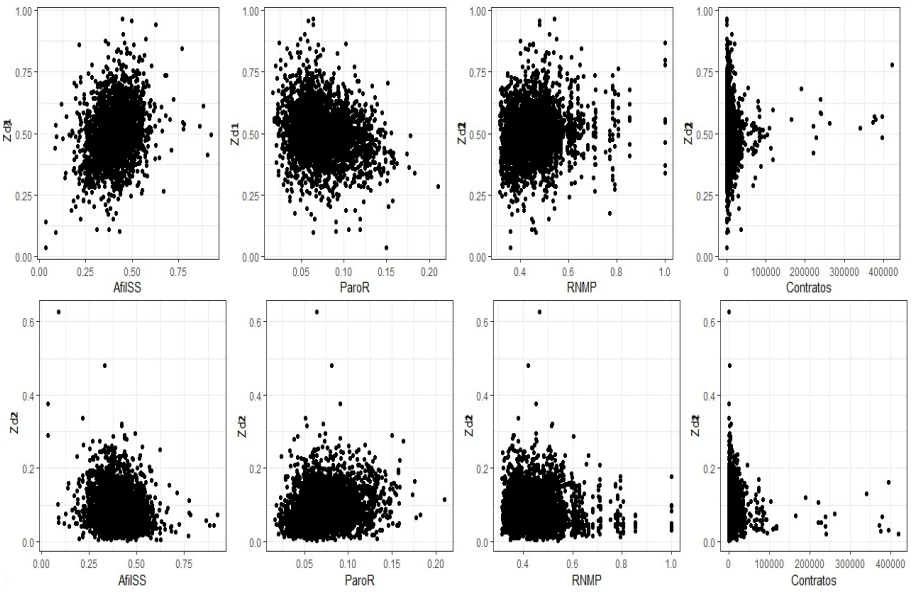


Figura 5.8: Gráficos de dispersión entre la variable objetivo ocupados, en la primera fila, y parados, en la segunda fila.

persona para los parados y ocupados, respectivamente. En el caso de los contratos, se observa que la correlación es prácticamente nula, pero se incluye porque en el modelo, tanto bajo transformación alr como sin la misma, para los ocupados ($k = 2$), es una variable auxiliar significativa. Estas relaciones se expondrán en gráficos de dispersión (Fig. 5.8) para ver su comportamiento. En la primera fila están los ocupados y en la segunda fila están los parados o desempleados.

En el Cuadro 5.5 se presentan los resultados del modelo BFH ajustado para el cuarto trimestre del año 2022. El análisis se estructura en dos bloques: el primero muestra los resultados sin aplicar transformación alr; el segundo, con la transformación alr descrita previamente. El objetivo es valorar comparativamente si la transformación mejora la capacidad explicativa y la significación estadística de las covariables.

Se consideran dos variables objetivo: z_1 (o y_1 tras la transformación), correspondiente a la proporción de ocupados, y z_2 (o y_2), correspondiente a la proporción de parados. Para cada una de ellas, se muestran los estimadores de regresión $\widehat{\beta}_{kl}$ con sus respectivos intervalos de confianza al 95 %, errores estándar y $p-$valores, asociados a la prueba de hipótesis nula $H_0 : \beta_{kl} = 0$, donde $k = 1, 2$ indica la variable objetivo y l recorre las distintas covariables.

Sin la transformación alr, en el caso de z_1, se observa que RNMP y AfilSS resultan estadísticamente significativas, mientras que la variable Contratos no lo es ($p = 0{,}36$). Para z_2, todas las covariables, incluyendo el intercepto, presentan una fuerte significación. En particular, la tasa de paro registrada ($ParoR$) muestra un efecto positivo pronunciado, mientras que RNMP actúa como factor inverso.

Con la transformación alr se aprecia una mejora sustancial en la significación de las covariables. Contratos, que no era significativa sin transformar, presenta ahora un coeficiente negativo y significativo para y_1. Asimismo, aumentan las magnitudes de los efectos estimados y se reduce la varianza relativa, indicando mayor estabilidad y poder explicativo del modelo. El intercepto de y_2 se mantiene negativo y significativo, lo cual es esperable tras la transformación log-ratio.

Así, la transformación mejora la calidad estadística del modelo bivariante FH, potenciando la detección de relaciones significativas entre las covariables y las variables respuesta. Esto respalda el uso de metodologías composicionales en estos contextos con proporciones, como es el caso del análisis conjunto de empleo y desempleo a nivel territorial.

En general, se observa que las covariables son significativas en el modelo BFH cuando se emplea la transformación alr. Esta es una de las ventajas del modelo propuesto, ya que permite una mejor interpretación de los resultados y mayor estabilidad en la estimación de parámetros.

5.2.2. Diagnosis del modelo

En cuanto a los residuos estandarizados (Fig. 5.9), se observa un comportamiento más razonable para los tres estimadores, con una mayor presencia de valores atípicos solo en el caso de los parados. Los residuos bajo la transformación alr son prácticamente insesgados en cuanto a la mediana y no presentan demasiados problemas con los valores atípicos, estando la mayoría de los residuos entre los valores -3 y 3.

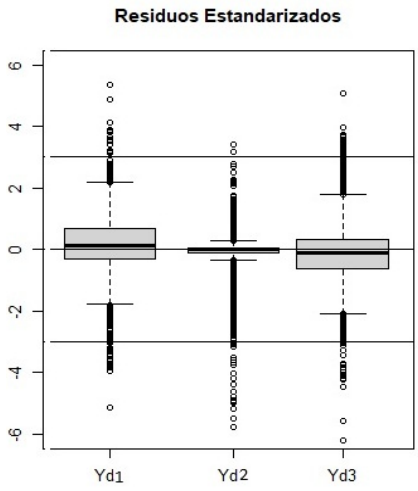


Figura 5.9: Residuos estandarizados del modelo BFH con alr para la proporción de ocupados, parados e inactivos, respectivamente.

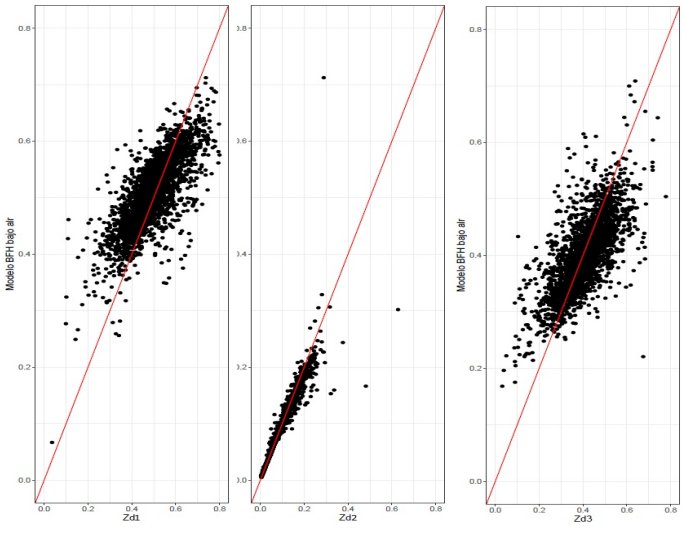


Figura 5.10: Estimadores directos frente a los composicionales del modelo BFH bajo alr para la proporción de ocupados, parados e inactivos.

Posteriormente, en la Fig. 5.10 se muestran las estimaciones composicionales frente a las estimaciones directas de las proporciones de ocupados (izquierda), parados (centro) e inactivos (derecha). Se observa que los estimadores basados en modelos toman valores de manera bastante simétrica en torno a las estimaciones directas. Como los estimadores directos de proporciones son prácticamente insesgados, lo observado sugiere que los estimadores composicionales comparten parcialmente esta propiedad. Esto es también lo esperado como ya se vio en la sección 5.1.

En la Fig. 5.11 se muestra cómo evoluciona el error a medida que aumenta el tamaño de la muestra en términos del RMSE para los casos del estimador directo, el modelo FH univariante y el BFH, para el cuarto trimestre de 2022. Se debe tener en cuenta el caso BFH sin y con alr. Como es evidente, el uso del modelo BFH bajo alr, representado en rojo, tiene un menor error para todo tamaño de muestra y además menor variabilidad en comparación con el estimador directo. La primera gráfica hace referencia a la proporción de ocupados, y la segunda a la de parados.

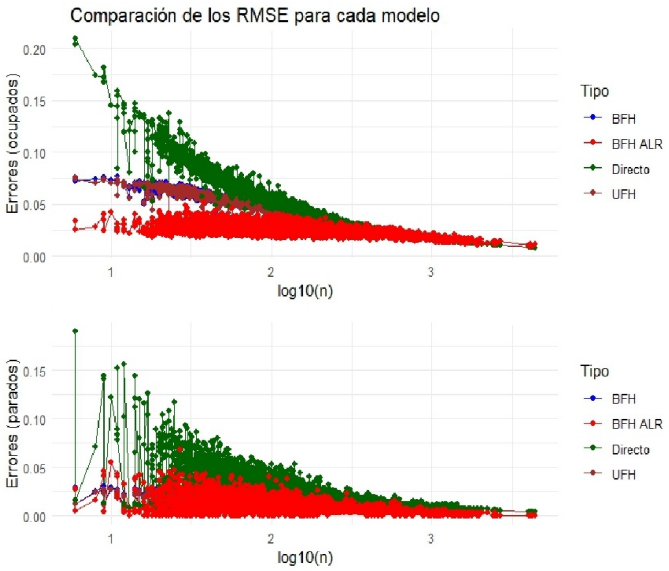


Figura 5.11: RMSE de la estimación de la proporción de ocupados (gráfico superior) y parados (inferior) a medida que aumentamos el tamaño de muestra (escala semilogarítmica)

Además, se observa cómo el RMSE de los modelos a considerar se comporta dentro de lo esperado conforme aumenta el tamaño muestral, para los estimadores directos y FH univariante (detallado específicamente en la Fig. 5.11), así como para los estimadores composicionales bajo el modelo BFH, tanto sin como con transformación alr.

Finalmente, se muestran el RRMSE para las estimaciones de las proporciones de parados y ocupados, así como para la tasa de desempleo, en las Figs. 5.12 y 5.13, respectivamente. Se observa en ambos casos una gran mejora del BFH en contraste con el estimador directo, especialmente en el caso de la proporción de parados. En cuanto a la estimación de las proporciones, se observa que la mayoría de los municipios están por debajo de la frontera del 20 % de CV (ONS, 2006).

En cuanto a la RRMSE de la tasa de paro (Fig. 5.13), descrito su MSE en la ecuación (3.2), con el modelo BFH bajo alr mejora considerablemente, lo que lleva a concluir que los predictores construidos bajo el modelo BFH con la transformación alr son una buena alternativa a los estimadores sin dicha transformación, ya sean directos o usando el modelo FH univariante. En cuanto al modelo FH univariante, esto se corrobora con estudios preliminares, que mostraron menores RRMSE con el BFH bajo transformación alr que con el caso FH univariante, así como un modelo con mayor significación en sus covariables.

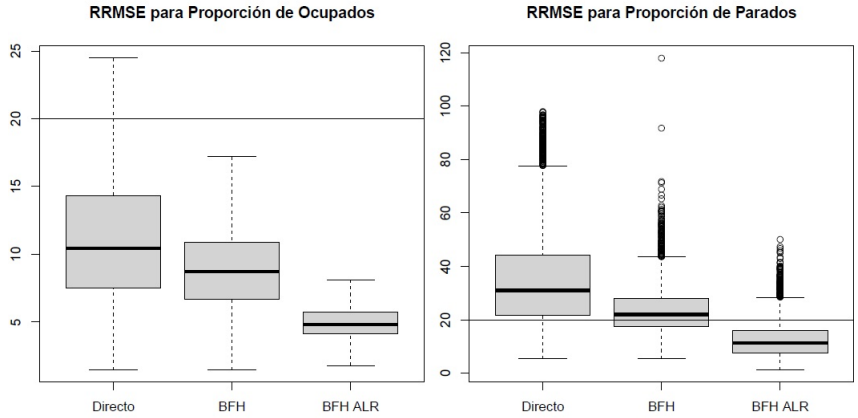


Figura 5.12: RRMSE de la proporción de ocupados y parados.

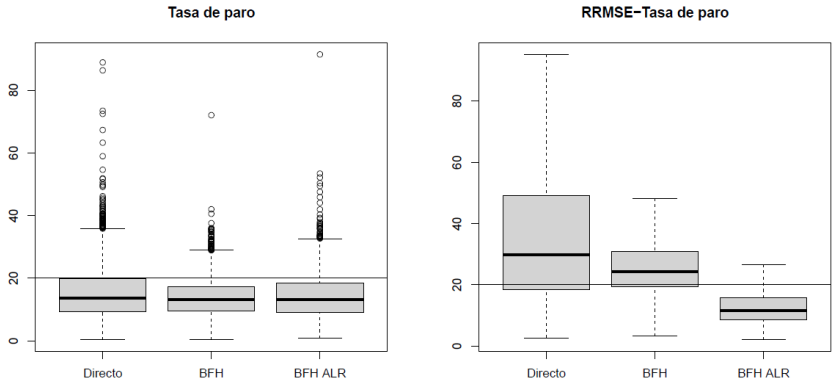


Figura 5.13: Boxplot de la estimación de la tasa de paro (izquierda) y RRMSE de la tasa de paro (derecha).

Y así, se exponen algunos municipios ordenados por tamaño de muestra para el cuarto trimestre de 2022 y el valor de sus predicciones, tanto para los parados y ocupados (Cuadro 5.6) como para los inactivos y la tasa de desempleo (Cuadro 5.7).

Como en el caso FH univariante, se estudia la estabilidad del modelo BFH frente al tiempo (en este caso, los trimestres). Se incluyen en el Cuadro 5.8 los deciles de las desviaciones típicas (RMSE) de las predicciones realizadas bajo el modelo BFH (arriba) y las estimaciones directas (abajo) para los ocho periodos de estudio, que abarcan desde el primer trimestre de 2021 hasta el cuarto del 2022, calculadas a nivel municipal. Se observa que la variabilidad temporal de las predicciones de la proporción de ocupados, parados e inactivos se reduce considerablemente. Además, la mejora es similar para las tasas de paro, lo que indica que el modelo BFH propuesto ofrece resultados más robustos a lo largo del tiempo.

Municipios	n_d	$\widehat{N}_d$	$\widehat{Z}_{d1}^{H}$	$\widehat{Z}_{d1}^{BFH}$	$\widehat{CV}_{d1}^{H}$	$\widehat{CV}_{d1}^{BFH}$	$\widehat{Z}_{d2}^{H}$	$\widehat{Z}_{d2}^{BFH}$	$\widehat{CV}_{d2}^{H}$	$\widehat{CV}_{d2}^{BFH}$
Madrid	3623	2758537	1532628	1324891	1.75	2.34	187564	191373	7.19	2.17
Cáceres	448	78156	41911	41039	5.09	3.44	5901	5918	19.10	5.81
Fuenlabrada	211	147322	85061	85412	6.98	3.50	9554	9524	29.78	9.85
Marbella	145	140030	73117	72303	9.36	4.24	16095	15652	28.18	11.10
Realejos, Los	112	43945	19534	22662	12.37	3.24	6340	6088	27.03	11.08
Salt	92	42763	24771	24746	11.17	3.96	1870	1902	50.59	17.48
Santurtzi	75	33560	15919	17249	13.74	3.62	2024	2035	49.02	15.28
Novelda	62	44163	23591	26172	12.92	3.43	1447	1465	69.73	28.11
Galapagar	49	25313	17836	15648	10.67	3.16	635	700	98.63	19.21
Villavic. De Odón	36	29648	13054	13034	21.43	5.20	570	584	99.89	25.09
Ibi	24	15197	10408	7300	19.93	4.52	1740	1795	59.84	18.99
Alfàs Del Pi, L'	5	3981	2458	2514	41.87	3.46	0	151		15.27

Cuadro 5.6: Estimadores directo y basado en el modelo BFH para los totales de ocupados y parados en algunos municipios junto a sus coeficientes de variación porcentuales

5.2.3. Mapas, resultados numéricos principales y conclusiones del Objetivo 1

Por último, se resumen todos los valores obtenidos en el estudio a través de mapas por municipios. En este caso, se considera el cuarto trimestre de 2022. Se presentan tanto la estimación como el coeficiente de variación para los cuatro casos de interés: la proporción de ocupados (Fig. 7.1), la proporción de parados (Fig. 7.2), la proporción de inactivos (Fig. 7.3), y finalmente, la tasa de desempleo (Fig. 7.4). Dichas figuras se insertan en el anexo por razones de espacio.

Con todo, y al igual que se ha ido observando en la resolución del Objetivo 2 del INE, a lo largo de la presente sección se ha abordado el problema de la estimación de la proporción de parados, ocupados e inactivos, así como de la tasa de paro, utilizando el modelo FH para el caso bivariante, foco del Objetivo 1. Si bien, en la mayoría de los casos, los distintos métodos conducen a soluciones similares en términos de estimación, se observa una considerable variabilidad en cuanto a las varianzas y los errores relativos asociados a cada enfoque. El estimador directo, siendo el más sencillo de aplicar, presenta limitaciones evidentes en situaciones donde los tamaños muestrales son pequeños, lo que genera elevados errores relativos y una mayor variabilidad en los resultados obtenidos.

Frente a este desafío, nuevamente se puede emplear la metodología de estimación en áreas pequeñas, la cual ofrece soluciones efectivas a los problemas derivados de los elevados errores relativos en muestras pequeñas (Morales et al. 2021; Prasad et al. 1990). Esta metodología se ha mostrado particularmente útil cuando el tamaño de la muestra es inferior a 30, una situación común en estudios de estimación a nivel local o municipal. En este contexto, se ha optado por la implementación de un modelo BFH

Municipios	n_d	$\widehat{N}_d$	$\widehat{Z}_{d3}^{H}$	$\widehat{Z}_{d3}^{BFH}$	$\widehat{CV}_{d3}^{H}$	$\widehat{CV}_{d3}^{BFH}$	$\widehat{R}^{H}$	$\widehat{R}^{BFH}$	$\widehat{CV}^{H}$	$\widehat{CV}^{BFH}$
Madrid	3623	2758537	1038345	1242273	2.38	3.19	10.90	12.62	6.91	2.90
Cáceres	448	78156	30345	31200	6.52	6.75	12.34	12.60	18.13	6.82
Fuenlabrada	211	147322	52708	52386	10.03	10.45	10.10	10.03	28.64	9.97
Marbella	145	140030	50818	52074	12.33	11.88	18.04	17.80	26.29	10.64
Realejos, Los	112	43945	18071	15194	13.09	11.54	24.50	21.18	23.77	10.16
Salt	92	42763	16122	16114	15.27	17.92	7.02	7.14	48.73	18.90
Santurtzi	75	33560	15616	14276	13.89	15.70	11.28	10.55	46.73	14.30
Novelda	62	44163	19124	16526	15.60	28.32	5.78	5.30	68.44	28.36
Galapagar	49	25313	6841	8965	24.13	19.47	3.44	4.28	97.63	19.24
Villavic. De Odón	36	29648	16024	16030	18.33	25.62	4.19	4.29	97.95	27.05
Ibi	24	15197	3049	6103	41.03	19.52	14.33	19.73	55.05	17.28
Alfàs Del Pi, L'	5	3981	1523	1316	60.71	15.66	0.00	5.67		14.84

Cuadro 5.7: Estimadores directo y basado en el modelo BFH para los totales de inactivos y la tasa de paro en algunos municipios, junto a sus coeficientes de variación porcentuales

bajo una transformación alr de los datos composicionales, lo cual no solo ha conducido a mejores predicciones en muestras pequeñas, sino que también ha aportado una mayor estabilidad temporal en las estimaciones. Este enfoque ha resultado en errores relativos significativamente más bajos que los obtenidos mediante los estimadores directos, alcanzando valores por debajo del umbral máximo admisible de publicación, establecido en un coeficiente de variación inferior al 20 % (ONS, 2006).

Cabe destacar que la transformación alr ha demostrado ser un factor clave en la mejora de los resultados obtenidos, ya que proporciona estimaciones más precisas y estables en comparación con los métodos tradicionales que no aplican esta transformación. Este hallazgo subraya la importancia de considerar transformaciones adecuadas para los datos composicionales, lo cual tiene un impacto directo en la fiabilidad de las estimaciones, especialmente cuando se trabaja con pequeñas muestras y estructuras de datos complejas.

Un desafío adicional que se presenta en la estimación bajo modelos con bajos RRMSE es la identificación y el uso adecuado de información auxiliar. En este estudio, se ha recurrido a diversas fuentes de datos auxiliares, tales como los afiliados a la Seguridad Social, los parados registrados en el SEPE, la renta neta media por hogar y la información sobre contratos registrados, con el objetivo de ajustar los modelos de regresión necesarios para la estimación de las proporciones de ocupados, parados e inactivos. La correcta integración de estas variables auxiliares ha sido fundamental para mejorar la precisión y estabilidad de las predicciones.

En conclusión, se puede afirmar que el modelo BFH bajo la transformación alr ha demostrado ser un enfoque robusto y eficaz para la estimación de proporciones y tasas en áreas pequeñas, superando en estabilidad y precisión a los estimadores directos. Este modelo no solo satisface los objetivos

Estimadores BFH (alr)	q_0	$q_{0,1}$	$q_{0,2}$	$q_{0,3}$	$q_{0,4}$	$q_{0,5}$	$q_{0,6}$	$q_{0,7}$	$q_{0,8}$	$q_{0,9}$	q_1
$\widehat{\overline{Z}}_{d1}^{BFH}$	0.00	0.01	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.06	0.23
$\widehat{\overline{Z}}_{d2}^{BFH}$	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.06	0.14
$\widehat{\overline{Z}}_{d3}^{BFH}$	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.06	0.13
$\widehat{R}_d^{BFH}$	0.01	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.08	0.29
Estimadores Directos	q_0	$q_{0,1}$	$q_{0,2}$	$q_{0,3}$	$q_{0,4}$	$q_{0,5}$	$q_{0,6}$	$q_{0,7}$	$q_{0,8}$	$q_{0,9}$	q_1
$\widehat{\overline{Z}}_{d1}^{H}$	0.00	0.01	0.02	0.02	0.03	0.03	0.03	0.04	0.05	0.06	0.25
$\widehat{\overline{Z}}_{d2}^{H}$	0.00	0.02	0.03	0.03	0.04	0.05	0.05	0.07	0.08	0.10	0.22
$\widehat{\overline{Z}}_{d3}^{H}$	0.00	0.02	0.02	0.03	0.04	0.04	0.05	0.06	0.07	0.10	0.24
$\widehat{R}_d$	0.01	0.02	0.03	0.04	0.04	0.05	0.06	0.07	0.08	0.11	0.29

Cuadro 5.8: Deciles de la RMSE temporal para los ocho periodos de estudio. Arriba, el modelo propuesto (BFH bajo alr) y debajo el directo.

planteados, sino que también ofrece una herramienta adecuada para abordar los problemas inherentes a la estimación con pequeñas muestras y datos composicionales. Se establece, por lo tanto, que el modelo propuesto es apto para su aplicación en estudios similares y presenta ventajas significativas en términos de fiabilidad y consistencia temporal, en analogía a lo obtenido en el caso FH univariante para el segundo objetivo.

Capítulo 6

Conclusiones

El presente estudio ha abordado, con rigor analítico y metodológico, la problemática inherente a la estimación en áreas pequeñas en el contexto de la EPA, valiéndose de un andamiaje teórico sólido sustentado en la modelización estadística avanzada y el empleo de información auxiliar estructurada. La investigación se ha centrado en la aplicación y validación empírica de modelos mixtos, particularmente los modelos de Fay-Herriot en sus variantes univariantes, bivariantes, con especial énfasis en la transformación alr aplicada a datos composicionales. A continuación serán expuestas todas las conclusiones derivadas de este extenso trabajo y las futuras y posibles líneas de investigación de interés dentro del gran área de la estadística relativa a la investigación dentro de la inferencia y estimación en áreas pequeñas.

6.1. Conclusiones y discusión final

Los resultados obtenidos confirman que los modelos propuestos, los modelos de Fay-Herriot, superan ampliamente a los estimadores directos, tanto en términos de eficiencia como de estabilidad. La reducción del error cuadrático medio en los estimadores modelizados, en comparación con los métodos directos, constituye una evidencia empírica incontestable de la superioridad del enfoque basado en modelos. Esta disminución del error ha sido sistemáticamente observada en el estudio de las tasas de empleo, desempleo e inactividad, así como en el análisis de la proporción de individuos clasificados como «ninis» en la población de 16 a 24 años. La utilización de datos auxiliares, tales como afiliaciones a la Seguridad Social, parados registrados en el SEPE, renta neta media y contratos laborales, ha permitido refinar significativamente la precisión de las predicciones, en tanto que la integración de estos factores en los modelos ha conducido a una mayor coherencia estadística y a la reducción del error cuadrático medio.

Desde una perspectiva metodológica, la investigación ha evidenciado que la transformación alr resulta de ser una herramienta indispensable en la modelización de datos composicionales, evitando los problemas derivados de restricciones inherentes a la naturaleza de las proporciones dentro del marco de la SAE. Los modelos que han incorporado dicha transformación han exhibido menor variabilidad y una mayor robustez en sus estimaciones, lo que sugiere que su adopción es recomendable en futuras aplicaciones de este tipo. En particular, los modelos BFH bajo alr han manifestado una notable capacidad para proporcionar estimaciones estables a lo largo del tiempo, lo que los hace idóneos para la producción oficial de estadísticas periódicas en el ámbito laboral.

Además, la investigación ha demostrado que la estabilidad temporal de los modelos FH univariante y BFH es superior a la de los estimadores directos, reflejándose en la menor dispersión de los errores de predicción a lo largo de los distintos periodos evaluados. Se ha comprobado, asimismo, que la inclusión de *benchmarking* en los modelos no solo conserva la precisión de las estimaciones, sino que además proporciona una calibración adicional que mejora la interpretabilidad de los resultados en términos de

consistencia con las estadísticas oficiales.

Desde una óptica aplicada, el estudio de la distribución geográfica de las tasas de desempleo ha permitido detectar patrones regionales que habrían sido difíciles de discernir mediante estimaciones directas, dada la elevada variabilidad de estas en áreas con tamaños muestrales reducidos. El análisis de la distribución espacial de los coeficientes de variación ha puesto de manifiesto que los modelos propuestos cumplen con lo esperado en el contexto español en la estimación de indicadores socioeconómicos, consolidando su aplicabilidad en contextos de estadística oficial.

En lo concerniente a las implicaciones prácticas, los resultados obtenidos en esta investigación ofrecen una contribución sustantiva a la literatura sobre la SAE y su aplicación en el ámbito de las encuestas laborales. La metodología desarrollada no solo es extrapolable a otros contextos en los que se requiera la estimación de parámetros en dominios pequeños en cuanto a tamaño muestral se refiere, sino que también sienta las bases para la futura implementación de estos modelos en sistemas de producción estadística de gran escala. La validación empírica realizada confirma que la combinación de modelos mixtos específico para áreas pequeñas con información auxiliar apropiada es una estrategia eficaz para mejorar la precisión de las estimaciones en escenarios con limitaciones muestrales.

En definitiva, el presente estudio ha demostrado que la modelización en áreas pequeñas mediante el uso de modelos de Fay-Herriot representa un avance significativo en la inferencia estadística aplicada a encuestas laborales. La integración de la transformación alr ha resultado ser un factor clave en la mejora de la precisión y estabilidad de las estimaciones, constituyendo un hallazgo de gran relevancia metodológica. Las conclusiones extraídas en este trabajo proporcionan un marco de referencia sólido para futuras investigaciones en el ámbito de la estadística oficial, abriendo nuevas líneas de exploración en el desarrollo de metodologías aún más refinadas para la estimación de indicadores socioeconómicos en dominios de pequeña escala.

6.2. Líneas futuras de investigación en la SAE

A continuación se describen diversas líneas de investigación futuras que pueden ampliar el campo de la estimación en áreas pequeñas, y que se plantean de ideas a vista de una posible futura tesis en estadística e investigación operativa.

En primer lugar, la incorporación de datos funcionales representa una vía de gran potencial para la estimación en áreas pequeñas. Al tratar variables cuya observación es una curva o función en el tiempo, por ejemplo, series de consumo eléctrico horarias o trayectorias de temperatura, se podría capturar la dinámica temporal completa en cada dominio, lo que exige desarrollar versiones funcionales de los modelos Fay-Herriot capaces de manejar alta dimensionalidad y dependencias intrínsecas de las curvas (Esteban et al. 2020; Benavent y Morales 2016). Este enfoque, si bien novedoso, presenta el reto de obtener suficientes observaciones de alta frecuencia para cada área y de diseñar métodos de reducción de dimensión, como por ejemplo, descomposición en bases de tipo ondulatorio como Fourier, o análisis en componentes principales, que mantengan la información esencial sin incurrir en sobreajuste.

Otra línea que es especialmente interesante dentro del campo de la estadística aplicada a las ciencias sociales, y sobre todo a la economía, es la extensión de los modelos a un marco de supervivencia para datos censurados o truncados (Slud y Maiti 2011), lo que permitiría estimar tasas de ocurrencia de eventos, como por ejemplo, la duración del desempleo para determinados sectores sociodemográficos, en dominios con escasos datos. Al combinar enfoques de análisis de supervivencia, como el modelo de riesgos proporcionales de Cox con efectos aleatorios a nivel de área, se podría obtener predictores de tiempos de fallo o de terminación de desempleo adaptados a cada dominio. El principal desafío es integrar la censura dentro del proceso de inferencia, asegurando la coherencia con las estimaciones directas de supervivencia en muestras pequeñas.

La introducción de predictores basados en cuantiles, los modelos MQ a estos efectos, constituye otra vía de investigación de gran actualidad. En lugar de centrarse en la media condicional, los modelos M-cuantil estiman cuantiles de la distribución de la variable de interés, ofreciendo una mayor robustez ante colas pesadas o distribuciones asimétricas (Tzavidis et al. 2008; Bugallo et al. 2024b). Desarrollar

versiones a nivel de área requerirá redefinir los predictores empíricos óptimos y sus MSE, así como diseñar algoritmos de ajuste eficientes, dada la complejidad adicional que supone la función de pérdida cuantílica.

En el contexto multivariante surge el problema de la evaluación de los integrales necesarios para calcular los EBP cuando el número de variables objetivo crece. La dimensión del efecto aleatorio multivariante provoca que las integrales de predicción sean casi intratables numéricamente con métodos convencionales. Innovar en técnicas de integración de alta dimensión, como cuadraturas adaptativas (Zhong y Feng 2003) es esencial para que el modelo de Fay-Herriot multivariante sea computacionalmente viable y preciso.

Otra dirección relevante es el diseño de métodos robustos o semiparamétricos que releguen la suposición de normalidad de los errores y efectos aleatorios. Modelos basados en distribuciones t de Student o en aproximaciones semiparamétricas podrían incrementar la resistencia del estimador ante valores atípicos o violaciones de supuestos (Jiang 1996, 1997, 1998; Burgard et al. 2020).

Finalmente, la hibridación con técnicas de aprendizaje automático (Viljanen et al. 2022) promete capturar relaciones no lineales y heterogeneidades complejas en grandes bases administrativas. Esta integración requerirá, sin embargo, un esfuerzo significativo para garantizar la interpretabilidad y la coherencia con las propiedades de insesgadez y varianza mínima del paradigma de áreas pequeñas.

De la misma forma, técnicas relativas a poblaciones sintéticas pueden ser de gran uso para simulaciones de ciertos fenómenos demográficos o económicos, o generación e imputación de datos faltantes, un campo aún por explorar y que tiene mucho futuro debido a la frecuencia elevada de estas situaciones. De la misma forma, la generación de dichas poblaciones sintéticas (Rubin 1993) puede ser de excepcional ayuda a la hora de estudiar propiedades de modelos novedosos o estimadores sin tener que recurrir al coste monetario y de acceso de los datos oficiales, no teniendo que restringir la investigación a solamente simulaciones.

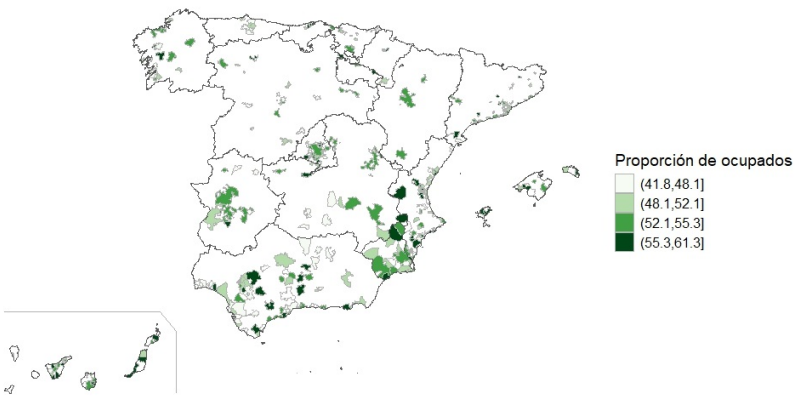
En suma, estas líneas de investigación combinan aspectos teóricos, computacionales y prácticos que, de desarrollarse exitosamente, permitirán ampliar el alcance de la estimación en áreas pequeñas a contextos de datos más diversos y complejos, consolidando su utilidad e interés matemático y aplicativo en la estadística oficial y aplicada.

Capítulo 7

Anexo 1: Resultados numéricos y mapas

7.1. Mapas del Objetivo 2

Se presentan los mapas relacionados con la distribución por municipios de la proporción de ocupados, parados, e inactivos y tasa de paro en España, del Objetivo 2.



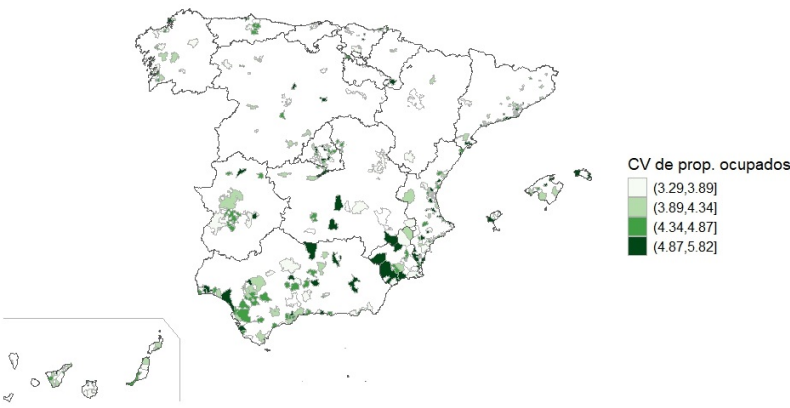
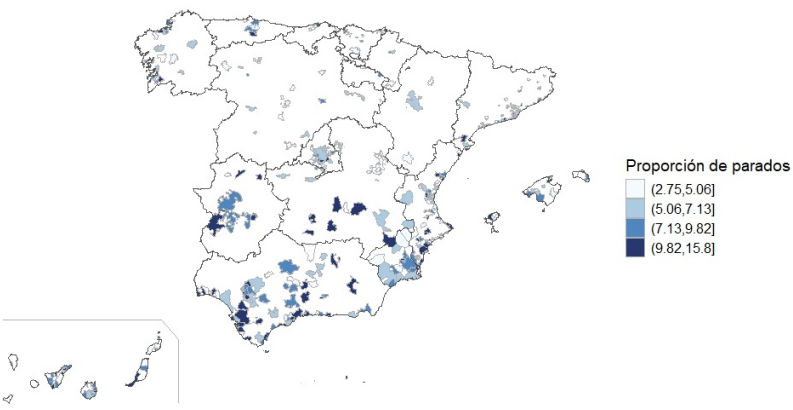


Figura 7.1: Mapas de la proporción de ocupados (izquierda) y su CV asociado (derecha) en España por municipios.



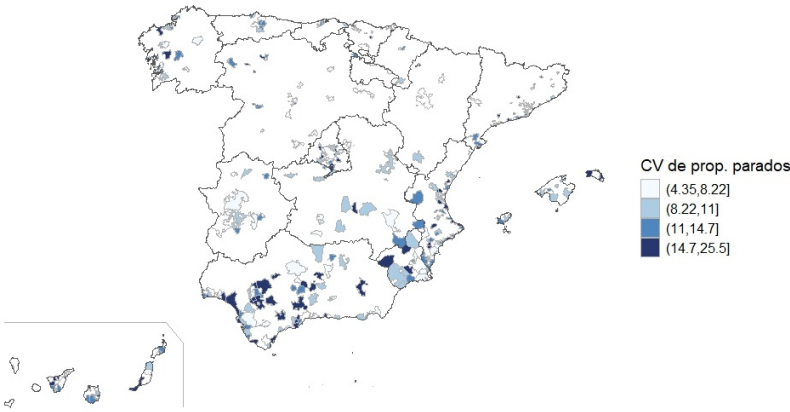
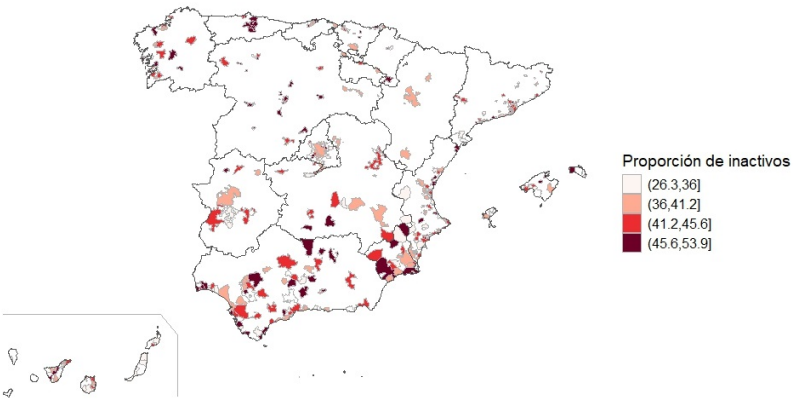


Figura 7.2: Distribución de la proporción de parados (izquierda) y su CV asociado (derecha) en España por municipios.



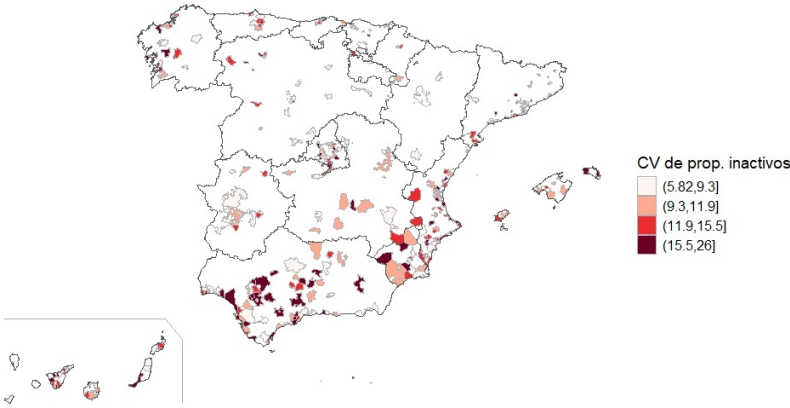


Figura 7.3: Distribución de la proporción de inactivos (izquierda) y su CV asociado (derecha) en España por municipios.

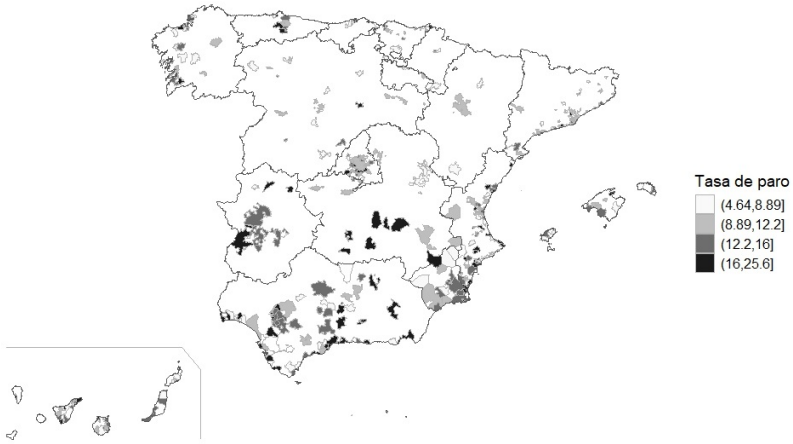


Figura 7.4: Distribución de la tasa de desempleo (izquierda) y su CV asociado (derecha) en España por municipios.

7.2. Resultados numéricos del Objetivo 2

Provincias	$n_{d,16a24}$	$\widehat{N}_{d,16a24}$	$\widehat{Z}_d^{\mathrm{H}}$	$\widehat{Z}_d^{\mathrm{FH}}$	$\widehat{Z}_d^{\mathrm{ben}}$	$\widehat{CV}_d^{\mathrm{H}}$	$\widehat{CV}_d^{\mathrm{FH}}$	$\widehat{CV}_d^{\mathrm{ben}}$
Ceuta	62	9510	1040	1064	1040	49.42	9.48	9.27
Melilla	70	10080	1785	1505	1785	47.07	7.31	8.67
Zamora	73	10992	358	388	398	75.51	18.22	18.67
Segovia	96	13846	1711	1540	1579	46.83	6.96	7.13
Lleida	97	39658	2139	2289	2291	59.79	12.34	12.35
Soria	102	8157	943	865	887	45.87	7.61	7.80
Palencia	105	12363	862	897	919	50.01	8.75	8.97
Araba/Álava	106	28128	1396	1464	1459	57.23	12.36	12.33
Ávila	112	13350	1012	1060	1087	54.82	10.38	10.64
Huesca	115	22913	2357	2280	2292	40.05	8.94	8.98
Ourense	128	19402	959	990	1001	54.83	9.94	10.06
Burgos	130	33460	2837	2846	2918	43.72	7.86	8.06
Huelva	135	55025	7621	6920	7144	33.76	6.60	6.81
Salamanca	136	25009	1097	1150	1179	50.87	14.15	14.51
Teruel	138	12863	908	926	931	46.05	9.32	9.36
Cuenca	140	17973	2111	1983	2015	41.59	8.15	8.29
León	145	30872	3363	3165	3245	42.27	7.49	7.68
Guadalajara	150	25020	1415	1486	1510	43.82	12.58	12.79
Gipuzkoa	156	67386	3954	4023	4011	50.76	8.43	8.41
Tarragona	158	78298	6933	6916	6923	36.63	8.00	8.01
Albacete	159	40872	3855	3866	3928	38.20	9.09	9.24
Lugo	163	23510	1660	1674	1694	44.86	9.27	9.37
Almería	167	70705	6962	7212	7445	40.10	10.09	10.42
Córdoba	191	78806	5042	5224	5393	34.84	11.57	11.95
Castellón/Castelló	194	53702	5451	5217	5418	35.53	6.79	7.06
Girona	197	73962	7671	7542	7549	32.00	7.78	7.79
Valladolid	202	41429	3257	3204	3284	37.77	7.74	7.93
Ciudad Real	213	44842	4172	4215	4283	32.98	8.03	8.16
Bizkaia	219	89578	8194	7940	7917	33.88	7.19	7.17
Cáceres	221	35924	4666	4398	4520	30.80	6.11	6.27
Jaén	223	60829	9657	8693	8974	28.08	6.52	6.73
Alicante/Alacant	229	172243	20437	19224	19968	32.99	6.88	7.15
Granada	233	96145	8971	9050	9342	29.43	8.88	9.17
La Rioja	249	27858	2616	2563	2616	29.61	6.72	6.86
Badajoz	256	63373	5238	5346	5495	30.68	9.78	10.06
Cádiz	256	121523	10810	10975	11329	29.48	8.50	8.78
Toledo	256	69527	8992	8664	8804	28.36	6.57	6.68
Málaga	275	171757	15224	15254	15746	28.49	7.73	7.98
Asturias	290	72430	7122	6733	7122	32.21	5.41	5.73
Zaragoza	296	83228	6437	6496	6529	30.73	7.90	7.94
Palmas, Las	300	117109	18059	16428	17627	27.50	6.42	6.89
Santa Cruz de Tenerife	303	96894	11079	10659	11437	30.54	7.36	7.89
Cantabria	313	49791	3506	3498	3506	31.37	7.16	7.18
Navarra	344	65185	5296	5291	5296	28.47	7.80	7.81
Balears, Illes	359	110934	8861	9014	8861	31.26	8.10	7.97
Coruña, A	395	79947	8045	7820	7912	26.32	5.91	5.98
Pontevedra	452	82179	8781	8522	8622	22.87	5.79	5.85
Murcia	456	161003	12021	12155	12021	22.82	7.44	7.36
Sevilla	461	185901	25520	24265	25048	20.88	5.58	5.76
Valencia/València	472	245861	19264	19336	20084	22.97	7.83	8.14
Barcelona	537	549179	62851	61655	61716	18.94	6.59	6.59
Madrid	772	634298	47640	47828	47640	19.62	6.04	6.02

Cuadro 7.1: Población de ninis por provincias, para el caso directo y el modelo FH univariante, al igual que el modelo bajo *benchmarking*, con sus coeficientes de variación porcentuales para el segundo trimestre de 2022 ordenados por tamaño muestral (de 16 a 24 años).

Provincias	$n_{d,16a24}$	$\widehat{N}_{d,16a24}$	$\widehat{Z}_d^{\mathrm{H}}$	$\widehat{Z}_d^{\mathrm{FH}}$	$\widehat{Z}_d^{\mathrm{ben}}$	$\widehat{CV}_d^{\mathrm{H}}$	$\widehat{CV}_d^{\mathrm{FH}}$	$\widehat{CV}_d^{\mathrm{ben}}$
Melilla	55	10293	2141	1358	2141	48.10	3.98	6.28
Ceuta	67	11401	1350	1563	1350	46.07	5.58	4.82
Zamora	73	11994	744	887	953	72.74	6.52	7.01
Segovia	82	14019	1843	1563	1680	48.98	4.08	4.38
Soria	91	7238	1114	834	896	50.05	4.05	4.35
Ávila	98	13173	749	853	917	58.08	7.51	8.06
Huesca	103	21216	1406	1541	1459	49.74	6.09	5.76
Lleida	106	43270	4346	4454	4536	43.99	4.45	4.53
Araba/Álava	110	25834	2524	2152	2284	50.10	4.24	4.50
Gipuzkoa	115	67132	4156	4229	4489	53.00	4.50	4.77
Palencia	115	12850	916	921	989	47.74	4.53	4.87
Teruel	122	11356	1032	1082	1025	49.12	5.18	4.90
Cuenca	125	19269	2272	2172	2180	46.11	4.47	4.49
Ourense	131	21519	2516	2334	2397	45.32	4.75	4.88
Huelva	134	60402	13723	9205	10160	30.23	3.47	3.83
León	136	32210	1898	2018	2168	52.02	5.44	5.84
Salamanca	137	26273	3250	2662	2861	42.54	3.70	3.97
Burgos	140	28388	2312	2321	2493	50.01	5.13	5.51
Lugo	144	22707	1931	1911	1963	43.11	4.67	4.79
Albacete	157	37412	3021	3122	3133	40.24	4.89	4.91
Guadalajara	159	29101	3313	3198	3209	34.94	4.48	4.50
Cáceres	169	31759	3090	3095	3151	36.40	4.84	4.92
Almería	171	70885	14686	13257	14633	33.05	4.47	4.94
Córdoba	180	78628	6752	6913	7630	36.90	4.75	5.25
Castellón/Castelló	181	56217	8120	7989	8572	32.23	4.92	5.28
Ciudad Real	182	43685	5301	5401	5420	34.55	4.89	4.91
Tarragona	193	82601	11185	10109	10296	30.94	4.07	4.14
Valladolid	196	44895	3143	3175	3411	37.57	4.72	5.07
Girona	198	69215	4678	5022	5115	37.72	6.19	6.31
Bizkaia	201	93610	4412	4540	4818	39.91	4.86	5.16
Jaén	217	60245	7631	7178	7923	29.70	3.85	4.25
Granada	223	92774	11985	11436	12623	27.82	4.18	4.62
Toledo	234	68815	6858	7013	7039	31.21	5.43	5.45
Cádiz	240	124089	14623	14595	16110	27.79	4.26	4.70
La Rioja	247	29182	2247	2281	2247	32.45	4.33	4.27
Alicante/Alacant	263	183040	32803	27884	29919	26.53	3.37	3.62
Palmas, Las	266	112528	14462	13906	13980	28.90	4.26	4.29
Málaga	270	168867	13734	14859	16401	29.71	5.55	6.12
Santa Cruz de Tenerife	285	101486	11910	12279	12345	32.00	4.83	4.86
Asturias	292	73194	8691	7993	8691	28.37	3.78	4.11
Zaragoza	295	88255	6305	6483	6141	31.60	5.38	5.10
Badajoz	300	63538	7768	7511	7646	25.56	3.63	3.70
Cantabria	303	49218	3771	3770	3771	30.33	5.15	5.15
Balears, Illes	385	118808	17260	17067	17260	25.42	3.98	4.03
Navarra	394	66041	5544	5566	5544	27.74	4.50	4.48
Coruña, A	397	81663	6900	6857	7043	26.99	4.84	4.97
Pontevedra	405	81828	6794	6802	6986	24.06	4.22	4.34
Valencia/València	445	248043	23073	23291	24990	22.61	4.88	5.24
Sevilla	475	194916	18277	18161	20046	22.49	4.50	4.97
Murcia	476	166039	21557	21312	21557	19.44	4.15	4.19
Barcelona	545	570013	62298	61212	62339	18.46	3.65	3.72
Madrid	814	662707	79354	76376	79354	20.28	3.68	3.82

Cuadro 7.2: Población de ninis por provincias, para el caso directo y el modelo FH univariante, al igual que el modelo bajo *benchmarking*, con sus coeficientes de variación porcentuales para el cuarto trimestre de 2022 ordenados por tamaño muestral (de 16 a 24 años).

Capítulo 8

Anexo 2: Software empleado

8.1. Tutoriales del uso del software aplicado

En esta sección se incluyen dos documentos en formato `R Markdown` originalmente elaborados para el Instituto Nacional de Estadística (INE) como parte de una actividad formativa dirigida a su personal técnico. Estos materiales, que contienen análisis y ejemplos prácticos con datos similares a los utilizados en este trabajo, fueron diseñados expresamente para la capacitación del INE y no para el presente TFM; aquí se presentan únicamente con ligeras adaptaciones de formato. Su incorporación pretende complementar el desarrollo teórico con casos reales de aplicación de los modelos y metodologías de estimación en áreas pequeñas, reforzando así la comprensión de su utilidad práctica en el ámbito institucional. Se expondrá primero el modelo FH univariante, propio del Objetivo 2, y después el bivariante, el BFH, propio del Objetivo 1 (y 3 por extensión).

Aplicación del modelo FH Univariante para la estimación de ninis a nivel provincial (Objetivo 2)

Alexandro Aneiros Batista

Enero de 2025

En este documento se presentan las aplicaciones a datos reales de las estimaciones de ninis, en las provincias españoles (Objetivo 2).

Usaremos un procedimiento para calcular las estimaciones de Hájek, junto con sus varianzas y coeficientes de variación (CV).

A continuación, se aborda el ajuste de un modelo Fay-Herriot a nivel de área para el caso univariante, la obtención de predictores plug-in y la estimación del error cuadrático medio (MSE) mediante bootstrap paramétrico.

Breve descripción del objetivo 2.

Como se hizo en la parte teórica, empezaremos con el segundo objetivo, o más bien con el modelo Fay-Herriot univariante (UFH). Posteriormente, proseguiremos con el modelo bivariante (BFH).

Nos enfocaremos en la estimación de los jóvenes que ni estudian ni trabajan, llamados comúnmente como ninis, al igual que del conocimiento del estado educativo y laboral de esta población.

El objetivo de este caso de uso es obtener estimaciones del total de ninis y de su proporción al respecto de la población de 16 a 24 años, véase, en la juventud y edad adulta, cuya edad está entre los 16 y los 24 años, ambos inclusive, en las provincias de España a partir de los microdatos de la Encuesta de Población Activa (EPA).

Aplicación a datos reales:

Carga de librerías y base de datos, y variables de interés.

```
# Cargar librerías necesarias

# Lista de todas las librerías necesarias
paquetes <- c(
  "foreach", "doParallel", "dplyr", "openxlsx", "MASS",
  "sae", "Matrix", "ggplot2", "gridExtra", "mapSpain", "sf", "xtable"
)

# Instalación y carga de paquetes
for (paquete in paquetes) {
```

```
  if (!require(paquete, character.only = TRUE)) {
    install.packages(paquete, dependencies = TRUE)
    library(paquete, character.only = TRUE)

    # Mensaje de confirmación
    cat("Todos los paquetes se han cargado correctamente.\n")
  }
}
```

Para obtener una mayor velocidad de cálculo, se opta por emplear cálculo pararelo con el uso de las librerías *foreach* y *doParalell*.

```
# Detectar el número de núcleos disponibles
n_cores <- detectCores()

# Crear un clúster de procesamiento paralelo con n_cores - 1
cluster <- makeCluster(n_cores - 1)

# Registrar el clúster para procesamiento paralelo
registerDoParallel(cluster)

# Mensaje de confirmación
cat("Clúster configurado con", n_cores - 1, "núcleos para procesamiento paralelo.\n")
```

```
## Clúster configurado con 7 núcleos para procesamiento paralelo.
```

Ahora, carguemos la base de datos (BBDD), que contiene tanto los estimadores directos, como sus varianzas. También estará ahí la información auxiliar de interés.

```
load('DEF_datosagregados_ninis_16a24.Rdata') # Varianzas y proporciones muestrales, entre otros.
```

Desde la BBDD, vamos a introducir las variables alr, y aprovechamos para calcular alguna variable de interés a mayores, como la tasa de desempleo directa.

```
### Evitamos los ceros ###

# Crear una copia de la base de datos original para modificaciones
datos_saery <- datosAgr

# Identificar las filas donde p.nini.16a24 es igual a cero
which(datos_saery$p.nini.16a24 == 0) -> kappa

# Sustituir los valores cero por el mínimo valor no cero de p.nini.16a24
datos_saery$p.nini.16a24[kappa] <- min(datos_saery$p.nini.16a24[-kappa])

# Comentario: Esta operación asegura que no existan valores iguales a cero en p.nini.16a24,
# evitando problemas en transformaciones logarítmicas posteriores.


######## AÑADO ALR ########

datos_saery = datos_saery %>%
```

```r
  mutate(ydi_comp = log(p.nini.16a24/(1-p.nini.16a24)),
         ydi=p.nini.16a24,
         tp_epa = pparados/(pparados+pocupados)
         )
#Se realiza la transformación alr y se define la variable objetivo Yd.

datos_saery=data.frame(datos_saery)

##### Creamos la BBDD de trabajo y limpiamos el entorno de trabajo #####

# Crear una nueva base de datos específica para ninis
dfninis <- datos_saery

# Renombrar la primera columna como "TRIM" (trimestre)
colnames(dfninis)[1] <- "TRIM"

# Guardar la base de datos de ninis en un archivo .Rdata
save(dfninis, file = "BBDD_ninis.Rdata")

# Limpiar el entorno eliminando variables temporales
rm("datosAgr", "datos_saery", "kappa", "paquete", "paquetes", "n_cores")

# Mostrar las primeras filas de la base de datos creada
head(dfninis)
```

```
##   TRIM CPRO2     inmig afiliados estudios1 estudios2 estudios3   pension
## 1  121     1  44655.21 134959.80  35750.27 133857.45 107065.80  99158.34
## 2  121     2  32889.84 147429.98  61883.33 178914.03  88751.52  87649.01
## 3  121     3 344981.15 687090.68 245046.87 916475.82 450634.85 459502.96
## 4  121     4 210756.07 278383.42 203175.01 281623.44 109325.49 157114.32
## 5  121     5   9983.40  56334.94  32060.01  75533.78  29361.32  47627.66
## 6  121     6  27720.12 255911.95 127882.68 312858.65 128536.74 184315.31
##   muestra pestudios1 pestudios2 pestudios3     pinmig      fnini     f16a24
## 1    1423  0.1292146  0.4838101  0.3869752 0.13735861 0.09245922 0.07713673
## 2    1570  0.1877819  0.5429059  0.2693122 0.08559360 0.06237309 0.09393286
## 3    2459  0.1519993  0.5684778  0.2795228 0.18345284 0.16653141 0.08975996
## 4    1681  0.3419741  0.4740146  0.1840113 0.29462414 0.16942844 0.09861456
## 5    1301  0.2340914  0.5515222  0.2143865 0.06394814 0.20633128 0.08320541
## 6    2813  0.2246401  0.5495709  0.2257890 0.04183719 0.13906113 0.08953536
##          snini  snini_16a24 snini_nocorr      p.nini p.nini.16a24 pocupados
## 1 1.056316e-05 0.0021217173 1.056316e-05 0.008485631   0.09245922 0.4972178
## 2 6.746530e-06 0.0008273776 6.746530e-06 0.006910157   0.06237309 0.4666995
## 3 1.315771e-05 0.0024750938 1.315771e-05 0.017595885   0.16653141 0.4467606
## 4 2.198931e-05 0.0036025085 2.198931e-05 0.020423533   0.16942844 0.5080439
## 5 2.350879e-05 0.0068590487 2.350879e-05 0.019848201   0.20633128 0.4396123
## 6 8.436343e-06 0.0015421016 8.436343e-06 0.014677723   0.13906113 0.4316281
##     pparados pinactivos pmenor16 p.nini_nocorr p.nini.16a24_nocorr
## 1 0.06338691  0.4393952        0   0.008485631          0.09245922
## 2 0.11500217  0.4182984        0   0.006910157          0.06237309
## 3 0.10596731  0.4472721        0   0.017595885          0.16653141
## 4 0.10512256  0.3868335        0   0.020423533          0.16942844
## 5 0.08252518  0.4778625        0   0.019848201          0.20633128
## 6 0.11803343  0.4503385        0   0.014677723          0.13906113
```

```
##   pocupados_nocorr pparados_nocorr pinactivos_nocorr pmenor16_nocorr    n
## 1        0.4972178      0.06338691         0.4393952               0 1207
## 2        0.4666995      0.11500217         0.4182984               0 1329
## 3        0.4467606      0.10596731         0.4472721               0 2091
## 4        0.5080439      0.10512256         0.3868335               0 1366
## 5        0.4396123      0.08252518         0.4778625               0 1127
## 6        0.4316281      0.11803343         0.4503385               0 2370
##         pop pop_16a24  ocupados   parados inactivos niniTotal n_nini n_16a24
## 1  273239.5  25077.11 135859.58  17319.81 120060.16   2318.61      8     121
## 2  325797.2  36094.25 152049.40  37467.39 136280.45   2251.31      8     148
## 3 1597491.7 168792.66 713696.33 169281.89 714513.45  28109.28     29     239
## 4  585205.3  70542.82 297310.00  61518.28 226377.02  11951.96     25     161
## 5  135034.9  12989.79  59363.01  11143.78  64528.12   2680.20     20     113
## 6  562049.1  59323.55 242596.15  66340.58 253112.32   8249.60     32     235
##     cvnini cvnini_nocorr cvnini_16a24 Nombre_Provincial CCAA  ydi_comp
## 1 38.30124      38.30124     49.81883       Araba/Álava  EUS -2.283971
## 2 37.58827      37.58827     46.11632          Albacete  CLM -2.710218
## 3 20.61479      20.61479     29.87443  Alicante/Alacant  CVA -1.610412
## 4 22.96016      22.96016     35.42552           Almería  AND -1.589683
## 5 24.42834      24.42834     40.13903             Ávila  CYL -1.347183
## 6 19.78875      19.78875     28.23909           Badajoz  EXT -1.823110
##          ydi    tp_epa
## 1 0.09245922 0.1130688
## 2 0.06237309 0.1976996
## 3 0.16653141 0.1917169
## 4 0.16942844 0.1714421
## 5 0.20633128 0.1580526
## 6 0.13906113 0.2147384
```

De esta forma, adelantándonos a lo que haremos posteriormente, tenemos para nuestro interés en el dataframe, *dfninis*, un total de 47 columnas, y que se resumen en:

Variables relacionadas con el estimador directo:

- **Proporción de ninis**, denotada como *ydi_comp*, con la transformación logit.
- **Varianza de la estimación de la proporción de ninis**, denotada como *snini_16a24_alr*.
- Dicha proporción pero sin alr, necesaria para la conformidad.
- **Tasa de paro**, denotada por *tp_epa*, resultado de la proporción de parados frente a ocupados y parados.

Covariables para el modelo:

- **Proporción de individuos con estudios primarios, secundarios y superiores con respecto a la población de 16 y más años**, denotadas por *pestudios1*, *pestudios2*, y *pestudios3*, respectivamente.
- **Proporciones de ocupados, parados e inactivos, con respecto la población anterior**, denotadas por *pcoupados*, *pparados* y *pinactivos*.
- **Proporción de población extranjera con respecto a la población de 16 y más años**, denotada como *pinmig.*

A su vez, dentro de la BBDD existen más variables que se han usado para el estudio. Por ejemplo:

- **TRIM**, que indica el trimestre en formato X2Y, donde X corresponde a los trimestres 1, 2, 3, o 4, e Y al año, 1, o 2 para indicar 2021 o 2022.

- **CPRO2**, que indica el código provincial, o *Nombre_Provincial*, que son los nombres de las provincias.
- Diversos totales, como *inmig*, para población inmigrante, *afiliados*, *estudios*, entre otros.
- Tamaños muestrales y poblacionales, como *muestra*, que indica la muestra total existente de la EPA, o *pop_16a24* que indica la población objetivo (jóvenes de entre 16 a 24 años).
- Entre otras, como los nombres de las provincias, las varianzas indicadas por la letra *s* inicial o los coeficientes de variación, por *cv*, o los totales de pensionistas, indicados por *pension*.

Análisis de la función *UFH.NINIS* para el cuarto trimestre de 2022.

Una vez tenemos la base de datos creada, *dfninis*, empezamos con el estudio de la proporción de NINIS por provincias españolas. La función que se encarga y que fue entregada al INE para ajustar este modelo fue la llamada *UFH.NINIS*, la cual ajusta a dicho modelo y ofrece tanto los errores, como las correlaciones y resultados finales del ajuste y los totales. En otras palabras, todo lo necesario para su puesta en uso y análisis.

En este caso, y por razones didácticas, iremos desgranando la función anterior línea a línea para ver lo que hace por dentro. Esto lo haremos para la función UFH, ya que para la BFH el algoritmo es análogo.

Primero, seleccionemos el trimestre de interés, denotado como *TAU*. En este caso, 422 por ser el 4º trimestre de 2022. A su vez, añadiremos dos variables más que nos permite la función: *excels*, para sacar los resultados también en formato Excel; y *B_boot*, que indica en número de réplicas bootstrap que usaremos para nuestros errores. Recomendamos que éstas estén entre 100 y 1000.

```
TAU = 422; B_boot = 100;

dfninis.temp <- (dfninis[dfninis$TRIM==TAU,])
```

Ahora bien, se propone desde la UDC una transformación alr, que en una dimensión se reduce a una transformación logística. Ello provoca que tendremos que cambiar ligeramente nuestros valores de la matriz de varianzas-covarianzas, como se indica en el siguiente bloque de código.

Recordemos de la teoría que, para ello, tenemos que hacer un cambio de base de la forma:

$$\widehat{\text{var}}_\pi(y_d) \approx H_0 \, \widehat{\text{var}}_\pi(z_d) \, H_0'$$

Así, tomando $z_0 = H0.el = q^{-1}1_{q-1}$, tendremos:

$$H(z_0 = q^{-1}1_{q-1}) = q(\boldsymbol{I_{q-1}} + \boldsymbol{1_{q-1}1'_{q-1}})$$

```
### Evitamos los ceros sustituyéndolos por el valor mínimo.

  sel.pprop <- dfninis.temp$snini_16a24>0
  dfninis.temp$snini_16a24[!sel.pprop] <- min(dfninis.temp$snini_16a24[sel.pprop])

### Establecemos el número de categorías totales (q),
### que son 2 (ninis) y no ninis.y construimos nuestra matriz
### de varianzas covarianzas, Ved_alr y la cambiamos por Ved por
### simplicidad.

  q = 2;

  # Nuestra z_0
  H0.el = q*(diag(q-1) + matrix(rep(1,q-1), nrow=q-1)%*%matrix(rep(1,q-1), ncol=q-1))
```

```r
Ved <- list();
D = dim(dfninis.temp)[1]
H0 <- list();
Ved_alr <- list();
snini_16a24_alr = c();

# En un bucle, creamos la matriz de varianzas-covarianzas para la transformada.
for(d in 1:D){
  H0[[d]] <- H0.el
  Ved[[d]] <- NA
  Ved[[d]] <- as.numeric(c(dfninis.temp$snini_16a24[d]))
  Ved_alr[[d]] = H0[[d]]*Ved[[d]]*t(H0[[d]])
  snini_16a24_alr[d] <- (1)*Ved_alr[[d]]
}

dfninis.temp$snini_16a24_alr <- snini_16a24_alr;
Ved <- Ved_alr

rm(H0); rm(Ved_alr); rm(H0.el)
```

Como acabamos de ver, tenemos que eliminar todo posible cero porque estaríamos, entonces, ante un caso degenerado que no es realista con nuestros datos. De esta forma, una manera de solventar esto es una solución de sustituir estos valores por otros estadísticos. En este caso, se opta por el mínimo, pero en otros casos, para ganar suavidad, se usa la media o mediana como posición más conservadora.

Ahora, definamos nuestras covariables y variables de interés de forma más sencilla.

```r
# Bloque de variables.

directy1 <- dfninis.temp$ydi_comp
vardiry1 <- dfninis.temp$snini_16a24_alr
directy2 <- dfninis.temp$ydi;
vardiry2 <- dfninis.temp$snini_16a24

# Bloque de covariables.

pestudios1 = dfninis.temp$pestudios1
pestudios2 = dfninis.temp$pestudios2
pestudios3 = dfninis.temp$pestudios3
pparados = dfninis.temp$pparados
pocupados = dfninis.temp$pocupados
pinactivos = dfninis.temp$pinactivos
pinmig = dfninis.temp$pinmig
tp_epa = dfninis.temp$tp_epa

# Las agrupamos en un data.frame.

covariables = data.frame(pestudios1, pestudios2, pestudios3,
                         pparados, pocupados, pinactivos,
                         pinmig,
                         tp_epa)

variables = data.frame(directy1, covariables)
```

Por interés y como una manera de ver nuestras covariables y su relación con la respuesta, hagamos un pequeño estudio de correlaciones.

```
# Hacemos un estudio de las correlaciones entre las variables.

correlaciones.y1 = data.frame(cor(variables)[-c(1),1])
corrs = cbind(data.frame(colnames(covariables)), correlaciones.y1)
colnames(corrs) <- c('Covariables', 'Variable Nini')
print(corrs)
```

```
##            Covariables Variable Nini
## pestudios1  pestudios1     0.2774922
## pestudios2  pestudios2     0.2642089
## pestudios3  pestudios3    -0.4612484
## pparados      pparados     0.4636113
## pocupados    pocupados    -0.1408043
## pinactivos  pinactivos    -0.2193556
## pinmig          pinmig     0.2644535
## tp_epa          tp_epa     0.4498569
```

Como podemos observar, los signos de las covariables entran dentro de lo que uno podría esperarse del grupo ninis: generalmente una correlación positiva en estudios primarios y secundarios, pero negativo en superiores. De la misma forma, tiende a haber una relación positiva con la proporción de desempleo y extranjería.

Con esto, pasamos a la regresión como tal. Para ello, emplearemos la función *mseFH* de la librería *sae*. Como regresoras, usaremos las variables anteriores. Eso sí, hemos de tener en cuenta que, por ejemplo, *pestudios* es una variable composicional. Por ende, su suma da unidad. Si incluimos tal cual las tres, tendremos una matriz singular que el algoritmo no podrá invertir. Por ende, habremos de escoger, para *q'* categorías de la variable composicional, *q'-1*.

```
# Bloque de regresión bajo UFH

fmod <- mseFH(directy1 ~
            + pestudios1
            + pestudios2
            + pocupados
            + pinactivos
            + pinmig
            - 1
            , vardiry1, MAXITER = 1e5)

fit.FH.sae1 <- fmod

fit.FH.sae1$est$fit$estcoef -> fit1

rbind(fit1) -> fitty

print(fitty)
```

```
##                  beta std.error    tvalue       pvalue
## Xpestudios1  2.266648 0.8071244  2.808301 4.980365e-03
## Xpestudios2  1.779935 0.7272844  2.447371 1.439025e-02
## Xpocupados  -4.569113 0.7553768 -6.048786 1.459416e-09
```

```
## Xpinactivos -3.688395 0.9884345 -3.731552 1.903038e-04
## Xpinmig      2.437263 0.8578082  2.841268 4.493460e-03
```

Si observamos la salida de *fitty*, todas las variables son significativas al 5% (incluso al 1% excepto por la proporción de individuos con estudios secundarios).

Si lo queremos ver un poco mejor en una tabla con los parámetros estadísticamente significativos, ejecutaríamos el siguiente bloque de código, donde mejoramos la presentación y añadimos la parte de varianza de los efectos aleatorios del modelo.

```
names2 <- data.frame(c('Primarios', 'Secundarios', 'Ocupados', 'Inactivos', 'Inmig'))
colnames(names2) <- ''
tabla_significacion_ufh <- data.frame(names2,
                                      fitty$beta -
                                        abs(qt(0.025,
                                               df = dim(dfninis.temp)[1]-1))*
                                        fitty$std.error,
                                      fitty$beta,
                                      fitty$beta +
                                        abs(qt(0.025,
                                               df = dim(dfninis.temp)[1]-1))*
                                        fitty$std.error,
                                      fitty$std.error,
                                      fmod$est$fit$refvar,
                                      fitty$pvalue,
                                      Sig=fitty$pvalue<0.05)
colnames(tabla_significacion_ufh) <- c('Covars', 'Lím. Inf.', 'Beta', 'Lím. Sup.',
                                       'Error Estándar'
                                       , 'Sigma_u^2'
                                       , 'p-valor', 'Sign.')

print(tabla_significacion_ufh)
```

```
##        Covars  Lím. Inf.      Beta Lím. Sup. Error Estándar  Sigma_u^2
## 1   Primarios  0.6462784  2.266648  3.887018      0.8071244 0.05224255
## 2 Secundarios  0.3198506  1.779935  3.240019      0.7272844 0.05224255
## 3    Ocupados -6.0855949 -4.569113 -3.052630      0.7553768 0.05224255
## 4   Inactivos -5.6727596 -3.688395 -1.704030      0.9884345 0.05224255
## 5       Inmig  0.7151408  2.437263  4.159385      0.8578082 0.05224255
##        p-valor Sign.
## 1 4.980365e-03  TRUE
## 2 1.439025e-02  TRUE
## 3 1.459416e-09  TRUE
## 4 1.903038e-04  TRUE
## 5 4.493460e-03  TRUE
```

Pasemos, pues, a la predicción y deducción de los CVs de las estimaciones. Para los EBLUP, usaremos la función del paquete *sae* llamada *eblupFH*. Los resultados son los mismos que si los sacásemos de *mseFH*, pero esta función calcula a su vez los MSE, y ello hace que se ralentice el proceso. Por ello, optamos por usar la función dedicada a lo que nos interesa.

Después, reorganizamos la matriz de covariables $\boldsymbol{X}$ y el vector de estimaciones directas, $\boldsymbol{y_d}$, para predecir los errores aleatorios y concretamente los predictores *plug-in*, deshaciendo alr.

```r
# Calculamos los EBLUPs.

eblup <- eblupFH(directy1 ~
            + pestudios1
            + pestudios2
            + pocupados
            + pinactivos
            + pinmig
            - 1
            , vardiry1, MAXITER = 1e5)

# Reorganizamos los datos.

X <- lapply(1:D, function(d) as.matrix(
    t(bdiag(as.numeric(c(pestudios1[d], pestudios2[d], pocupados[[d]],
                         pinactivos[d], pinmig[d])))))

y <- lapply(1:D,
            function(d)
              matrix(c(directy1[d])));
```

A partir de esta función, calculamos los EBLUEs y el EBLUP siguiendo las expresiones vistas en la teoría. Recordemos que estas son:

$$\widehat{\beta} = (\boldsymbol{X'}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \quad \widehat{\boldsymbol{u}} = \frac{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{u}}^{\boldsymbol{2}}}{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{u}}^{\boldsymbol{2}} + \boldsymbol{\sigma}_{\boldsymbol{d}}^{\boldsymbol{2}}}(\boldsymbol{y_d} - \boldsymbol{x_d}\widehat{\beta}),$$

y donde μ lo podemos escribir como:

$$\widehat{\mu}_d = \boldsymbol{x_d}\widehat{\boldsymbol{\beta}} + \widehat{u}_d = \boldsymbol{x_d}\widehat{\boldsymbol{\beta}} + \frac{\sigma_d^2}{\widehat{\sigma}_u^2 + \sigma_d^2}(y_d - \boldsymbol{x_d}\widehat{\boldsymbol{\beta}}) = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \sigma_d^2}y_d + \frac{\sigma_d^2}{\widehat{\sigma}_u^2 + \sigma_d^2}\boldsymbol{x_d}\widehat{\boldsymbol{\beta}}$$

Finalmente, recordemos que estamos empleando un modelo con una trasformación alr, como sigue:

$$\boldsymbol{y_d} = h(\boldsymbol{z_d}) = \log\left(\frac{z_{d1}}{1 - z_{d1}}\right), \quad d = 1, \ldots, D$$

y para deshacerla, emplearíamos:

$$\boldsymbol{z_{d1}} = \frac{\exp\{y_{d1}\}}{1 + \exp\{y_{d1}\}}, \quad d = 1, \ldots, D$$

```r
### Obtenemos de las predicciones estimadas con alr.

sigma_u <- eblup$fit$refvar; # Estimación de la varianza de los random effects.
beta_tilde <- eblup$fit$estcoef$beta # Estimación del EBLUE-beta.

u.teor <- list()

for (d in 1:D) { #EBLUP-u
  u.teor[[d]] <- (sigma_u/(sigma_u+Ved[[d]]))*(y[[d]] - X[[d]]%*%beta_tilde)
}
```

```r
  ### Deshacemos alr y sacamos los predictores plug-in.

  mudb <- pidkb <- list(); pd1 <- pd2 <- c()
  for (d in 1:D) {
    mudb[[d]] <- X[[d]]%*%beta_tilde + u.teor[[d]]
    pidkb[[d]] <- exp(mudb[[d]])/(1+sum(exp(mudb[[d]])))
    pd1 <- c(pd1, pidkb[[d]][1,]);
    pd2 <- 1-pd1
  }

  mu.aux <- sapply(mudb, function(matriz) matriz[,1])

  head(mu.aux)
```

```
## [1] -2.398420 -2.396458 -1.716384 -1.469511 -2.669758 -2.009497
```

Ahora, lo que nos resta es obtener el error cometido por la estimación, los MSE. Para ello, emplearemos la función *BOOT.compo1d*. Necesitaremos, pues, la estimación de la varianza de los efectos aleatorios y la estimación de la beta, los EBLUE.

Dicha función toma de entrada la matriz de coeficientes, X, el número de dominios, D, la matriz de estimaciones de las varianzas muestrales, *Ved*, y la de los errores aleatorios, *theta.0* $= \widehat{\sigma}_u^2$ (aquí es solo un valor), junto con los EBLUE, *beta.hat*. También se puede añadir el máximo de iteraciones, *MAX_ITER*, o el número de realizaciones bootstrap, *B_boot*.

```r
source("BOOT_compo1d.R")

  thetas.0 <- c(sigma_u)
  beta.hat <- beta_tilde

  mses = BOOT.compo1d(X, D, Ved, thetas.0, beta.hat, B = B_boot,
                      categs = 1, MAX_ITER = 1000)
```

Dentro de *mses*, obtenemos dos listas: la primera relacionada con los MSEs del EBLUP de μ_d y la segunda, la que nos importa en este caso, relativa a los MSEs del predictor *plug-in*.

Conformidad (*Benchmarking*)

Si bien el modelo remata aquí, solo faltando juntar todo lo obtenido y sacar los CV, desde el INE es de interés hacer un encaje entre los valores que se obtienen de los estimadores directos y los predichos; en otras palabras, tenemos que resolver un problema de conformidad. Tomando de dominios las provincias, y de ahí los supradominios como Comunidades Autónomas (CC.AA.), calculamos los parámetros de conformidad con las estimaciones sin alr en el directo (de ahí que las dejásemos en la base de datos desde el inicio). Se toma, de referencia, el libro "A course on small area estimation and mixed models", capítulo 3.7.1, pág. 58-60.

Recordemos, pues, que el problema se resuelve como

$$\widehat{Z}_d^{ben} = \lambda \widehat{Z}_d^{UFH}, \quad \text{donde} \quad \lambda = \frac{\sum_{d\in\Omega} \widehat{Z}_d^{dir}}{\sum_{d\in\Omega} \widehat{Z}_d^{UFH}},$$

siendo Ω nuestro supradominio.

En general, si λ está en torno a la unidad, podemos realizar la aproximación de que el MSE del estimador bajo conformidad es

$$MSE_b \approx MSE \cdot \lambda^2 .$$

```
lambdas <- data.frame(ccaa = dfninis.temp$CCAA, dir = directy2,
                      ind=pd1, CPRO2 = dfninis.temp$CPRO2)
lambdas.2 <- lambdas %>% group_by(ccaa) %>%
  summarise(L = sum(dir, na.rm = T)/sum(ind, na.rm = T))
lambdas.3 <- left_join(lambdas, lambdas.2, by = c("ccaa"))

Ved_bench <- list()
directy3 <- c(); directy4 <- c()
vardiry3 <- c();
varb <- c()

for (d in 1:D) {
    directy3[d] <- log(pd1[d]*lambdas.3$L[d]/(1-pd1[d]*lambdas.3$L[d]))
    varb[d] <- mses[[2]][d]*(lambdas.3$L[d])^2
    Ved_bench[[d]] <- varb[d]
    vardiry3[d] <- Ved_bench[[d]]
    directy4[d] <- pd1[d]*lambdas.3$L[d]
}
```

Con ello, ya obtenemos tanto las proporciones como sus errores para nuestra variable de interés, tanto sin como con *benchmarking*.

```
pred_nini <- pd1
pred_bench <- pd1_bench <- directy4

mse_nini <- mses[[2]]
mse_nini_bench <- mses[[2]]*(lambdas.3$L)^2
```

Finalmente, calculamos los CV y ponemos todos los resultados en una tabla.

```
CVdir1 <- round(100*sqrt(vardiry2)/abs(directy2), 10)

CV.fh.1 <- round(100*sqrt(mse_nini)/abs(pred_nini), 10)

CV.bench <- CV.fh.1*lambdas.3$L

output <- data.frame( Prov=dfninis.temp$CPRO2,
                      ccaa = dfninis.temp$CCAA,
                      trim=dfninis.temp$TRIM,
                      nd = dfninis.temp$muestra,
                      nini_nd = dfninis.temp$n_nini,
                      nd_16a24 = dfninis.temp$n_16a24,
                      L = lambdas.3$L,
                      DIR=round(directy2, 25),
                      Vdir=round(vardiry2, 25),
```

```
                        CVdir = CVdir1,
                        mufh = round(mu.aux, 25),
                        EBfh = round(pred_nini, 25),
                        MSEfh = round(mse_nini, 25),
                        CVfh = CV.fh.1,
                        pdbench = pred_bench,
                        msebench = round(mse_nini_bench, 25),
                        CVbench = CV.bench,
                        Nombre=dfninis.temp$Nombre_Provincial,
                        Poblacion_EPA=dfninis.temp$pop,
                        Poblacion_16a24=dfninis.temp$pop_16a24)
head(output, 3)
```

```
##   Prov ccaa trim   nd nini_nd nd_16a24        L        DIR        Vdir    CVdir
## 1    1  EUS  422 1266       9      110 1.061296 0.09769449 0.002395710 50.10107
## 2    2  CLM  422 1508      12      157 1.003677 0.08074673 0.001055860 40.24188
## 3    3  CVA  422 2594      35      263 1.072974 0.17921353 0.002259898 26.52611
##        mufh       EBfh        MSEfh     CVfh    pdbench     msebench  CVbench
## 1 -2.398420 0.08329325 1.185551e-05 4.133807 0.08839881 1.335344e-05 4.387193
## 2 -2.396458 0.08344317 1.672225e-05 4.900682 0.08375001 1.684546e-05 4.918703
## 3 -1.716384 0.15233754 3.329910e-05 3.787994 0.16345430 3.833639e-05 4.064421
##             Nombre Poblacion_EPA Poblacion_16a24
## 1      Araba/Álava      273196.5        25833.80
## 2         Albacete      326628.3        37411.67
## 3 Alicante/Alacant     1630340.1       183040.14
```

Describamos las variables de la salida:

1. **Prov**:
   Provincia a la que pertenece el dato (puede ser un código o nombre de la provincia).
2. **ccaa**:
   Comunidad Autónoma asociada a la provincia.
3. **trim**:
   Trimestre del año.
4. **nd**:
   Tamaño muestral en el análisis.
5. **nini_nd**:
   Tamaño muestral de "ninis" entre los datos analizados.
6. **nd_16a24**:
   Tamaño muestral correspondientes al rango de edad de 16 a 24 años.
7. **L**:
   Parámetro de conformidad, utilizado para evaluar la calidad o ajuste del modelo en el análisis.
8. **DIR**:
   Estimación directa para la variable de interés.
9. **Vdir**:
   Varianza asociada a la estimación directa (**DIR**).

10. **CVdir**:
    Coeficiente de variación asociado a la estimación directa (**DIR**), calculado como:

$$\text{CVdir} = \frac{\sqrt{\text{Vdir}}}{|\text{DIR}|}$$

11. **mufh**:
    Media estimada por el modelo para la variable de interés, relacionado con la estimación en ALR (no importante en nuestro caso).

12. **EBfh**:
    Estimación para la variable de interés

13. **MSEfh**:
    Error cuadrático medio estimado para la estimación indirecta (**EBfh**).

14. **CVfh**:
    Coeficiente de variación asociado a la estimación indirecta (**EBfh**), calculado como:

$$\text{CVfh} = \frac{\sqrt{\text{MSEfh}}}{|\text{EBfh}|}$$

15. **pdbench**:
    Estimación ajustada de un benchmark para la variable de interés.

16. **msebench**:
    Error cuadrático medio asociado al benchmark ajustado (**pdbench**).

17. **CVbench**:
    Coeficiente de variación asociado al benchmark ajustado (**pdbench**).

18. **Nombre**:
    Nombre de la provincia.

19. **Poblacion_EPA**:
    Población total estimada según la Encuesta de Población Activa.

20. **Poblacion_16a24**:
    Población total en el rango de edad de 16 a 24 años según la EPA.

Como podemos observar, los valores obtenidos de CV son bastante aceptables si usamos como referencia lo recomendado por la ONS (2006), que indica que los resultados tienen la calidad suficiente como para ser publicados si la estimación tiene un CV por debajo del 20%.

De todas formas, hagamos algún estudio de cómo son los residuos y de su estabilidad temporal y distribución geográfica.

Análisis de los datos obtenidos para todos los trimestres.

Para ello, lógicamente, habremos de aplicar el modelo anterior para todos los trimestres. Así pues, ejecutemos el siguiente bloque de código.

En este código está la función **UFH.NINIS**, que resume todo lo anterior en una sola función, que solo necesita las realizaciones bootstrap de entrada y los trimestres, ya que, previamente definida la BBDD, tiene en cuenta todo lo anterior (X, Ved. . .).

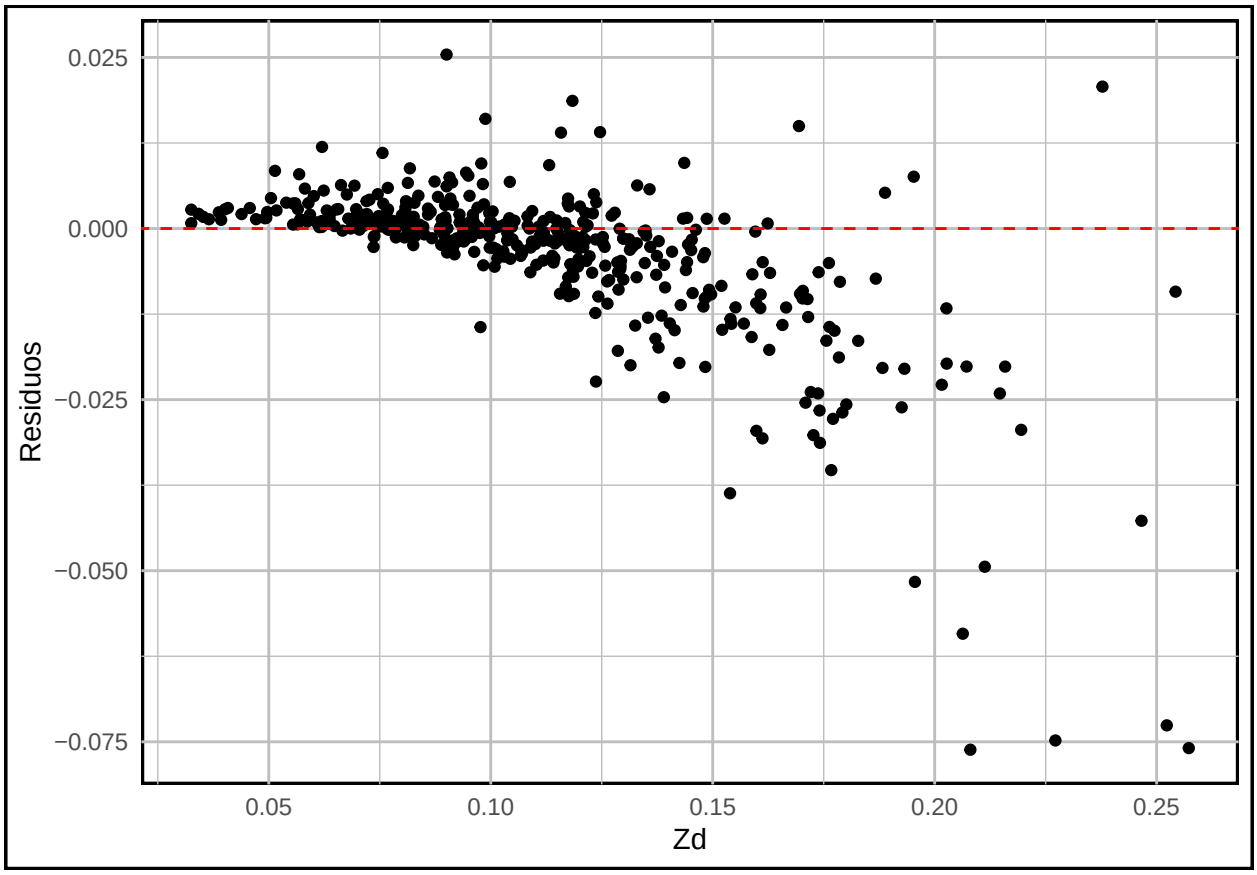
```r
### Bloque de "limpieza" de R y de librerías ###
# rm(list = ls())
# freshr::freshr()
# try(dev.off(dev.list()["RStudioGD"]), silent=TRUE)
# cat("\014")
```

```r
##############################
library(foreach); library(doParallel)
library(dplyr); library(openxlsx); library(MASS)
library(sae); library(Matrix); library(ggplot2);
library(gridExtra); library(mapSpain); library(sf)
library(xtable)
##############################

Trimestrales = list()
Trims = c(121,221,321,421,122,222,322,422)
source("UFH_NINIS.R")
load("BBDD_ninis.Rdata")

n_cores <- detectCores();
cluster <- makeCluster(n_cores - 1)
registerDoParallel(cluster)
```

```r
#### Ejecución paralela del modelo UFH para cada trimestre ####

Trimestrales <- foreach(t = seq_along(Trims)) %dopar% {
  library(sae); library(Matrix); library(dplyr);
  library(ggplot2); library(gridExtra); library(openxlsx)
  source("UFH_NINIS.R")
  load("BBDD_ninis.Rdata")
  T.aux = UFH.NINIS(Trims[t], B_boot = 100)
  Trimestrales[[t]] <- T.aux
}


T1 <- Trimestrales[[1]];
T2 <- Trimestrales[[2]];
T3 <- Trimestrales[[3]];
T4 <- Trimestrales[[4]];
T5 <- Trimestrales[[5]];
T6 <- Trimestrales[[6]];
T7 <- Trimestrales[[7]];
T8 <- Trimestrales[[8]];


y1 <- rbind(T1$NINIS, T2$NINIS, T3$NINIS, T4$NINIS,
            T5$NINIS, T6$NINIS, T7$NINIS, T8$NINIS)

corr <- cbind(T1$TC, T2$TC, T3$TC, T4$TC, T5$TC, T6$TC, T7$TC, T8$TC)
sign_ufh <- rbind(T1$TS.UFH, T2$TS.UFH, T3$TS.UFH, T4$TS.UFH,
                  T5$TS.UFH, T6$TS.UFH, T7$TS.UFH, T8$TS.UFH)

datos_ninis <- list(y1 = y1, corr = corr, sign = sign_ufh)
```

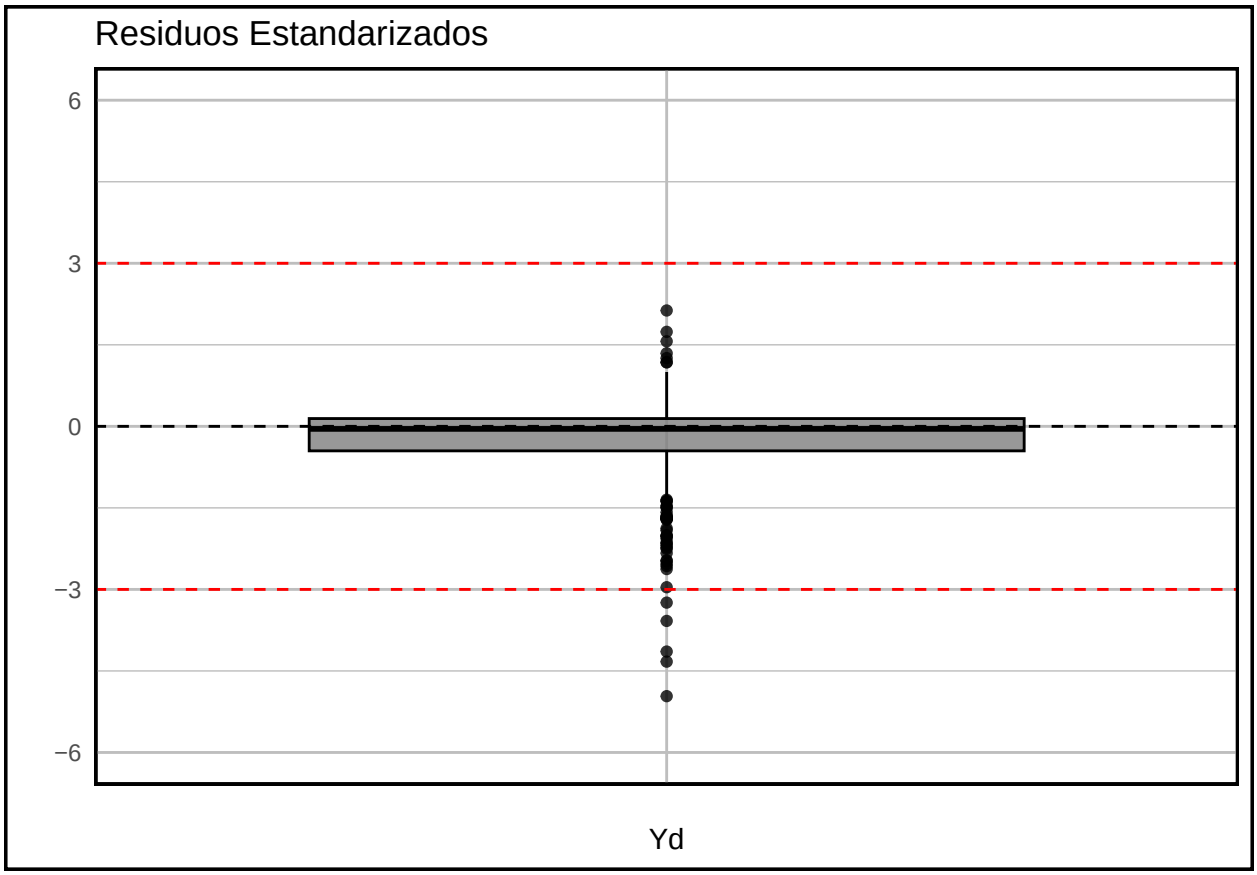
Miremos, por ejemplo, los residuos del modelo para el conjunto de datos. No nos fijemos en el código, sino que ejecutémoslo y miremos directamente lo que obtenemos en la salida.

```
res2 = cbind(y1$EBfh - y1$DIR, (1-y1$EBfh) - (1-y1$DIR))

resmarg1 = ggplot(data = NULL, aes(x = y1$DIR, y = y1$EBfh - y1$DIR)) +
  geom_point() +
    geom_abline(intercept = 0, slope = 0, color = "red", size = .5, lty=2) +
  labs(x = "Zd", y = "Residuos")+theme_minimal()+
 theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey")
  )
resmarg1
```




```
############################

resid_estandarizados <- res2[, 1] / sd(res2[, 1])

# Crear un data frame con los datos
dfq1 <- data.frame(Yd1 = resid_estandarizados)

# Crear el boxplot con ggplot
```

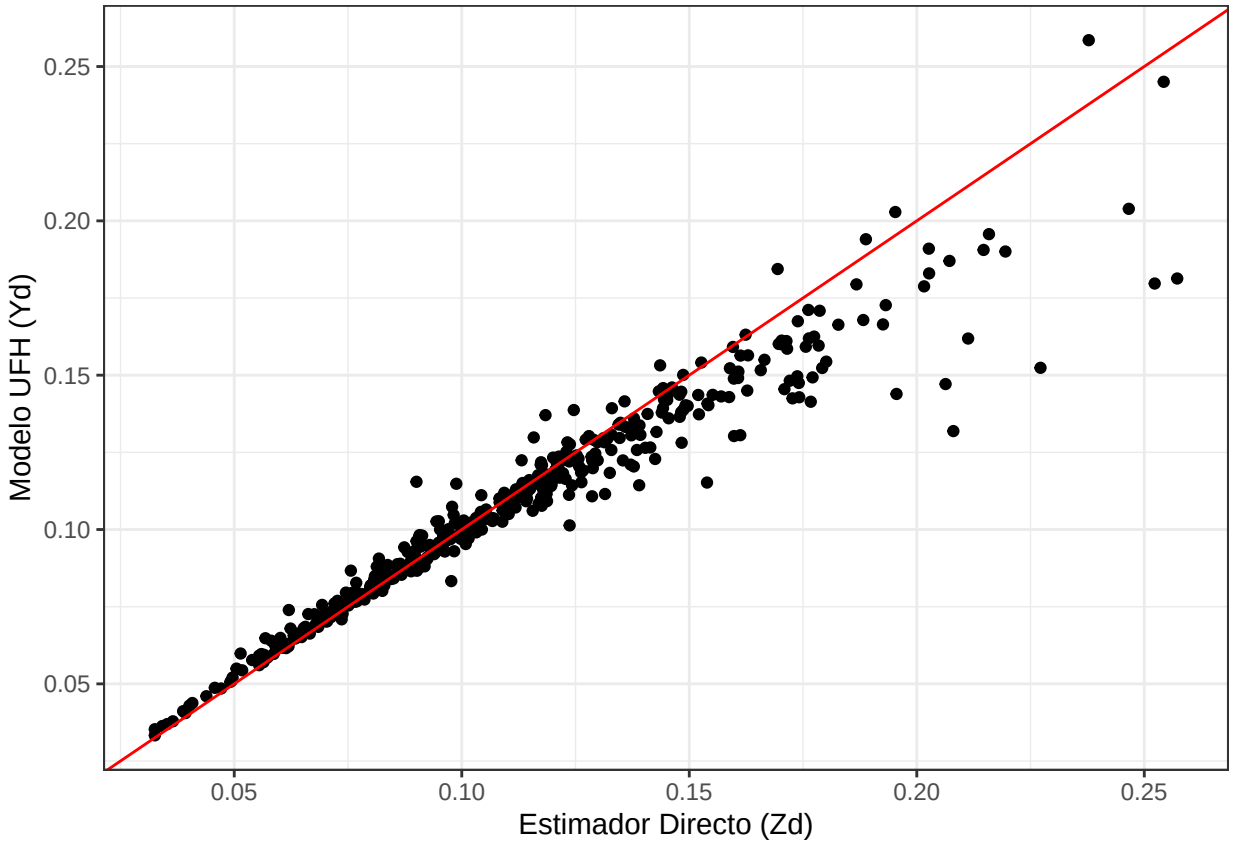
```r
box <- ggplot(dfq1, aes(x = "", y = Yd1)) +
  geom_boxplot(fill = "grey50", color = "black", alpha = 0.8) +
  labs(title = "Residuos Estandarizados", x = "Yd", y = "") +
  scale_y_continuous(limits = c(-6, 6)) +
  geom_hline(yintercept = 3, color = "red", lty = 2) +
  geom_hline(yintercept = 0, color = "black", linetype = "dashed") +
  geom_hline(yintercept = -3, color = "red", lty = 2) +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey"),
    panel.border = element_rect(color = "black", fill = NA, size = 1)  # Borde del panel
  )

box
```



Veamos los residuos de otra forma: observando la dispersión de los valores del modelo frente a los directos.

```r
data1 <- data.frame(x = y1$DIR, y = y1$EBfh)

dirvseblup1 = ggplot(data1, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", size = .5) +
```

```
  labs(x = "Estimador Directo (Zd)", y = "Modelo UFH (Yd)") +theme_bw()
dirvseblup1
```

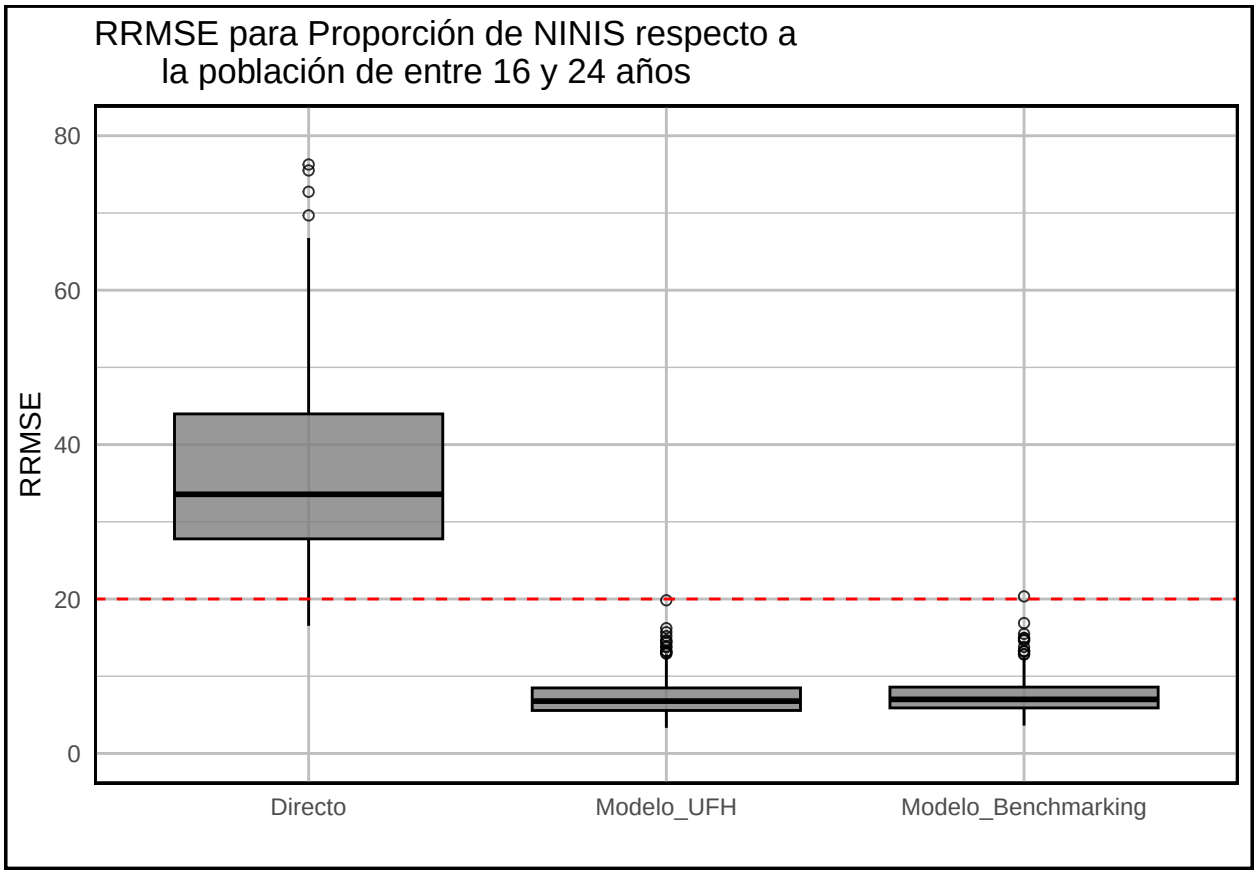


En este caso, vemos que los valores tienden a agruparse en torno a la bisectriz, salvo algunos puntos donde el directo sobreestima los ninis (o el modelo infraestima). Dado que los estimadores directos de proporciones son esencialmente insesgados, este patrón sugiere que los estimadores composicionales comparten parcialmente esta propiedad.

A continuación, presentamos los boxplots para la proporción estimada de ninis en comparación con la estimación directa. Como podemos observar, los valores de CV son mucho más reducidos que los proporcionados por la estimación directa. de forma que concluimos que el modelo tiene un mejor desempeño que dicho directo. Además, todos los valores salen por debajo del 20%, usando como criterio nuevamenente la ONS (2007), lo cual da un aliciente a emplear los modelos FH en comparación con la estimación directa.

```
res2.RMSE <- data.frame(
  Directo = y1$CVdir,
  Modelo_UFH = y1$CVfh,
  Modelo_Benchmarking = y1$CVfh*y1$L
)

res2.RMSE_long <- tidyr::pivot_longer(res2.RMSE,
                                      cols = everything(),
                                      names_to = "Método",
                                      values_to = "RRMSE")
res2.RMSE_long$Método <- factor(res2.RMSE_long$Método,
                                levels = c("Directo", "Modelo_UFH",
                                           "Modelo_Benchmarking"))
```

```
rrmse <- ggplot(res2.RMSE_long, aes(x = Método, y = RRMSE)) +
  geom_boxplot(fill = "grey50", color = "black",
               alpha = 0.8, outlier.shape = TRUE) +
  geom_hline(yintercept = 20, color = "red",
             linetype = "dashed") +
  labs(title = "RRMSE para Proporción de NINIS respecto a
       la población de entre 16 y 24 años",
       x = "", y = "RRMSE") +
  scale_y_continuous(limits = c(0, 80)) +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey"),
    panel.border = element_rect(color = "black", fill = NA, size = 1)
  )

rrmse
```



Finalmente, resumimos toda la información trimestral en la siguiente tabla resumen, junto con los totales. Mostramos en la última línea el último trimestre de 2021 como ejemplo.

```
tabla1 <- data.frame(Provincia = y1$Nombre, ccaa=y1$ccaa,
                     Prov=y1$Prov, Trim=y1$trim,
```

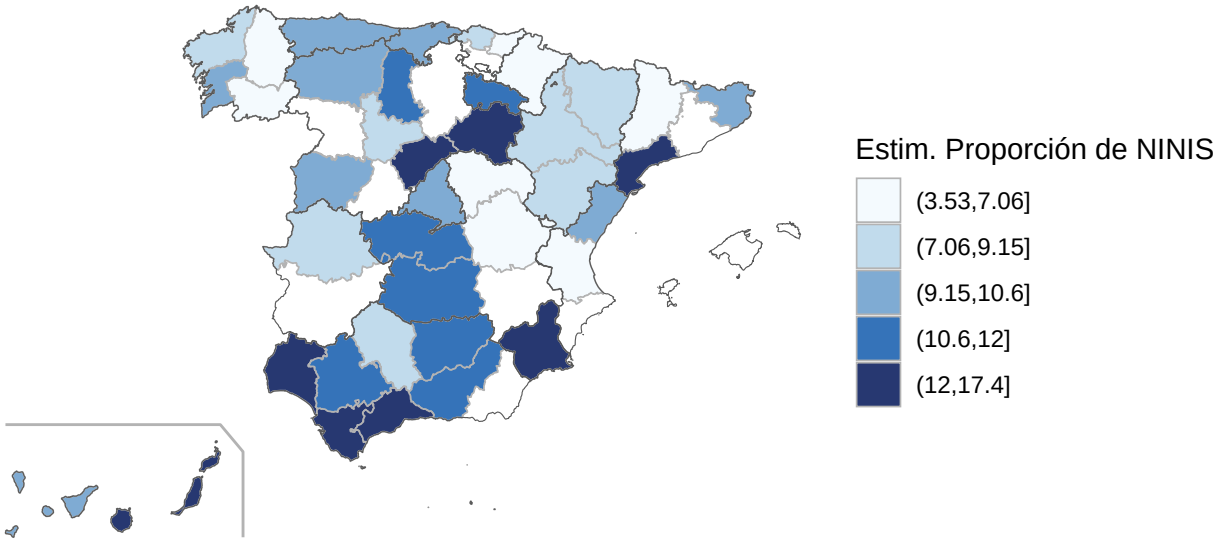
```
                      nd=y1$nd, nini_nd= y1$nini_nd, nd_16a24= y1$nd_16a24,
                      Nd = y1$Poblacion_EPA, Nd16a24 = y1$Poblacion_16a24,
                      ninidir=100*y1$DIR, cvninidir = y1$CVdir,
                      niniind=100*y1$EBfh, cvniniind=y1$CVfh,
                      ninibench = 100*y1$pdbench, cvninibench = y1$CVfh*y1$L
                      )

tabla_final <- tabla1 %>% mutate(Total.Nini.Dir = round(ninidir*Nd16a24/100),
                                 Total.Nini.UFH = round(niniind*Nd16a24/100),
                                 Total.Nini.Bench = round(ninibench*Nd16a24/100))

tabla_final_421 <- tabla_final %>% filter(Trim==421)

# print((tabla_final_421))
```

Finalizamos este estudio con el mapeo de los ninis por provincias. Para ello, emplearemos el siguiente código.

```
# Para los cortes a nivel de cuantil.

aa  =seq(0.025, 0.975, (0.975-0.025)/5)
bb = seq(0, 1, (1-0)/5)

#Cargo los códigos de los municipios

munic <- esp_get_prov()

#Represento ninis

ninis.mapa=y1
ninis.mapa=ninis.mapa%>%filter(trim==122)
ninis.mapa$Prov <- as.character(ninis.mapa$Prov)
representar=left_join(munic,ninis.mapa,by=c("cpro"="Prov"))%>%
  filter(!is.na(nd))%>%mutate(ninis=pdbench*100)

#Cargo el mapa de las CCAA para establecer las siluetas de las CCAA
ccaa_sf <- esp_get_ccaa()
can <- esp_get_can_box()

#Calculo el intevalo de la leyenda

representar$cuts <- cut(representar$ninis, quantile(representar$ninis,probs = bb))

estimador_bench=ggplot(representar)+
  geom_sf(aes(fill = cuts), color = "grey70", linewidth = .3) +
  geom_sf(data = ccaa_sf, fill = NA)+
  geom_sf(data = can, color = "grey70") +
  scale_fill_manual(
    values = rev(hcl.colors((length(table(representar$cuts))),
                            "blues")),na.translate = FALSE,
    guide = guide_legend(title = "Estim. Proporción de NINIS",
                         direction = "vertical"))+   theme_void() +
  labs(title = "Mapa de la Proporción de NINIS")
representar=left_join(munic,ninis.mapa,by=c("cpro"="Prov"))%>%
```

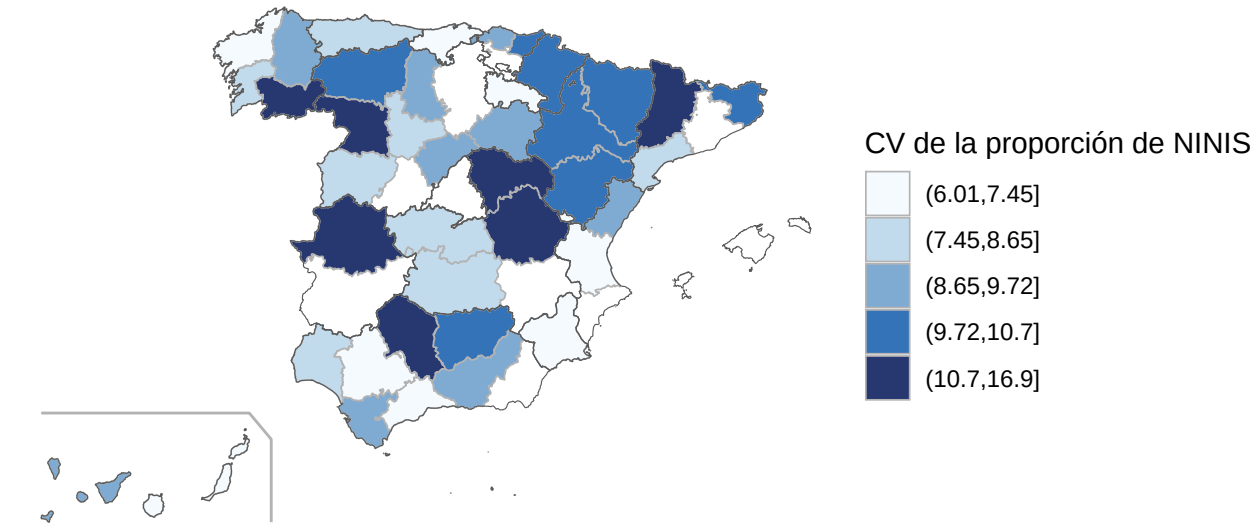
```
  filter(!is.na(nd))%>%mutate(ninis=CVfh*L)

#Cargo el mapa de las CCAA para establecer las siluetas de las CCAA
ccaa_sf <- esp_get_ccaa()
can <- esp_get_can_box()

#Calculo el intevalo de la leyenda

representar$cuts <- cut(representar$ninis, quantile(representar$ninis,probs = bb))

cv_est_bench=ggplot(representar)+
  geom_sf(aes(fill = cuts), color = "grey70", linewidth = .3) +
  geom_sf(data = ccaa_sf, fill = NA)+
  geom_sf(data = can, color = "grey70") +
  scale_fill_manual(
    values = rev(hcl.colors((length(table(representar$cuts))), "blues")),
    na.translate = FALSE,
    guide = guide_legend(title = "CV de la proporción de NINIS",
                         direction = "vertical")) +
  theme_void() +
  labs(title = "Mapa del Coeficiente de Variación (CV) de la Proporción de NINIS")

estimador_bench
```

Mapa de la Proporción de NINIS



En el mapa con la leyenda *Estim. Proporción de NINIS*, las proporciones de ninis con respecto de la población de 16 a 24 años objetivo oscilan entre 4.85% y 20.8%.

```
cv_est_bench
```



El mapa de la posición inferior, con la leyenda *CV de la proporción de NINIS*, muestra el coeficiente de variación de las estimaciones de la proporción de ninis al respecto de la población de 16 a 24 años, cuyos valores varían desde 3.60% hasta 7.21%.

Es notable que las provincias con mayor proporción de ninis no siempre coinciden con las de mayor variabilidad, sugiriendo que algunas regiones, aunque tienen altas proporciones de ninis, también presentan estimaciones más estables y viceversa.

Algo a comentar es que las zonas de mayor concentración de ninis estén en la zona sur y levantina de España, pero también en la llamada España Vacía, siendo Soria, Ciudad Real o Cuenca grandes ejemplos de este caso con una estimación de ninis respecto a la población de 16 a 24 años considerablemente elevada.

```
stopCluster(cl = cluster)
```

Conclusiones

La implementación de los métodos UFH es relativamente sencilla. En efecto, y gracias a la librería *sae*, podemos obtener las estimaciones de forma rápida. A su vez, el análisis y obtención del error, al tratarse de un modelo univariante, es rápida también incluso empleando iteraciones bootstrap elevadas.

Lo que observamos es que el grueso del trabajo es más la organización de los datos y la transformación alr. En efecto, de haber obviado esta transformación, el código se habría reducido considerablemente.

Sea como fuere, y vistos los resultados comparativos entre el modelo UFH y las estimaciones directas, el uso de dicho modelo es especialmente recomendado para con el fin de obtener resultados con bajo CV y estimaciones análogas (gracias a la conformidad) a las directas.

Ahora, con todo esto explicado, pasaríamos al modelo bivariante, donde estimaremos simultáneamente tanto la ocupación como el desempleo y la correlación entre ambas variables laborales.

Aplicación del modelo BFH para la estimación de ocupados, parados, inactivos y tasa de desempleo por sexo, nacionalidad y área geográfica

Alexandro Aneiros Batista

Enero de 2025

En este documento se presentan las aplicaciones a datos reales de las estimaciones de ocupados, parados e inactivos, así como la tasa de paro, en diversos dominios de interés, utilizando datos del último trimestre de la Encuesta de Población Activa del 2021. El análisis se centra en las provincias españolas.

Usaremos un procedimiento para calcular las estimaciones de Hájek, junto con sus varianzas y coeficientes de variación (CV).

A continuación, se aborda el ajuste de un modelo Fay-Herriot a nivel de área para el caso bivariante, la obtención de predictores plug-in y la estimación del error cuadrático medio (MSE) mediante bootstrap paramétrico.

Breve descripción de esta práctica.

Nos enfocaremos en la estimación de ocupados, parados, inactivos y tasa de desempleo por sexo, nacionalidad y provincia españolas. Los datos sobre los que trabajaremos serán una mezcla de los casos del Objetivo 1 y del Objetivo 3, ya que ambos usan el modelo BFH para datos composicionales, y nos interesa ver ambos objetivos.

El objetivo de este caso de uso es obtener estimaciones de los totales anteriores haciendo subconjuntos a tres niveles: nivel geográfico, tomando las provincias; a nivel de sexo; y a nivel de extranjería, mirando si el individuo es o no extranjero; y todo ello a partir de los microdatos de la EPA y datos auxiliares de interés que mencionaremos a continuación. También tendremos en cuenta la variable temporal trimestral, donde tomaremos el año 2021.

La estructura que seguiremos es análoga, salvo particularidades y versiones resumidas, del caso UFH.

Aplicación a datos reales:

Carga de librerías y base de datos, y variables de interés.

```
# Cargar librerías necesarias

# Lista de todas las librerías necesarias
paquetes <- c(
  "future", "future.apply", "dplyr", "openxlsx",
  "sae", "Matrix", "ggplot2", "gridExtra" #, "mapSpain", "sf", "xtable"
```

```
)

# Instalación y carga de paquetes
for (paquete in paquetes) {
  if (!require(paquete, character.only = TRUE)) {
    install.packages(paquete, dependencies = TRUE)
    library(paquete, character.only = TRUE)
  }
}

# Mensaje de confirmación
cat("Todos los paquetes se han cargado correctamente.\n")
```

Para obtener una mayor velocidad de cálculo, se opta por emplear cálculo paralelo con el uso de las librerías *future* y *future.apply*, análogas a las ya usadas *doParalell* y *foreach*.

Ahora, carguemos la base de datos (BBDD), que contiene tanto los estimadores directos, como sus varianzas. También estará ahí la información auxiliar de interés.

```
load('DEF_BBDD_prov.RData')
# Varianzas y proporciones muestrales, entre otros, datos auxiliares, EPA...
glimpse(df)
```

```
## Rows: 1,664
## Columns: 39
## $ CPRO2             <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01"~
## $ TRIM              <dbl> 121, 121, 121, 121, 122, 122, 122, 122, 221, 221, 22~
## $ EXT               <dbl> 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0~
## $ SEXO              <dbl> 1, 6, 1, 6, 1, 6, 1, 6, 1, 6, 1, 6, 1, 6, 1, 6, 1, 6~
## $ afiliadosfe       <dbl> 61226.52, 54888.75, 9150.45, 9694.08, 64044.95, 6064~
## $ pensionfe         <dbl> 45092.09, 48548.10, 3074.02, 2444.13, 42726.05, 4160~
## $ epa16a25fe        <dbl> 13036.13, 9422.52, 2282.74, 3240.00, 15381.93, 12175~
## $ epa26a45fe        <dbl> 29650.68, 29062.22, 9232.12, 11471.41, 28678.54, 294~
## $ epa46a64fe        <dbl> 42284.63, 42406.43, 4630.43, 4283.30, 45465.21, 4442~
## $ epa65ymas         <dbl> 31131.71, 39247.64, 689.79, 1167.80, 26842.05, 35824~
## $ paradossepe       <dbl> 2515.667, 2515.667, 2515.667, 2515.667, 2179.636, 21~
## $ contratossepe     <dbl> 3684.290, 3684.290, 3684.290, 3684.290, 4314.727, 43~
## $ pop_padron        <dbl> 36847.19, 36847.19, 36847.19, 36847.19, 35550.88, 35~
## $ pension_media     <dbl> 1258.411, 1258.411, 1258.411, 1258.411, 1261.618, 12~
## $ rnmp              <dbl> 14601.06, 14601.06, 14601.06, 14601.06, 14212.82, 14~
## $ rnmh              <dbl> 35137.68, 35137.68, 35137.68, 35137.68, 34063.12, 34~
## $ s11               <dbl> 0.0004860543, 0.0004824160, 0.0050903238, 0.00395336~
## $ s12               <dbl> -5.547322e-05, -3.858011e-05, -2.981478e-03, -1.0632~
## $ s13               <dbl> -0.0004305811, -0.0004438359, -0.0021088456, -0.0028~
## $ s22               <dbl> 9.830571e-05, 7.993796e-05, 3.852084e-03, 2.085873e-~
## $ s23               <dbl> -4.283249e-05, -4.135785e-05, -8.706057e-04, -1.0225~
## $ s33               <dbl> 0.0004734136, 0.0004851937, 0.0029794512, 0.00391267~
## $ pocupados         <dbl> 0.5264277, 0.4665026, 0.5930052, 0.4320548, 0.551597~
## $ pparados          <dbl> 0.04449879, 0.04272383, 0.22051336, 0.16407729, 0.03~
## $ pinactivos        <dbl> 0.4290735, 0.4907735, 0.1864814, 0.4038679, 0.412587~
## $ pocupados_nocorr  <dbl> 0.5264277, 0.4665026, 0.5930052, 0.4320548, 0.551597~
## $ pparados_nocorr   <dbl> 0.04449879, 0.04272383, 0.22051336, 0.16407729, 0.03~
## $ pinactivos_nocorr <dbl> 0.4290735, 0.4907735, 0.1864814, 0.4038679, 0.412587~
```

```
## $ n                 <int> 544, 542, 52, 69, 484, 515, 51, 67, 507, 522, 44, 67~
## $ pop               <dbl> 116103.15, 120138.81, 16835.08, 20162.51, 116367.73,~
## $ ocupados          <dbl> 61119.91, 56045.07, 9983.29, 8711.31, 64188.10, 6161~
## $ parados           <dbl> 5166.45, 5132.79, 3712.36, 3308.21, 4167.73, 4537.28~
## $ inactivos         <dbl> 49816.79, 58960.95, 3139.43, 8142.99, 48011.90, 5569~
## $ cv1               <dbl> 4.187971, 4.708220, 12.031346, 14.552739, 4.342379, ~
## $ cv2               <dbl> 22.28133, 20.92697, 28.14576, 27.83527, 26.04595, 22~
## $ cv3               <dbl> 5.070942, 4.488244, 29.270660, 15.488071, 5.734442, ~
## $ cv1_nocorr_sivar  <dbl> 4.187971, 4.708220, 12.031346, 14.552739, 4.342379, ~
## $ cv2_nocorr_sivar  <dbl> 22.28133, 20.92697, 28.14576, 27.83527, 26.04595, 22~
## $ cv3_nocorr_sivar  <dbl> 5.070942, 4.488244, 29.270660, 15.488071, 5.734442, ~
```

Como podemos observar, muchas variables son las mismas que en el UFH. Esto es porque se usa la misma base de datos. De todas formas, ahora viene agregada la edad desde los datos de la EPA por grupos etarios, que trataremos posteriormente.

Desde la BBDD, vamos a introducir las variables alr al igual que hicimos en el caso UFH, y alguna proporción de interés.

```
### Tomamos los datos de la BBDD anterior y añadimos algunas ###

df <- df %>%
  mutate(dom = paste(CPRO2, SEXO, EXT, TRIM, sep = "")) %>%
  dplyr::select(dom, everything())

df <- df %>% mutate(dom = as.double(dom)) %>% rename(PK_TRIM = TRIM)
df <- df %>%
  mutate(pparados = pparados/(pparados+pocupados+pinactivos),
         pocupados = pocupados/(pparados+pocupados+pinactivos),
         pinactivos = pinactivos/(pparados+pocupados+pinactivos),
         porcentaje_afiliadosss_epa = afiliadosfe / pop,
         porcentaje_16a25_epa = epa16a25fe / pop,
         porcentaje_26a45_epa = epa26a45fe / pop,
         porcentaje_46a65_epa = epa46a64fe / pop,
         porcentaje_65ymas_epa = epa65ymas / pop,
         yd1 = log(pparados/pinactivos),
         yd2 = log(pocupados/pinactivos),
)
```

Puede ser que en algún caso haya problemas de ceros al tratarse de más de una categoría práctica (como en el caso de ninis, donde la otra variable simplemente era el complementario). Aquí, al tener tres categorías prácticas (ocupados, parados e inactivos), es posible que todas las observaciones caigan en una categoría. Ante esto, podemos hacer un redondeo como el que aparece a continuación para solventar el problema. También construimos la matriz **H** necesaria para el cambio de base del espacio con nuestras variables naturales al caso con la transformación alr.

Es interesante estudiar la transformación alr antes de empezar el método, ya que a veces los problemas de convergencia pueden venir por problemas en la transformación de la matriz de varianzas-covarianzas.

Si hubiese algún problema con elementos de esta matriz transformada muy próximos a cero, les podemos añadir, por ejemplo, un pequeño término épsilon que puede ser la varianza del primer decil, o, por suavidad, la media o mediana.

```
###### SOLUCIÓN DEL PROBLEMA DE CEROS #####

D=nrow(df)
datos2 <- df %>% dplyr::select(yd1, yd2)

for (i in 1:D){
  if (min(abs(datos2[i,1:2]))<=0.001){
    k=which(abs(datos2[i,1:2])<=0.001)
    u=(1/10)*min(datos2[i,-k], na.rm = T)
    suma2=2-length(k)

    datos2[i,1]=datos2[i,1]-u/suma2
    datos2[i,2]=datos2[i,2]-u/suma2
    datos2[i,k]=u
  }
}

df[,"yd1"]=datos2[,1]; df[,"yd2"]=datos2[,2]

df <- df %>%
  mutate( # por si aparece algún infinito
    yd1 = ifelse(yd1 == Inf | yd1 == -Inf, min(df$yd1), yd1),
    yd2 = ifelse(yd2 == Inf | yd2 == -Inf, min(df$yd2), yd2)
  )

df.all <- df #copia de la BBDD

# ############## Construcción apriorística de la matriz H
# #############       y de las Var. ALR ##################

q = 3;
H0.el = q*(diag(q-1) + matrix(rep(1,q-1), nrow=q-1)%*%matrix(rep(1,q-1), ncol=q-1))

Ved <- list();
D = dim(df)[1]
H0 <- list();
Ved_alr <- list();
s11_alr = c(); s12_alr = c(); s22_alr = c()

  for(d in 1:D){
    H0[[d]] <- H0.el
    Ved[[d]] <- matrix(NA, nrow=2, ncol=2)
    sup <- as.numeric(df$s12[d])
    Ved[[d]][upper.tri(Ved[[d]])] <- sup
    Ved[[d]][lower.tri(Ved[[d]])] <- sup
    diag(Ved[[d]]) <- as.numeric(c(df$s11[d], df$s22[d]))
    Ved_alr[[d]] = H0[[d]]%*%Ved[[d]]%*%t(H0[[d]])
    s11_alr[d] <- Ved_alr[[d]][1,1]
    s12_alr[d] <- Ved_alr[[d]][1,2]
    s22_alr[d] <- Ved_alr[[d]][2,2]
  }

Ved <- Ved_alr
```

```r
df$s11_alr <- s11_alr; df$s12_alr <- s12_alr; df$s22_alr <- s22_alr;

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

rm(list = c("Ved", "H0", "Ved_alr", "H0.el",
            "sup", "q", "d", "D", "paquete", "paquetes",
            "suma2", "u", "i", "k", "eps", "datos2"))
```

Notar que hacemos una copia del data.frame y un cambio de nombres. Esto no afectará a los resultados, es simplemente por coherencia interna del algoritmo usado en el entregable. Ejecutaremos el algoritmo presente en el archivo *BFH_PAR_OCU.R*. La estructura de éste es muy similar al visto para el caso UFH. A continuación, comentaremos las covariables que emplearemos en este caso:

1. **16-25**: Proporción de individuos de entre 16 y 25 años (marco de personas del INE).

2. **26-45**: Proporción de individuos de entre 26 y 45 años (marco de personas del INE).

3. **65+**: Proporción de individuos mayores de 65 años (marco de personas del INE).

4. **RNMP**: Renta neta media por persona (datos elaborados por el INE).
5. **Afiliados (AfilSS)**: Porcentaje de individuos afiliados a la Seguridad Social (obtenidos de los microdatos proporcionados por el INE).

Evidentemente, dentro de la BBDD se usaron más covariables para el estudio, pero para este caso específico éstas eran las más significativas a lo largo de los trimestres, y de ahí que las usemos.

```r
source("BFH_PAR_OCU.R")

tini<-proc.time()

#################################### EJECUTAMOS BFH

BFH <- function(t) {Trims_vector <- c(121, 221, 321, 421);
library(sae)
library(Matrix)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(openxlsx)
return(BFH.PAR.OCU(Trims_vector[t], F, 200))}

library(future)
library(future.apply)

plan(multisession)  # Usa todos los núcleos disponibles

Trimestrales <- future_lapply(1:4, BFH)

elapsed_time <- proc.time() - tini
print(elapsed_time)

#######################################
# Juntamos datos
```

```
T1 <- Trimestrales[[1]];
T2 <- Trimestrales[[2]];
T3 <- Trimestrales[[3]];
T4 <- Trimestrales[[4]];

y1 <- rbind(T1$PAR, T2$PAR, T3$PAR, T4$PAR)
y2 <- rbind(T1$OCU, T2$OCU, T3$OCU, T4$OCU)

df.all <- df.all %>%
  dplyr::select(dom, SEXO, everything()) %>%
  filter(PK_TRIM %in% c(121, 221, 321, 421))
#filtramos porque solo queremos el 2021.

### Hacemos algunos arreglos de nombres y juntamos todo en un solo data.frame.

y1 <- as_tibble(y1)
y1 <- y1 %>% arrange(CPRO2)
df.all <- df.all %>% arrange(PK_TRIM, CPRO2)
y1 <- y1 %>% rename(dom = dominio)
y2 <- as_tibble(y2)
y2 <- y2 %>% rename(dom = dominio)

df_plot <- left_join(df.all, y1, by="dom")
df_plot2 <- left_join(df_plot, y2, by="dom")
df_plot <- df_plot2
```

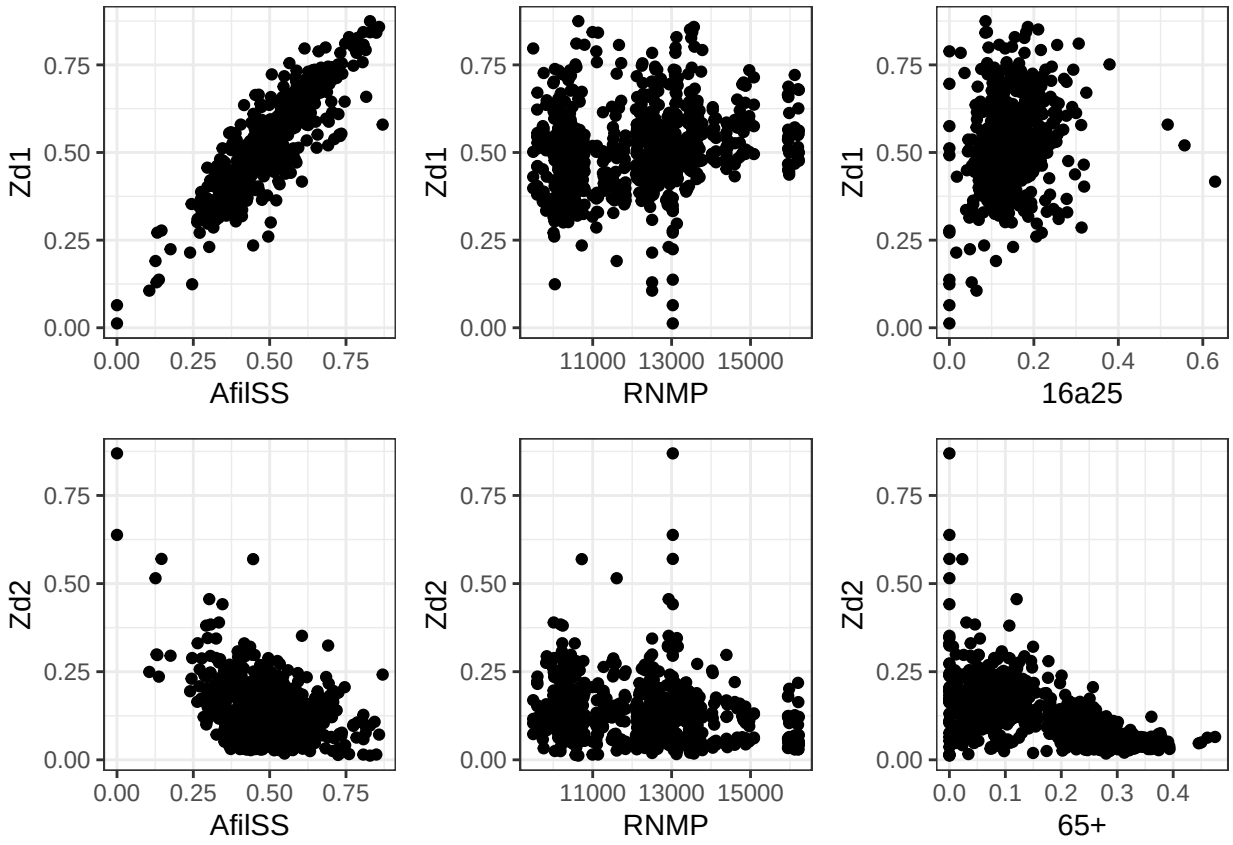
Gráficos de análisis de los resultados.

Una vez que tengamos todos los datos reunidos, procederemos a estudiar la salida a través de diversos gráficos, como hicimos en el caso UFH.

Por ejemplo, observemos la relación de las variables directas frente a las explicativas, al igual que los residuos y sus diferentes representaciones para estudiar el modelo ajustado.

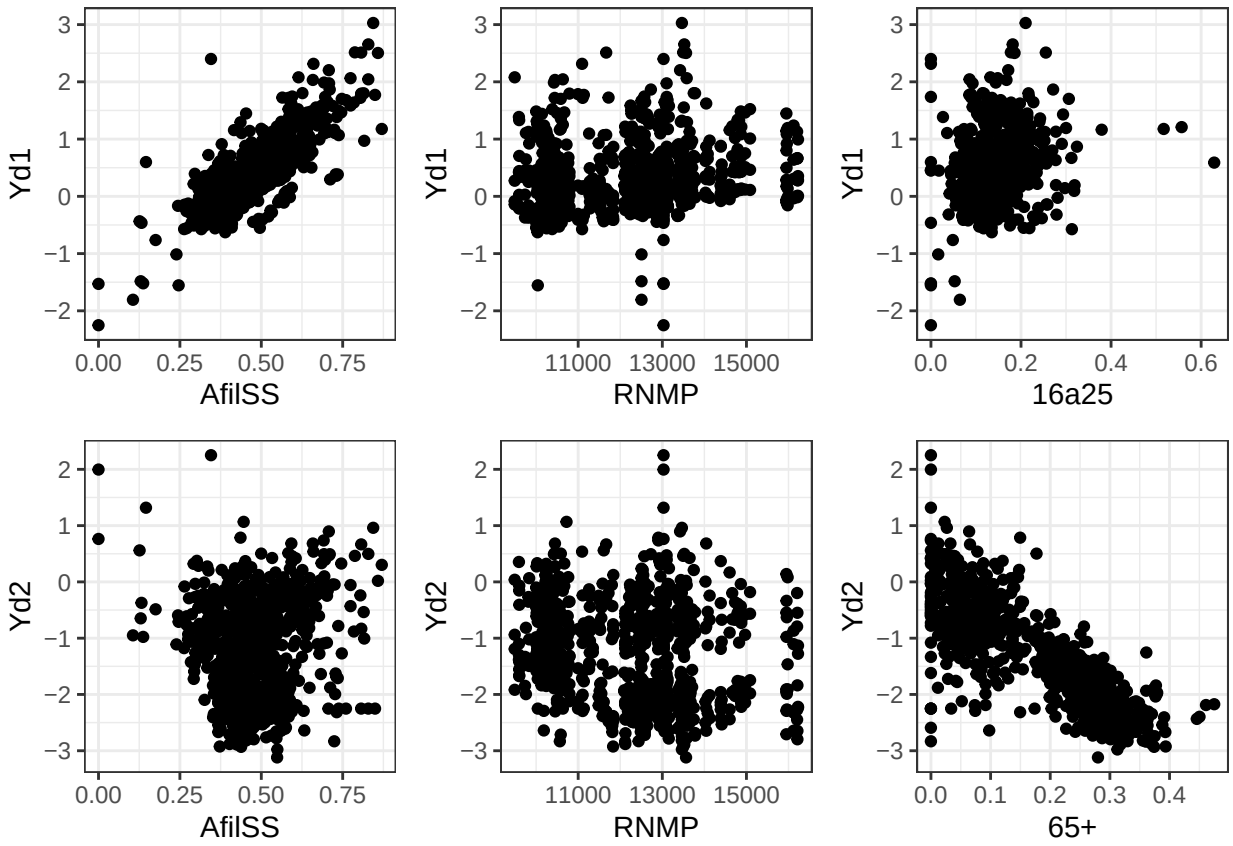
```
######## GRÁFICOS COVARIABLES #######

afilsszd1 = ggplot(data = NULL, aes(x = df_plot$porcentaje_afiliadosss_epa,
                                    y = df_plot$pparados)) +
  geom_point() +
  labs(x = "AfilSS", y = "Zd2")+theme_bw()

afilsszd2 = ggplot(data = NULL, aes(x = df_plot$porcentaje_afiliadosss_epa,
                                    y = df_plot$pocupados)) +
  geom_point() +
  labs(x = "AfilSS", y = "Zd1")+theme_bw()

rnmpzd1 = ggplot(data = NULL, aes(x = df_plot$rnmp,
                                  y = df_plot$pparados)) +
  geom_point() +
  labs(x = "RNMP", y = "Zd2")+theme_bw()

rnmpzd2 = ggplot(data = NULL, aes(x = df_plot$rnmp,
```

```
                                    y = df_plot$pocupados)) +
  geom_point() +
  labs(x = "RNMP", y = "Zd1")+theme_bw()

edad65pzd1 = ggplot(data = NULL, aes(x = df_plot$porcentaje_65ymas_epa,
                                     y = df_plot$pparados)) +
  geom_point() +
  labs(x = "65+", y = "Zd2")+theme_bw()

edad16_30zd2 = ggplot(data = NULL, aes(x = df_plot$porcentaje_16a25_epa,
                                       y = df_plot$pocupados)) +
  geom_point() +
  labs(x = "16a25", y = "Zd1")+theme_bw()


grid.arrange(afilsszd2, rnmpzd2,edad16_30zd2,
             afilsszd1, rnmpzd1, edad65pzd1, nrow = 2)
```




```
######## GRÁFICOS COVARIABLES TRANSFORMADAS #######

afilsszd1 = ggplot(data = NULL, aes(x = df_plot$porcentaje_afiliadosss_epa,
                                    y = df_plot$yd1)) +
  geom_point() +
  labs(x = "AfilSS", y = "Yd2")+theme_bw()

afilsszd2 = ggplot(data = NULL, aes(x = df_plot$porcentaje_afiliadosss_epa,
```

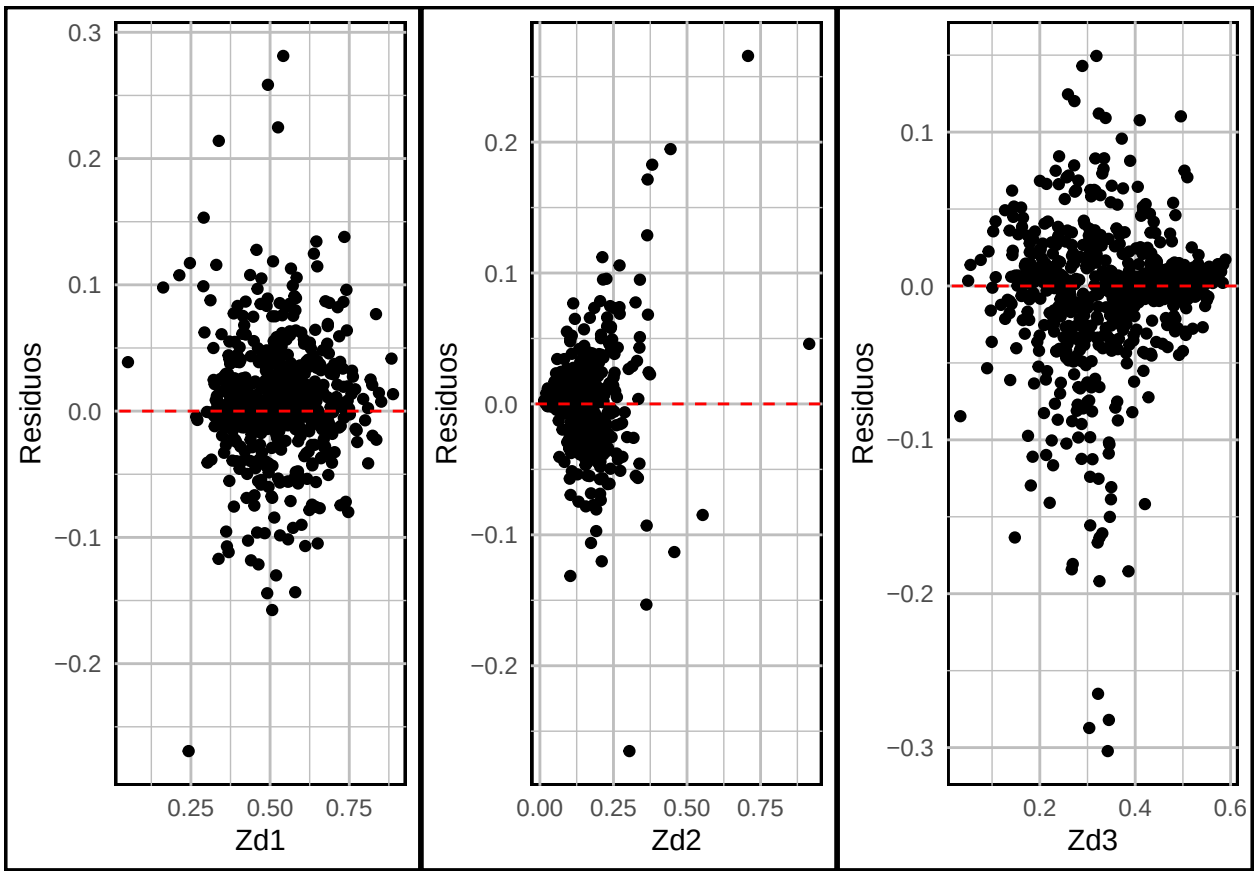
```
                                    y = df_plot$yd2)) +
  geom_point() +
  labs(x = "AfilSS", y = "Yd1")+theme_bw()

rnmpzd1 = ggplot(data = NULL, aes(x = df_plot$rnmp,
                                  y = df_plot$yd1)) +
  geom_point() +
  labs(x = "RNMP", y = "Yd2")+theme_bw()

rnmpzd2 = ggplot(data = NULL, aes(x = df_plot$rnmp,
                                  y = df_plot$yd2)) +
  geom_point() +
  labs(x = "RNMP", y = "Yd1")+theme_bw()

edad65pzd1 = ggplot(data = NULL, aes(x = df_plot$porcentaje_65ymas_epa,
                                     y = df_plot$yd1)) +
  geom_point() +
  labs(x = "65+", y = "Yd2")+theme_bw()

edad16_30zd2 = ggplot(data = NULL, aes(x = df_plot$porcentaje_16a25_epa,
                                       y = df_plot$yd2)) +
  geom_point() +
  labs(x = "16a25", y = "Yd1")+theme_bw()


grid.arrange(afilsszd2, rnmpzd2, edad16_30zd2,afilsszd1, rnmpzd1, edad65pzd1,
             nrow = 2)
```



De izquierda a derecha, tenemos las gráficas de cómo se comporta el estimador directo frente a algunas covariables, tanto para ocupados (fila superior) como para parados (fila inferior). También lo hacemos para la variable transformada.

```
######## PLOTS de residuos (homocedasticidad) ##########

df_plot$pd3 <- 1 - df_plot$pd1 - df_plot$pd2

res2 = cbind(df_plot$pd1 - df_plot$pparados,
             df_plot$pd2 - df_plot$pocupados,
             df_plot$pd3 - df_plot$pinactivos)

resmarg1 = ggplot(data = NULL, aes(x = df_plot$pd1,
                                   y = df_plot$pd1 - df_plot$pparados)) +
  geom_point() +
    geom_abline(intercept = 0, slope = 0, color = "red", size = .5, lty=2) +
  labs(x = "Zd2", y = "Residuos")+theme_minimal()+
 theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey")
  )


resmarg2 = ggplot(data = NULL, aes(x = df_plot$pd2,
                                   y = df_plot$pd2 - df_plot$pocupados)) +
```

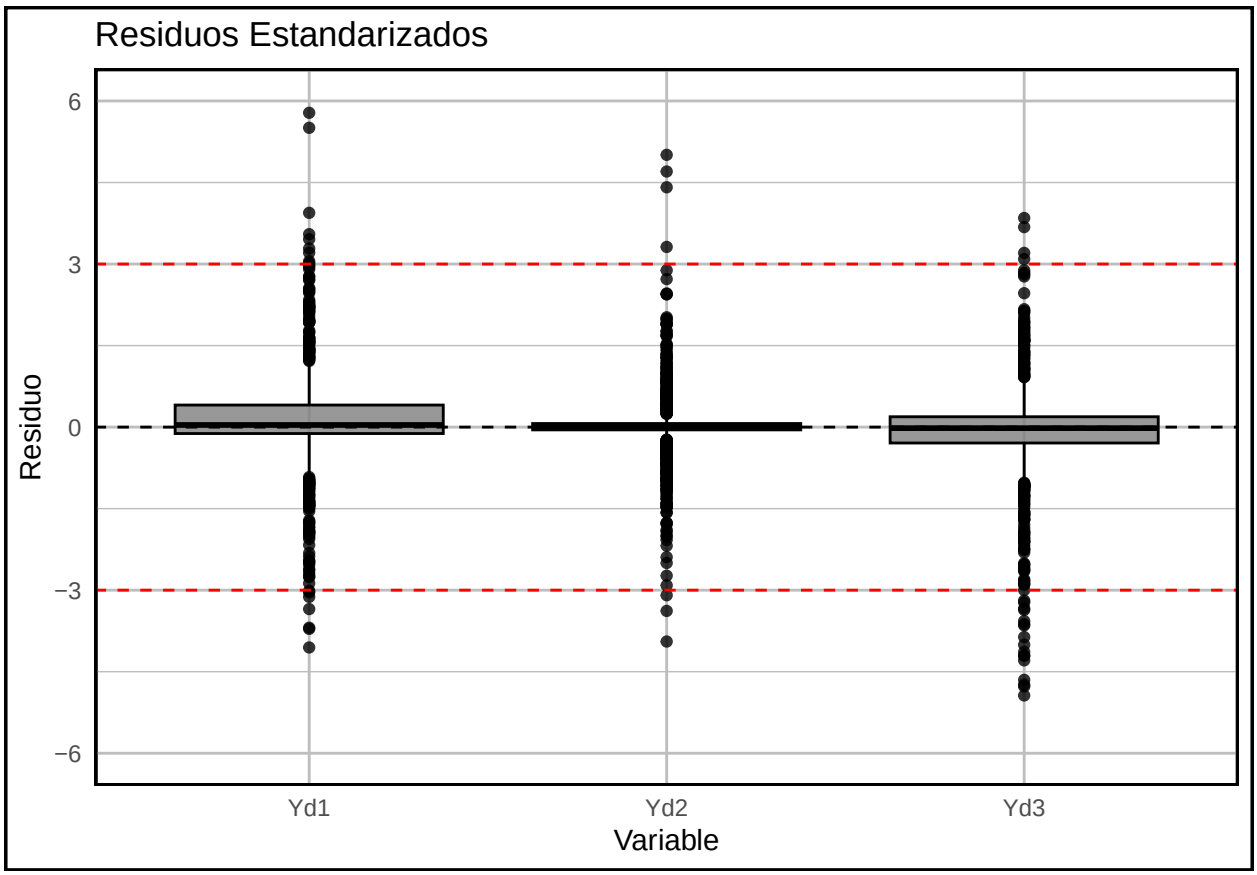
```
  geom_point() +
    geom_abline(intercept = 0, slope = 0, color = "red", size = .5, lty=2) +
  labs(x = "Zd1", y = "Residuos")+theme_minimal()+
 theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey")
  )

resmarg3 = ggplot(data = NULL, aes(x = df_plot$pd3,
                                   y = df_plot$pd3 - df_plot$pinactivos)) +
  geom_point() +
    geom_abline(intercept = 0, slope = 0, color = "red", size = .5, lty=2) +
  labs(x = "Zd3", y = "Residuos")+theme_minimal()+
 theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey")
  )

grid.arrange(resmarg2, resmarg1, resmarg3, nrow=1)
```



En este gráfico analizamos la dispersión de los datos en comparación con las estimaciones directas. Como se puede observar, los puntos tienden a agruparse formando una nube homogénea alrededor de la línea

de referencia, lo que indica una distribución consistente de los datos y una buena correspondencia entre las estimaciones directas y los valores modelados. Esta homogeneidad sugiere que el modelo logra captar adecuadamente las tendencias generales, minimizando la presencia de valores extremos o patrones anómalos.

```
#################### BOXPLOTS ##########################

resid_estandarizados <- res2 / sd(res2, na.rm = T)

library(tidyr)
dfq1 <- data.frame(Yd1 = resid_estandarizados[,2],
                   Yd2 = resid_estandarizados[,1],
                   Yd3 = resid_estandarizados[,3])

dfq1_long <- dfq1 %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Residuo")

# Crear el boxplot para ambas variables
box <- ggplot(dfq1_long, aes(x = Variable, y = Residuo)) +
  geom_boxplot(fill = "grey50", color = "black", alpha = 0.8) +  # Fondo gris oscuro
  labs(title = "Residuos Estandarizados", x = "Variable", y = "Residuo") +
  scale_y_continuous(limits = c(-6, 6)) +
  geom_hline(yintercept = 3, color = "red", lty = 2) +
  geom_hline(yintercept = 0, color = "black", linetype = "dashed") +
  geom_hline(yintercept = -3, color = "red", lty = 2) +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    panel.grid.major = element_line(color = "grey"),
    panel.grid.minor = element_line(color = "grey"),
    panel.border = element_rect(color = "black", fill = NA, size = 1)
  )

box
```

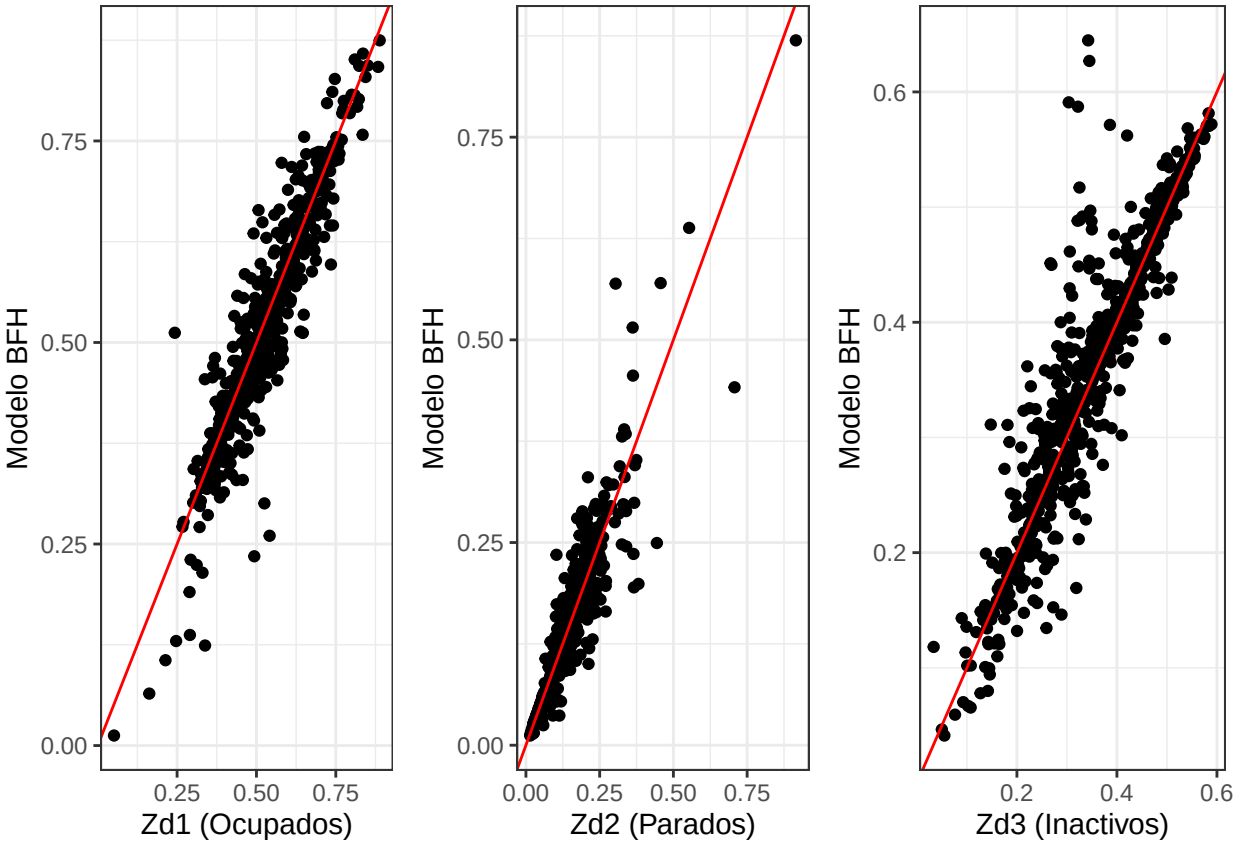


Al analizar los boxplots de los residuos, observamos que, con excepción de unos pocos valores atípicos, la mayoría de los residuos se encuentran dentro del rango $[-3, 3]$. Esto indica que el modelo ajustado logra capturar adecuadamente la variabilidad de los datos y que los errores están mayoritariamente distribuidos de manera razonable.

La concentración de los residuos en este intervalo sugiere que no existen grandes discrepancias entre los valores observados y los predichos por el modelo, lo que refuerza su calidad y ajuste. Los pocos datos fuera de este rango pueden ser indicativos de áreas específicas o patrones particulares que podrían requerir un análisis más detallado para comprender mejor las fuentes de variación en esos casos concretos.

```
######################## MODELO VS DIRECTO ##############################

data1 <- data.frame(x = df_plot$pd1, y = df_plot$pparados)

dirvseblup1 = ggplot(data1, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", size = .5) +
  labs(x = "Zd2 (Parados)", y = "Modelo BFH") +theme_bw()


data2 <- data.frame(x = df_plot$pd2, y = df_plot$pocupados)

dirvseblup2 = ggplot(data2, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", size = .5) +
  labs(x = "Zd1 (Ocupados)", y = "Modelo BFH") +theme_bw()
```

```r
df_plot$pd3 <- 1 - df_plot$pd1 - df_plot$pd2

data3 <- data.frame(x = df_plot$pd3, y = df_plot$pinactivos)

dirvseblup3 = ggplot(data3, aes(x = x, y = y)) +
  geom_point()  +
  geom_abline(intercept = 0, slope = 1, color = "red", size = .5) +
  labs(x = "Zd3 (Inactivos)", y = "Modelo BFH") +theme_bw()


grid.arrange(dirvseblup2, dirvseblup1, dirvseblup3, ncol=3)
```



Además, al analizar los gráficos que comparan los estimadores directos frente a los obtenidos mediante el modelo, observamos que los puntos tienden a agruparse en torno a la bisectriz. Esto indica una buena concordancia entre ambos métodos, lo que refuerza la validez del modelo ajustado para reflejar la realidad de los datos. La alineación cercana a la bisectriz sugiere que el modelo es capaz de corregir posibles sesgos y mejorar la precisión en las estimaciones.

Por otro lado, al examinar los boxplots de las predicciones de las proporciones de ocupados, parados e inactivos y compararlos con los valores obtenidos de manera directa, se evidencian mejoras significativas. Las predicciones modeladas muestran menor dispersión y, en general, reducen la variabilidad asociada a los estimadores directos. Esto es especialmente importante en áreas con tamaños muestrales reducidos, donde los estimadores directos suelen ser menos precisos. Las mejoras observadas en los boxplots reflejan que el modelo logra captar mejor las relaciones subyacentes en los datos y proporciona estimaciones más consistentes y fiables.

En cuanto a la tasa de paro, observamos una clara mejoría, especialmente porque, al utilizar el modelo, prácticamente todos los valores cumplen con el criterio de la ONS, que establece que las estimaciones deben

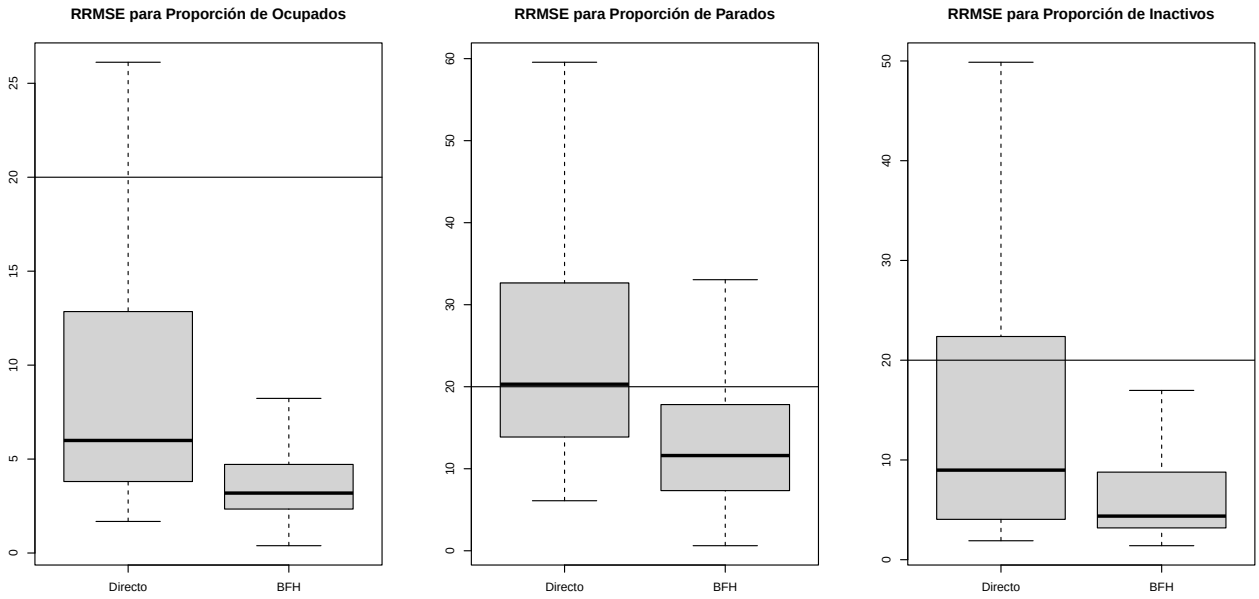


Figure 1: RRMSE de la proporción de parados y ocupados.

tener CVs menores al 20%.

```
### Escribimos los datos convenientemente.
s12 = df_plot$s12; s11 = df_plot$s11;  s22 = df_plot$s22
var.parados = s11; var.ocupados = s22; cov.paro = s12;
ocupados = df_plot$pocupados; parados = df_plot$pparados

### Aplicamos las expresiones y la fórmula de Särndal para la tasa de desempleo.
tparo.dir = df_plot$pparados_nocorr /
  (df_plot$pparados_nocorr + df_plot$pocupados_nocorr)
msetp.dir = ((ocupados^2*var.parados)/
               (parados + ocupados)^4)  +
  (((parados^2)*var.ocupados)/(parados + ocupados)^4)  -
  2*((ocupados*parados*cov.paro)/(parados + ocupados)^4)
cv.tp.dir = sqrt(msetp.dir)/tparo.dir

### Creamos la tabla con todo lo que nos interesa.
tabla1 = data.frame(Nacionalidad = df_plot$EXT.x,
                    Sexo = df_plot$SEXO,
                    PROV = as.numeric(df_plot$CPRO2.x),
                    Trimestre = df_plot$PK_TRIM,

                    nd = df_plot$nd.x,

                    par = df_plot$pparados_nocorr,  parBFH = df_plot$pd1,
                    CVdirpar = df_plot$cv2_nocorr_sivar,  CVinpar = df_plot$CVeb.x,

                    ocu = df_plot$pocupados_nocorr, ocuBFH = df_plot$pd2,
                    CVdirocu = df_plot$cv1_nocorr_sivar, CVinocu = df_plot$CVeb.y,

                    ina = df_plot$pinactivos_nocorr,
                    inaBFH = 1 - df_plot$pd1 - df_plot$pd2,
```

```
                      CVina = df_plot$cv3_nocorr_sivar,
                      CVinaBFH = 100*sqrt(df_plot$MSE_ina.x)/
                        abs(1-df_plot$pd1-df_plot$pd2),

                      tp = 100*df_plot$pparados_nocorr/
                        (df_plot$pparados_nocorr+df_plot$pocupados_nocorr),
                      tpBFH = 100*df_plot$pd1/(df_plot$pd1+df_plot$pd2),

                      cv.tp = 100*cv.tp.dir,
                      cv.tp.BFH = 100*sqrt(df_plot$MSEparo.y)/
                        (df_plot$pd1/(df_plot$pd1+df_plot$pd2)),
                      Nd = df_plot$pop, Nteor = df_plot$pop_padron,
                      Neff = df_plot$ocupados + df_plot$parados + df_plot$inactivos
)

### Añadimos el texto correspondiente relacionado con si es o no nacional
### y el nombre oficial de la provincia

tabla1 <- tabla1 %>%
  mutate(País = case_when(
    Nacionalidad == 0 ~ "Español",
    Nacionalidad == 1 ~ "Extranjero"))

tabla1 <- tabla1 %>%
  mutate(PROV_nombre = case_when(
    PROV == 2 ~ "Albacete",
    PROV == 3 ~ "Alicante/Alacant",
    PROV == 4 ~ "Almería",
    PROV == 1 ~ "Araba/Álava",
    PROV == 33 ~ "Asturias",
    PROV == 5 ~ "Ávila",
    PROV == 6 ~ "Badajoz",
    PROV == 7 ~ "Balears, Illes",
    PROV == 8 ~ "Barcelona",
    PROV == 48 ~ "Bizkaia",
    PROV == 9 ~ "Burgos",
    PROV == 10 ~ "Cáceres",
    PROV == 11 ~ "Cádiz",
    PROV == 39 ~ "Cantabria",
    PROV == 12 ~ "Castellón/Castelló",
    PROV == 13 ~ "Ciudad Real",
    PROV == 14 ~ "Córdoba",
    PROV == 15 ~ "Coruña, A",
    PROV == 16 ~ "Cuenca",
    PROV == 20 ~ "Gipuzkoa",
    PROV == 17 ~ "Girona",
    PROV == 18 ~ "Granada",
    PROV == 19 ~ "Guadalajara",
    PROV == 21 ~ "Huelva",
    PROV == 22 ~ "Huesca",
    PROV == 23 ~ "Jaén",
    PROV == 24 ~ "León",
    PROV == 25 ~ "Lleida",
```

```
    PROV == 27 ~ "Lugo",
    PROV == 28 ~ "Madrid",
    PROV == 29 ~ "Málaga",
    PROV == 30 ~ "Murcia",
    PROV == 31 ~ "Navarra",
    PROV == 32 ~ "Ourense",
    PROV == 34 ~ "Palencia",
    PROV == 35 ~ "Palmas, Las",
    PROV == 36 ~ "Pontevedra",
    PROV == 26 ~ "Rioja, La",
    PROV == 37 ~ "Salamanca",
    PROV == 38 ~ "Santa Cruz de Tenerife",
    PROV == 40 ~ "Segovia",
    PROV == 41 ~ "Sevilla",
    PROV == 42 ~ "Soria",
    PROV == 43 ~ "Tarragona",
    PROV == 44 ~ "Teruel",
    PROV == 45 ~ "Toledo",
    PROV == 46 ~ "Valencia/València",
    PROV == 47 ~ "Valladolid",
    PROV == 49 ~ "Zamora",
    PROV == 50 ~ "Zaragoza",
    PROV == 51 ~ "Ceuta",
    PROV == 52 ~ "Melilla",
    TRUE ~ NA_character_
  ))
```

Ahora, construyamos la tabla final para visualizarla.

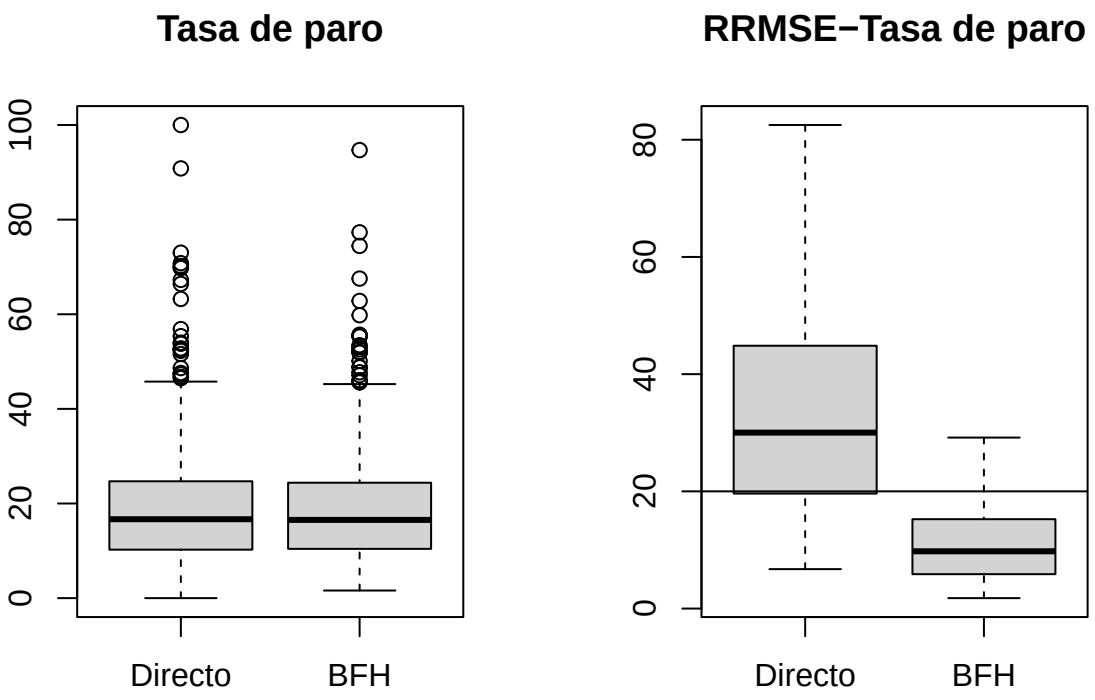
```
### Perfilamos la tabla.

tabla_final <- tabla1 %>% mutate(Total.Par = round(par*Neff),
                                 Total.Par.BFH = round(parBFH*Neff),
                                   Total.Ocu = round(ocu*Neff),
                                 Total.Ocu.BFH = round(ocuBFH*Neff),
                                   Total.Ina = round(ina*Neff),
                                 Total.Ina.BFH = round(inaBFH*Neff))

# write.xlsx(tabla_final, file = "datos_finales_todos2.xlsx")

head(tabla_final)
```

```
##   Nacionalidad Sexo PROV Trimestre  nd        par     parBFH CVdirpar CVinpar
## 1            0    1    1       121 544 0.04449879 0.04530882 22.28133    8.70
## 2            0    6    1       121 542 0.04272383 0.04292070 20.92697    8.43
## 3            1    1    1       121  52 0.22051336 0.19031913 28.14576   14.77
## 4            1    6    1       121  69 0.16407729 0.18728590 27.83527   14.78
## 5            0    1    2       121 622 0.09523355 0.09559929 13.93885    8.85
## 6            0    6    2       121 653 0.10107089 0.10161243 12.60754    8.88
##         ocu    ocuBFH  CVdirocu CVinocu       ina    inaBFH     CVina  CVinaBFH
## 1 0.5264277 0.5317392  4.187971    2.57 0.4290735 0.4229520  5.070942  3.718850
## 2 0.4665026 0.4674284  4.708220    3.19 0.4907735 0.4896509  4.488244  3.553871
## 3 0.5930052 0.6051877 12.031346    5.07 0.1864814 0.2044932 29.270660 11.293528
```

```
## 4 0.4320548 0.5071319 14.552739    5.32 0.4038679 0.3055822 15.488071  9.999773
## 5 0.5253054 0.5264696  4.084675    2.89 0.3794611 0.3779311  5.400328  4.760079
## 6 0.4119613 0.4129654  5.104544    3.58 0.4869678 0.4854222  4.301993  3.920587
##          tp     tpBFH    cv.tp cv.tp.BFH        Nd    Nteor      Neff
## 1  7.794137  7.851828 46.15410  6.981455 116103.15 36847.19 116103.15
## 2  8.389947  8.410068 47.47214  6.034025 120138.81 36847.19 120138.81
## 3 27.106125 23.924261 29.27008 14.469550  16835.08 36847.19  16835.08
## 4 27.523645 26.970206 31.42674 12.186107  20162.51 36847.19  20162.51
## 5 15.346910 15.367958 19.98500  7.416916 149621.64 25429.78 149621.64
## 6 19.700694 19.746756 17.63392  6.682475 151408.87 25429.78 151408.87
##         País PROV_nombre Total.Par Total.Par.BFH Total.Ocu Total.Ocu.BFH
## 1    Español Araba/Álava      5166          5260     61120         61737
## 2    Español Araba/Álava      5133          5156     56045         56156
## 3 Extranjero Araba/Álava      3712          3204      9983         10188
## 4 Extranjero Araba/Álava      3308          3776      8711         10225
## 5    Español    Albacete     14249         14304     78597         78771
## 6    Español    Albacete     15303         15385     62375         62527
##   Total.Ina Total.Ina.BFH
## 1     49817         49106
## 2     58961         58826
## 3      3139          3443
## 4      8143          6161
## 5     56776         56547
## 6     73731         73497
```

Y veamos el boxplot comparativo de la tasa de desempleo del directo y del modelo.

```
### Hacemos los boxplots de la tasa de desempleo.

par(mfrow=c(1,2))
boxplot(100*tparo.dir,tabla1$tpBFH, names=c("Directo","BFH"),
        main="Tasa de paro", ylim=c(0,100))
boxplot(tabla1$cv.tp, tabla1$cv.tp.BFH, names=c("Directo","BFH"),
        main="RRMSE-Tasa de paro",outline=F)
abline(a=20,b=0)
```




```
par(mfrow=c(1,1))
```

Por último, procedemos a analizar los resultados obtenidos a través de los mapas generados. Antes de ello, asegurémonos de cargar las librerías necesarias, instalándolas previamente en caso de no estar disponibles en el entorno de trabajo. Este paso garantizará que todas las herramientas requeridas para la visualización y análisis estén correctamente configuradas.

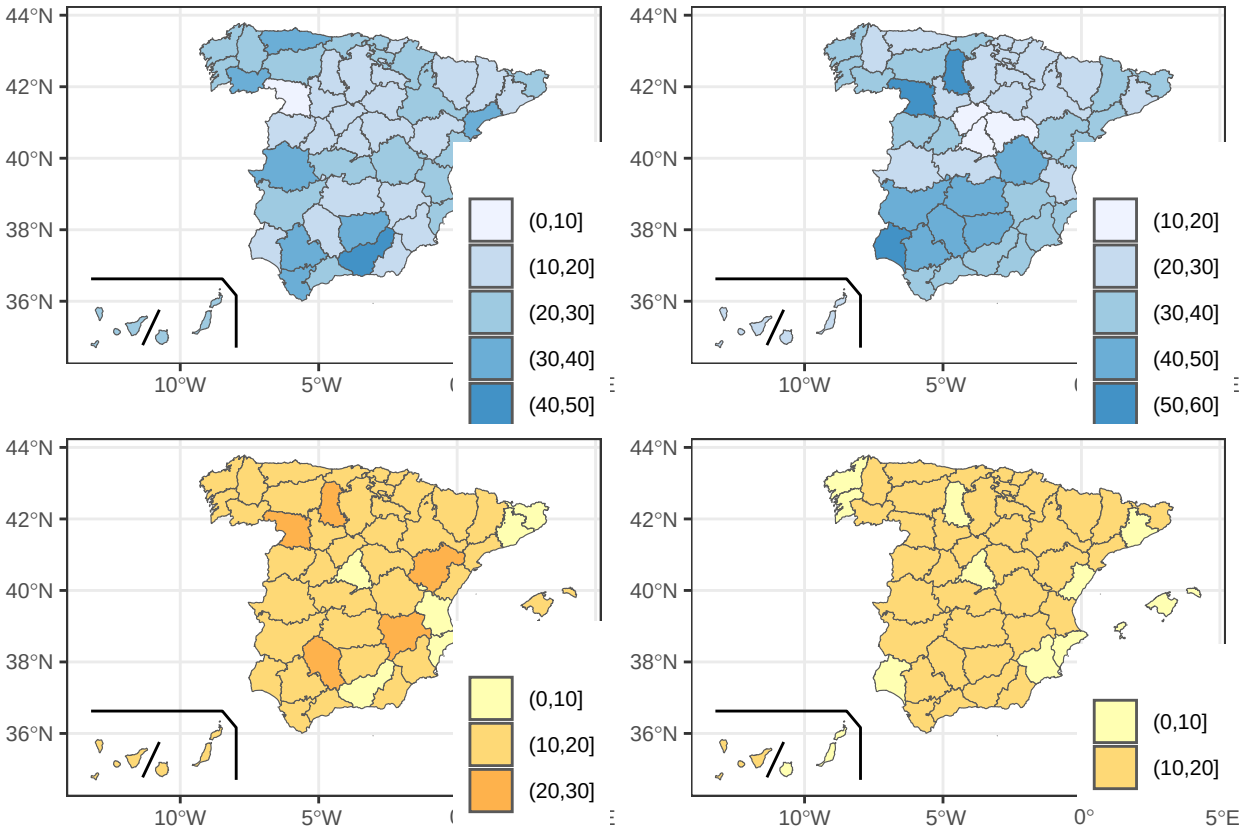
```
# Cargar librerías necesarias

# Lista de todas las librerías necesarias
paquetes <- c(
  "mapSpain", "tidyverse", "RColorBrewer", "openxlsx", "gridExtra"
)

# Instalación y carga de paquetes
for (paquete in paquetes) {
  if (!require(paquete, character.only = TRUE)) {
    install.packages(paquete, dependencies = TRUE)
    library(paquete, character.only = TRUE)
  }
}

# Mensaje de confirmación
cat("Todos los paquetes se han cargado correctamente.\n")
```

Tomaremos como ejemplo el primer trimestre de 2021 para nuestro mapeo.

```
## Estimacion de totales de ocupados, parados e inactivos y tasas de paro
## segun la list.results$Nacionalidad extranjera
```

```r
library(mapSpain)
library(tidyverse)
library(RColorBrewer)
library(openxlsx)
library(gridExtra)

df.orig <- read.xlsx("datos_finales_todos2.xlsx")

list.results <- rbind(df.orig[df.orig$Trimestre==121,])
list.results <- list.results %>%   filter(!(PROV %in% c(51, 52)))

######### MAPEO DEL CASO: EXTRANJERO ############

nuts2.aux <- data.frame('cpro' = list.results$PROV,
                        'pred.umpl.rate'= list.results$tpBFH,
        'breaks.pred.umpl.rate' = cut(
          list.results$tpBFH,
          breaks=seq(0, 70,
                     by=10)))[list.results$Sexo==1 &
                                list.results$Nacionalidad==1, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

extr_h <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"Blues")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

####
####

nuts2.aux <- data.frame('cpro' = list.results$PROV,
                        'pred.umpl.rate'= list.results$tpBFH,
                        'breaks.pred.umpl.rate' = cut(
                          list.results$tpBFH,
breaks=seq(0, 70, by=10)))[list.results$Sexo==6 & list.results$Nacionalidad==1, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

extr_m <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"Blues")[1:7]))+
```

```
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

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



nuts2.aux <- data.frame('cpro' = list.results$PROV,
                        'pred.umpl.rate'= list.results$cv.tp.BFH,
                        'breaks.pred.umpl.rate' = cut(
                          list.results$cv.tp.BFH,
breaks=seq(0, 50, by=10)))[list.results$Sexo==1 & list.results$Nacionalidad== 1, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

cvextr_h <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"YlOrRd")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

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



nuts2.aux <- data.frame('cpro' = list.results$PROV,
                        'pred.umpl.rate'= list.results$cv.tp.BFH,
                        'breaks.pred.umpl.rate' = cut(
                          list.results$cv.tp.BFH,
breaks=seq(0, 50, by=10)))[list.results$Sexo==6 & list.results$Nacionalidad== 1, ]
nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

cvextr_m <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"YlOrRd")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

grid.arrange(extr_h, extr_m, cvextr_h, cvextr_m, ncol=2, nrow=2)
```



Vemos que, para el caso de la población no nacional, empezando por la fila superior que estima la tasa de desempleo,tenemos que:

– Columna izquierda (Hombres): El mapa muestra la tasa de desempleo de los hombres por provincia. Se observa que las tasas varían entre las diferentes provincias, con algunas áreas más oscuras que indican tasas de desempleo más altas (por ejemplo, superiores al 40%). Las provincias con tonos más claros tienen tasas de desempleo más bajas, situándose en el rango del 0-10%.

– Columna derecha (Mujeres): En el caso de las mujeres, las tasas de desempleo también presentan variabilidad geográfica. Las tasas más altas (40-50% o más) están concentradas en ciertas provincias, mientras que otras presentan valores más bajos. En general, se puede observar un patrón de desigualdad geográfica, con algunas provincias más afectadas que otras.

En cuanto a la columna inferior, relacionada con los CVs, tenemos que:

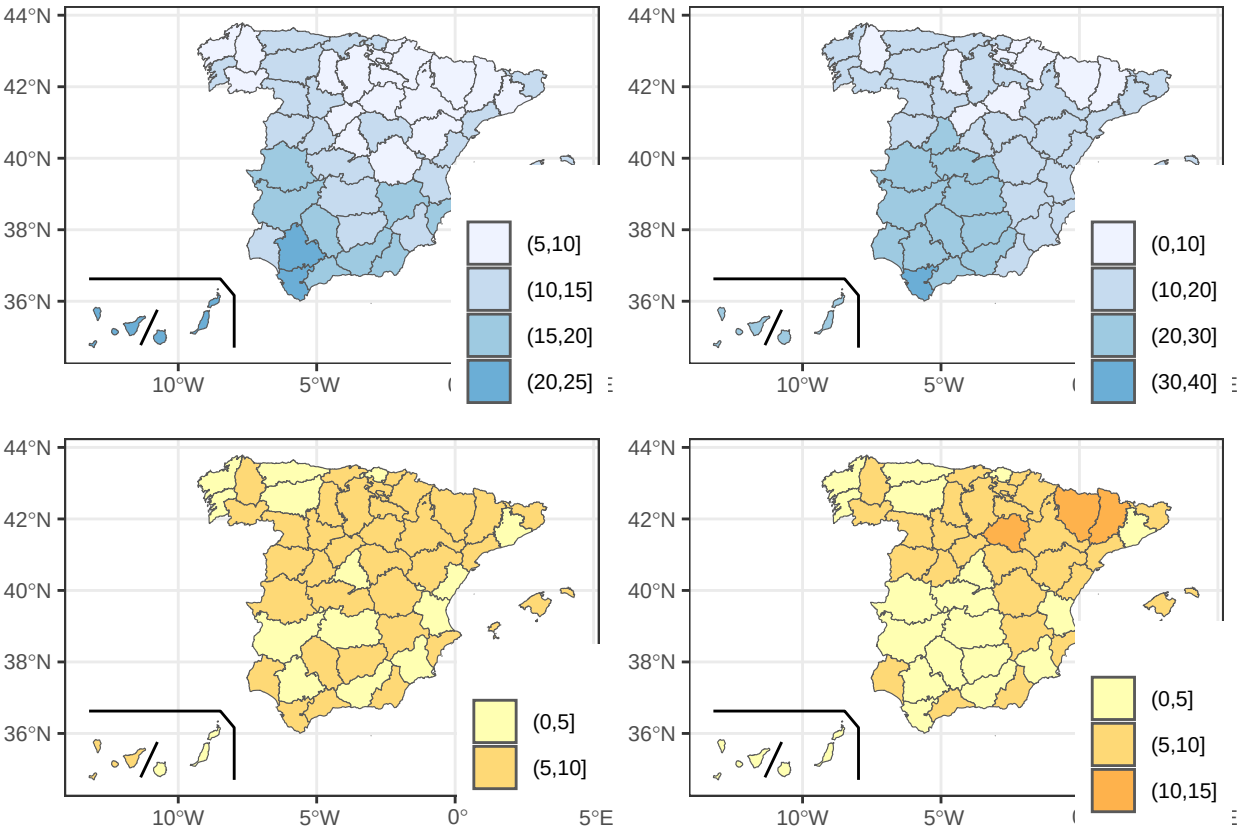
– Columna izquierda (Hombres): El mapa representa los coeficientes de variación de las estimaciones de desempleo masculino. Un CV más bajo (tonos claros) indica mayor precisión en las estimaciones, mientras que los valores más altos (tonos oscuros) reflejan mayor incertidumbre. La mayoría de las provincias presentan valores en el rango del 10-20%, lo que sugiere que las estimaciones son relativamente precisas en general, aunque algunas áreas tienen valores más elevados.

– Columna derecha (Mujeres): Para las mujeres, los CVs también muestran variabilidad entre provincias. Se observa un patrón similar al de los hombres, con muchas provincias en el rango del 10-20%. Sin embargo, en algunas áreas específicas, los CVs son más altos, lo que indica una mayor incertidumbre en las estimaciones de desempleo femenino en esas provincias.

```
######### MAPEO DEL CASO: NACIONAL ###########


nuts2.aux <- data.frame('cpro' = list.results$PROV,
```

```
                        'pred.umpl.rate'= list.results$tpBFH,
                        'breaks.pred.umpl.rate' = cut(
                          list.results$tpBFH,
breaks=seq(0, 35, by=5)))[list.results$Sexo==1 & list.results$Nacionalidad==0, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

no_extr_h <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"Blues")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

####
####

nuts2.aux <- data.frame('cpro' = list.results$PROV,
                        'pred.umpl.rate'= list.results$tpBFH,
                        'breaks.pred.umpl.rate' = cut(
                          list.results$tpBFH,
breaks=seq(0, 40, by=10)))[list.results$Sexo==6 & list.results$Nacionalidad==0, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

no_extr_m <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"Blues")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

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



nuts2.aux <- data.frame('cpro' =list.results$PROV,
                        'pred.umpl.rate'= list.results$cv.tp.BFH,
                        'breaks.pred.umpl.rate' = cut(list.results$cv.tp.BFH,
breaks=seq(0, 20, by=5)))[list.results$Sexo==1 & list.results$Nacionalidad== 0, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)
```

```r
nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

cvno_extr_h <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"YlOrRd")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

####
####

nuts2.aux <- data.frame('cpro' = list.results$PROV,
                        'pred.umpl.rate'= list.results$cv.tp.BFH,
                        'breaks.pred.umpl.rate' = cut(
                          list.results$cv.tp.BFH,
breaks=seq(0, 20, by=5)))[list.results$Sexo==6 & list.results$Nacionalidad== 0, ]

nuts2 <- esp_get_prov(year='2021')
nuts2$cpro <- as.numeric(nuts2$cpro)

nuts2 <- merge(nuts2, nuts2.aux, by='cpro')

cvno_extr_m <- ggplot(nuts2) +
  geom_sf(data = esp_get_can_box()) +
  geom_sf(data = esp_get_can_provinces()) +
  geom_sf(aes(fill = breaks.pred.umpl.rate)) +
  scale_fill_manual(values = c(brewer.pal(7,"YlOrRd")[1:7]))+
  theme_bw() + theme(text = element_text(size=10), legend.position=c(.87,.2))+
  guides(fill=guide_legend(title=" "))

grid.arrange(no_extr_h, no_extr_m, cvno_extr_h, cvno_extr_m, ncol=2, nrow=2)
```



En cuanto al caso nacional, las conclusiones son análogas. Es interesante observar que, incluso aunque en las estimaciones directas no haya demasiados casos de CV elevado, el uso de la metodología de áreas pequeñas resulta en CVs muy pequeños, apenas superiores al 10% exceptuando casos puntuales, lo que incentiva su uso.

Conclusiones

Nuevamente, la implementación de los métodos BFH es sencilla al igual que lo fue bajo el modelo UFH. En efecto, y gracias a los códigos implementados, podemos obtener las estimaciones de forma rápida y con bajo CV.

Sea como fuere, y vistos los resultados comparativos entre el modelo BFH y las estimaciones directas, el uso de dicho modelo es especialmente recomendado para obtener resultados con bajo CV y estimaciones análogas (gracias a la conformidad) a las directas.

A su vez, la ventaja en este caso de los modelos de área es que podemos modelar la correlación de ambas variables de interés, por ejemplo parados y ocupados, lo que supone un enfoque más realista en este sentido que el modelo a nivel de individuo.

8.2. Funciones de R desarrolladas

En esta sección final las funciones empleadas escritas en R desarrolladas propiamente específicas para la estimación en áreas pequeñas. Para el resto de las relativas a la SAE, se remite al lector al GitHub del libro de Morales et al. (2021), donde se detallan: https://github.com/small-area-estimation/sae-book, o a las librerías de `R` y sus manuales del CRAN, como es el caso del paquete `sae`.

8.2.1. Código R del modelo FH univariante

A continuación se presenta el código realizado para la resolución del problema FH univariante (llamado en el código UFH por simplicidad).

8.2.1.1. Código R de la función `UFH.NINIS`

```
UFH.NINIS <- function(TAU, excels = F, B_boot = 100) {
  # Filtrar datos para el trimestre TAU
  dfninis.temp <- dfninis[dfninis$TRIM == TAU, ]

  ########## BLOQUE TRANSFORMACIÓN VARIANZA ALR ##########
  # Asegurar proporciones positivas antes de ALR
  sel.pprop <- dfninis.temp$snini_16a24 > 0
  dfninis.temp$snini_16a24[!sel.pprop] <-
  min(dfninis.temp$snini_16a24[sel.pprop])

  q     <- 2
  H0.el <- q * (diag(q - 1) +
  matrix(1, q - 1, q - 1) %*% matrix(1, q - 1, q - 1))

  Ved         <- list()
  D           <- nrow(dfninis.temp)
  Ved_alr     <- vector("list", D)
  snini_16a24_alr <- numeric(D)

  for (d in 1:D) {
    Ved[[d]]       <- as.numeric(dfninis.temp$snini_16a24[d])
    Ved_alr[[d]]   <- H0.el %*% Ved[[d]] %*% t(H0.el)
    snini_16a24_alr[d] <- Ved_alr[[d]]
  }

  dfninis.temp$snini_16a24_alr <- snini_16a24_alr
  rm(H0.el)
  ########################################################

  # Bloque de variables directas
  directy1 <- dfninis.temp$ydi_comp
  vardiry1 <- dfninis.temp$snini_16a24_alr
  directy2 <- dfninis.temp$ydi
  vardiry2 <- dfninis.temp$snini_16a24

  # Bloque de covariables
  covariables <- with(dfninis.temp, data.frame(
  pestudios1, pestudios2, pestudios3,
  pparados, pocupados, pinactivos, pinmig, tp_epa
  ))

```

```
variables  <- data.frame(directy1, vardiry1, covariables)
variables2 <- data.frame(directy2, vardiry2, covariables)

# Calcular correlaciones y exportar si es necesario
correlaciones.y1 <- cor(variables)[, 1]
correlaciones.y2 <- cor(variables2)[, 1]
corrs <- cbind(
names = colnames(variables),
Y1    = correlaciones.y1,
Y2    = correlaciones.y2
)
if (excels) {
  write.xlsx(corrs, file = paste0('Resultados/tabla_correlaciones_', TAU, '.xlsx'))
}

# Bloque de regresión UFH (Fay?Herriot)
fmod <- mseFH(
directy1 ~ pestudios1 + pestudios2 +
pocupados + pinactivos + pinmig - 1,
vardiry1,
MAXITER = 1e5
)

fit1  <- fmod$est$fit$estcoef
fitty <- as.data.frame(t(fit1))

names2 <- data.frame(Covars = c(
'Primarios', 'Secundarios', 'Ocupados', 'Inactivos', 'Inmig'
))

tabla_significacion_ufh <- data.frame(
names2,
Lím.Inf. = fitty$beta - abs(qt(0.025, D - 1)) * fitty$std.error,
Beta     = fitty$beta,
Lím.Sup. = fitty$beta + abs(qt(0.025, D - 1)) * fitty$std.error,
`Error Estándar` = fitty$std.error,
Theta    = fmod$est$fit$refvar,
`p-valor`= fitty$pvalue,
Sign.    = fitty$pvalue < 0.05
)
if (excels) {
  write.xlsx(
  tabla_significacion_ufh,
  file = paste0('Resultados/tabla_significaciones_ufh_ALR', TAU, '.xlsx')
  )
}

# Predicción EBLUP y cálculo de CVs
eblup <- eblupFH(
directy1 ~ pestudios1 + pestudios2 +
pocupados + pinactivos + pinmig - 1,
vardiry1,
MAXITER = 1e5
)

```

```r
sigma_u    <- eblup$fit$refvar
beta_tilde <- eblup$fit$estcoef$beta

X <- lapply(1:D, function(d) {
  t(bdiag(as.numeric(c(
  dfninis.temp$pestudios1[d],
  dfninis.temp$pestudios2[d],
  dfninis.temp$pocupados[d],
  dfninis.temp$pinactivos[d],
  dfninis.temp$pinmig[d]
  ))))
})
y <- lapply(directy1, function(v) matrix(v))

# Cálculo de efectos aleatorios y predicciones
u.teor <- lapply(1:D, function(d) {
  (sigma_u / (sigma_u + Ved[[d]])) * (y[[d]] - X[[d]] %*% beta_tilde)
})

mudb  <- pidkb <- vector("list", D)
pd1   <- pd2 <- numeric(D)
for (d in 1:D) {
  mudb[[d]]  <- X[[d]] %*% beta_tilde + u.teor[[d]]
  pidkb[[d]] <- exp(mudb[[d]]) / (1 + sum(exp(mudb[[d]])))
  pd1[d]      <- pidkb[[d]][1]
  pd2[d]      <- 1 - pd1[d]
}

# Bootstrap interno para MSEs
# Nota: BOOT.compo1d <- function(...) {   /* ... */     }
mses <- BOOT.compo1d(X, D, Ved, sigma_u, beta_tilde, B = B_boot)

# Bloque de benchmarking
lambdas <- data.frame(ccaa = dfninis.temp$CCAA, dir = directy2, ind=pd1,
    CPRO2 = dfninis.temp$CPRO2)
lambdas.2 <- lambdas %>% group_by(ccaa) %>% summarise(L = sum(dir, na.rm = T
    )/sum(ind, na.rm = T))
lambdas.3 <- left_join(lambdas, lambdas.2, by = c("ccaa"))

Ved_bench <- list()
directy3 <- c(); directy4 <- c()
vardiry3 <- c();
varb <- c()

for (d in 1:D) {
  directy3[d] <- log(pd1[d]*lambdas.3$L[d]/(1-pd1[d]*lambdas.3$L[d]))
  varb[d] <- mses[[2]][d]*(lambdas.3$L[d])^2
  Ved_bench[[d]] <- varb[d]
  vardiry3[d] <- Ved_bench[[d]]
  directy4[d] <- pd1[d]*lambdas.3$L[d]
}

# Cálculo de coeficientes de variación
CVdir1   <- round(100 * sqrt(vardiry2) / abs(directy2), 10)
CV.fh.1  <- round(100 * sqrt(mses[[2]]) / abs(pd1), 10)
CV.bench <- CV.fh.1 * lambdas.3$L
```

```
151
152   # Construir tabla de salida
153   output <- data.frame(
154   Prov            = dfninis.temp$CPRO2,
155   ccaa            = dfninis.temp$CCAA,
156   trim            = dfninis.temp$TRIM,
157   nd              = dfninis.temp$muestra,
158   nini_nd         = dfninis.temp$n_nini,
159   nd_16a24        = dfninis.temp$n_16a24,
160   L               = lambdas.3$L,
161   DIR             = round(directy2, 25),
162   Vdir            = round(vardiry2, 25),
163   CVdir           = CVdir1,
164   mufh            = round(sapply(mudb, '[', 1), 25),
165   EBfh            = round(pd1, 25),
166   MSEfh           = round(mses[[1]], 25),
167   CVfh            = CV.fh.1,
168   pdbench         = directy4,
169   msebench        = round(mses[[2]] * lambdas.3$L^2, 25),
170   CVbench         = CV.bench,
171   Nombre          = dfninis.temp$Nombre_Provincial,
172   Poblacion_EPA   = dfninis.temp$pop,
173   Poblacion_16a24= dfninis.temp$pop_16a24
174   )
175
176   if (excels) {
177     write.xlsx(output, file = paste0('Resultados/resultadosninis_', TAU, '.
          xlsx'))
178   }
179
180   cat("\nFin Trimestre", TAU, "\n")
181   return(list(
182   NINIS = output,
183   TC    = corrs,
184   TS.UFH= tabla_significacion_ufh
185   ))
186 }
```

Código 8.1: Función `UFH.NINIS` en R

8.2.1.2. Código R de la función `BOOT.compo1d`

```
1
2  BOOT.compo1d <- function(X, D, Vedb, thetasb, betasb, i = 0, B = 50, categs =
       1, MAX_ITER = 2000) {
3    # Preparar matrices y vectores
4    Xd             <- X
5    Vudb           <- thetasb                    # varianza entre áreas
6    edb <- udb    <- array(0, dim = c(D, categs)) # errores de muestreo y área
7    theta.gorro.ast <- numeric(B)                # varianzas estimadas bootstrap
8    beta.gorro.ast  <- vector("list", B)         # betas estimadas bootstrap
9    pidk.gorro.ast  <- mudk.gorro.ast <-
10   array(0, dim = c(categs, D, B))            # predicciones almacenadas
11
12   # Inicializar contadores y acumuladores
13   b       <- 0
```

```
  BadTot2 <- 0
  difmu  <- 0
  difpi  <- 0
  medias <- rep(0, categs)
  mudb   <- pidkb <-
  array(0, dim = c(categs, D, B))
  yb     <- vector("list", D)
  excepcion <- integer()
  u      <- vector("list", D)

  # Bucle de bootstrap
  while (b < B) {
    b <- b + 1

    # Simular errores de área (u) y de muestreo (e)
    for (d in 1:D) {
      edb[d, ] <- rnorm(categs, mean = medias, sd = sqrt(Vedb[[d]]))
      udb[d, ] <- rnorm(categs, mean = medias, sd = sqrt(Vudb))
    }

    # Simular y* y calcular mu db
    for (d in 1:D) {
      mudb[, d, b] <- X[[d]] %*% betasb + udb[d, ]
      yb[[d]]      <- mudb[, d, b] + edb[d, ]
    }

    # Calcular proporciones simuladas
    for (d in 1:D) {
      pidkb[, d, b] <- exp(mudb[, d, b]) / (1 + sum(exp(mudb[, d, b])))
    }

    ##### Ajuste del modelo en bootstrap
    # Crear variables de covariables a partir de X
    nombres_variables <- c("pestudios1", "pestudios2", "pocupados", "
        pinactivos", "pinmig")
    for (j in seq_along(nombres_variables)) {
      assign(nombres_variables[j],
      sapply(X, function(m) m[, j]))
    }

    # Preparar vector de respuesta
      yb.fit <- sapply(yb, function(matriz) matriz[,1])


    # Intentar ajustar modelo eblupFH
    fit2 <- try(
    eblupFH(yb.fit ~ pestudios1 + pestudios2 + pocupados +
    pinactivos + pinmig - 1,
    vardiry1, MAXITER = MAX_ITER, PRECISION = 0.01),
    silent = TRUE
    )

      if(class(fit2)=="try-error"){
      # Si falla, registrar excepción y repetir iteración
      excepcion <- c(excepcion, b)
      write.table(data.frame(error = class(fit2), D, j, b),
```

```
      file = "WARNING.txt", append = TRUE, col.names = FALSE)
      b <- b - 1
      BadTot2 <- BadTot2 + 1
    } else if (fit2$fit$iterations < MAX_ITER) {
      # Extraer varianza y betas estimadas
      theta.gorro.ast[b]   <- fit2$fit$refvar
      beta.gorro.ast[[b]]  <- fit2$fit$estcoef$beta
      sigma_u              <- theta.gorro.ast[b]

      # Calcular efectos aleatorios u para cada área
      for (d in 1:D) {
        u[[d]] <- (sigma_u / (sigma_u + Vedb[[d]])) *
        (yb[[d]] - X[[d]] %*% beta.gorro.ast[[b]])
      }

      # Guardar predicciones mudk y pidk
      for (d in 1:D) {
        mudk.gorro.ast[, d, b] <- X[[d]] %*% beta.gorro.ast[[b]] + u[[d]]
        pidk.gorro.ast[, d, b] <- exp(mudk.gorro.ast[, d, b]) /
        (1 + sum(exp(mudk.gorro.ast[, d, b])))
      }

      # Acumular diferencias para MSE
      difmu <- difmu + (mudk.gorro.ast[, , b] - mudb[, , b])^2
      difpi <- difpi + (pidk.gorro.ast[, , b] - pidkb[, , b])^2
    } else {
      # Iteración excede MAX_ITER, repetir
      b <- b - 1
      BadTot2 <- BadTot2 + 1
    }
  }

  # Calcular MSE promedio
  mse_mudk_ast <- difmu / B
  mse_pidk_ast <- difpi / B

  # Devolver listas con MSEs
  return(list(mse_mudk_ast, mse_pidk_ast))
}
```

Código 8.2: Función `BOOT.compo1d` en R

8.2.2. Código R del modelo BFH

A continuación, se presenta el código realizado para la resolución del problema FH bivariante, del BFH. Salvo las tres funciones que siguen, el resto se pueden obtener del GitHub que se puso anteriormente, del libro de Morales et al. (2021).

8.2.2.1. Código R de la función `BFH.PAR.OCU`

```
BFH.PAR.OCU <- function(TAU, flag = F, B_Boot = 1000) {
  # Cargar y filtrar datos para el trimestre TAU, y funciones externas.
  source('SAE19_BFH_Functions.R')

  df <- df[df$PK_TRIM == TAU, ]
```

```r
print(TAU)
D <- length(unique(df$dom))
d <- D

# Preparar vectores y listas de respuesta multivariada
y <- lapply(1:D, function(d) {
  matrix(c(df$yd1[d],
  df$yd2[d]))
})

# Evitar ceros en varianzas y lambdas
sel.pprop <- df$s11_alr > 0
sel.pgap  <- df$s22_alr > 0
df$s11_alr[!sel.pprop] <- min(df$s11_alr[sel.pprop])
df$s22_alr[!sel.pgap]  <- min(df$s22_alr[sel.pgap])

# Transformación ALR de la matriz de varianzas
q     <- 3
H0.el <- q * (diag(q-1) +
matrix(1, q-1, q-1) %*% matrix(1, q-1, q-1))

Ved      <- vector("list", D)
H0       <- vector("list", D)
Ved_alr  <- vector("list", D)
s11_alr  <- numeric(D)
s12_alr  <- numeric(D)
s22_alr  <- numeric(D)

for (d in 1:D) {
  H0[[d]]      <- H0.el
  Ved[[d]]     <- matrix(NA, 2, 2)
  sup          <- as.numeric(df$s12[d])
  Ved[[d]][upper.tri(Ved[[d]])] <- sup
  Ved[[d]][lower.tri(Ved[[d]])] <- sup
  diag(Ved[[d]]) <- c(df$s11[d], df$s22[d])

  Ved_alr[[d]] <- H0.el %*% Ved[[d]] %*% t(H0.el)
  s11_alr[d]   <- Ved_alr[[d]][1, 1]
  s12_alr[d]   <- Ved_alr[[d]][1, 2]
  s22_alr[d]   <- Ved_alr[[d]][2, 2]
}

Ved.log <- Ved
Ved     <- Ved_alr

# Definir variables directas y sus varianzas
directy1 <- df$yd1
directy2 <- df$yd2
vardiry1 <- df$s11_alr
vardiry2 <- df$s22_alr

# Coeficientes de variación directos
CVdir1 <- sqrt(vardiry1) / abs(directy1)
CVdir2 <- sqrt(vardiry2) / abs(directy2)

# Añadir columnas al data frame
```

```
df$directy1 <- directy1
df$vardiry1 <- vardiry1

# Definir covariables
afiliados_ss                <- df$porcentaje_afiliadosss_epa
renta_neta_media_persona    <- df$rnmp / max(df$rnmp)
p1625                       <- df$porcentaje_16a25_epa
p2645                       <- df$porcentaje_26a45_epa
p4664                       <- df$porcentaje_46a65_epa
p65plus                     <- df$porcentaje_65ymas_epa

covariables <- data.frame(
afiliados_ss, renta_neta_media_persona,
p1625, p2645, p4664, p65plus
)

# Correlaciones con cada variable objetivo
variables  <- data.frame(directy1, vardiry1, covariables)
variables2 <- data.frame(directy2, vardiry2, covariables)

correlaciones.y1 <- data.frame(cor(variables)[, 1])
correlaciones.y2 <- data.frame(cor(variables2)[, 1])
corrs <- cbind(
data.frame(colnames(variables)),
correlaciones.y1,
correlaciones.y2
)
colnames(corrs) <- c('names', 'Y1', 'Y2')

# Construcción de la lista de diseño X para BFH
X <- lapply(1:D, function(d) {
  as.matrix(t(bdiag(
  as.numeric(c(afiliados_ss[d],
  renta_neta_media_persona[d],
  p1625[d],
  p2645[d],
  p65plus[d])),
  as.numeric(c(1,
  afiliados_ss[d],
  renta_neta_media_persona[d],
  p1625[d],
  p2645[d],
  p65plus[d]))
  )))
})

# Nombres de covariables para tablas
names2 <- data.frame(
Covariables = c(
'%AfilSS', 'RentaNetaMediaPersona', '%Edad_16_25',
'%Edad_26_45', '%Edad_65_plus', 'Intercept',
'%AfilSS', 'RentaNetaMediaPersona',
'%EPA_Edad_16_25', '%Edad_26_45', '%Edad_65_plus'
)
)
```

```r
# Resolver valores iniciales de thetas vía índices de trimestre
thetas_bfh.k1 <- c(0.072, 0.3, 0.1, 0.1, .1, .1, .1, .1)
thetas_bfh.k2 <- c(0.02, 0.1, 0.06, 0.1, .1, .1, .07, .07)
Index <- switch(as.character(TAU),
'121' = 1, '122' = 2, '221' = 3, '222' = 4,
'321' = 5, '322' = 6, '421' = 7, '422' = 8
)
refvark1 <- thetas_bfh.k1[Index]
refvark2 <- thetas_bfh.k2[Index]
thetas.0 <- c(refvark1, refvark2, 0)
Vud <- UveU(thetas.0)

# Ajuste REML para BFH
fit <- try(REML.BFH(X, y, D, Ved, Vud, MAXITER = 1e10), TRUE)
cat("BFH model parameters", fit[[1]], "\n")
cat("Number of iterations", fit[[3]], "\n")
cat("Errors: ", fit[[4]], "\n")

# Extraer betas, thetas y efectos aleatorios
beta.u.hat <- BETA.U.BFH(X, y, D, Ved, fit[[1]])
beta.hat   <- beta.u.hat[[1]]
theta.hat  <- fit[[1]]
u          <- beta.u.hat[[2]]
pv         <- pvalue(beta.hat, fit)
betas      <- data.frame(beta.hat, pv, Sig = pv[[3]] < 0.05)
C.I        <- CI(beta.hat, fit)

# Tabla de betas con intervalos y significación
tabla_significacion <- data.frame(
names2,
Lím. Inf.      = C.I[[1]][, 1],
Beta           = beta.hat,
Lím. Sup.      = C.I[[1]][, 2],
`Error Estándar` = round(pv, 5),
Estadístico    = pv,
`p-valor`      = pv,
Sign.          = pv[[3]] < 0.05
)
colnames(tabla_significacion) <- c(
'Covars', 'Lím. Inf.', 'Beta', 'Lím. Sup.',
'Error Estándar', 'Estadístico', 'p-valor', 'Sign.'
)

# Tabla de thetas
names2_theta <- data.frame(Covars = c('Theta1', 'Theta2', 'Theta3'))
tabla_significacion_thetas <- data.frame(
names2_theta,
Lím. Inf. = C.I[[2]][, 1],
Theta     = theta.hat,
Lím. Sup. = C.I[[2]][, 2]
)

# Cálculo de EBLUPs
eblup.bfh   <- mapply("+", lapply(X, "%*%", beta.hat), u)
eblup.bfh.1 <- eblup.bfh[1, ]
eblup.bfh.2 <- eblup.bfh[2, ]
```

```r

# Bootstrap para MSEs
source("Boot_Compo.R")
library(MASS)
mses <- BOOT.compo2d(X, D, Ved, theta.hat, beta.hat,
i = 0, B = B_Boot, categs = 2)

# Cálculo de proporciones y CVs
categs <- 3
mudb  <- pidkb <- vector("list", D)
pd1   <- pd2 <- pd3 <- numeric(D)
for (d in 1:D) {
  mudb[[d]]  <- X[[d]] %*% beta.hat + u[[d]]
  pidkb[[d]] <- exp(mudb[[d]]) / (1 + sum(exp(mudb[[d]])))
  pd1[d]     <- pidkb[[d]][1]
  pd2[d]     <- pidkb[[d]][2]
  pd3[d]     <- 1 - pd1[d] - pd2[d]
}

# Coeficientes de variación directos y BFH
CVdir1   <- round(100 * sqrt(df$s11) / abs(df$pparados), 2)
CVdir2   <- round(100 * sqrt(df$s22) / abs(df$pocupados), 2)
CV.bfh.1 <- round(100 * sqrt(mses[[2]][1, ]) / abs(pd1), 2)
CV.bfh.2 <- round(100 * sqrt(mses[[2]][2, ]) / abs(pd2), 2)

# Preparar salidas
nd.true <- df$n
cvs <- cbind(
sexo = df$SEXO, CPRO2 = df$CPRO2, EXT = df$EXT,
nd = nd.true, CVdir1, CVdir2, CV.bfh.1, CV.bfh.2
)

estims <- cbind(df$pparados, pd1, df$pocupados, pd2)

# Construcción de tablas de resultados
output1 <- data.frame(
dominio   = df$dom, sexo = df$SEXO, CPRO2 = df$CPRO2,
EXT       = df$EXT, nd = nd.true, trim = df$PK_TRIM,
DIR       = round(df$yd1, 5),
dir1      = round(df$pparados, 5),
Vdir      = round(df$s11, 5),
CVdir     = CVdir1,
pd1       = pd1,
EB        = round(eblup.bfh.1, 5),
MSEeb     = round(mses[[2]][1, ], 5),
CVeb      = CV.bfh.1,
MSEparo   = mses[[3]],
MSE_ina   = mses[[4]]
)

output2 <- data.frame(
dominio   = df$dom, sexo = df$SEXO, CPRO2 = df$CPRO2,
EXT       = df$EXT, nd = nd.true, trim = df$PK_TRIM,
DIR       = round(df$yd2, 5),
Vdir      = round(df$s22, 5),
CVdir     = CVdir2,
```

```
  pd2        = pd2,
  EB         = round(eblup.bfh.2, 5),
  MSEeb      = round(mses[[2]][2, ], 5),
  CVeb       = CV.bfh.2,
  MSEparo    = mses[[3]],
  MSE_ina    = mses[[4]]
  )
  head(output1, 10)
  head(output2, 10)

  # Guardar resultados si flag = TRUE
  if (flag) {
    write.xlsx(corrs, file = paste0('results_alr/tabla_correlaciones_', TAU, '.xlsx'))
    write.xlsx(tabla_significacion, file = paste0('results_alr/tabla_significaciones_betas_bfh_', TAU, '.xlsx'))
    write.xlsx(tabla_significacion_thetas, file = paste0('results_alr/tabla_significaciones_thetas_bfh_', TAU, '.xlsx'))
    write.xlsx(output1, file = paste0('results_alr/resultadosalrparados_', TAU, '.xlsx'))
    write.xlsx(output2, file = paste0('results_alr/resultadosalrocupados_', TAU, '.xlsx'))
  }

  # Retornar lista de resultados
  return(list(
  PAR        = output1,
  OCU        = output2,
  TS         = tabla_significacion,
  TC         = corrs,
  TS_theta   = tabla_significacion_thetas
  ))
}
```

Código 8.3: Función `BFH.PAR.OCU` en R

8.2.2.2. Código R de la función `BOOT.compo2d`

```

BOOT.compo2d <- function(X, D, Vedb, thetasb, betasb, i, B = 50, categs = 3) {
  # Cargar funciones necesarias
  source("Estimacion␣BETA.R")
  source('SAE19_BFH_Functions.R')

  Xd   <- X
  Vudb <- UveU(thetasb)           # varianza de efectos aleatorios

  # Inicializar arrays para errores y predicciones
  edb   <- udb   <- array(0, dim = c(D, categs))
  theta.gorro.ast <- vector("numeric", B)
  beta.gorro.ast  <- vector("list", B)
  pidk.gorro.ast  <- mudk.gorro.ast <- array(0, dim = c(categs, D, B))

  # Contadores y acumuladores de errores
  b          <- 0
  BadTot2    <- 0
```

```
  difmu      <- 0
  difpi      <- 0
  difpi_ina  <- 0
  diftp      <- 0
  medias     <- rep(0, categs)

  # Arrays para almacenar simulaciones
  mudb         <- pidkb <- array(0, dim = c(categs, D, B))
  pidkb_ina    <- array(0, dim = c(1, D, B))
  pidk.gorro.ast_ina <- array(0, dim = c(1, D, B))
  yb           <- vector("list", D)
  excepcion    <- integer()

  # Calcular medias para evitar ceros extremos
  media.par  <- mean(sapply(Vedb, function(x) x[1, 1]), na.rm = TRUE)
  media.ocu  <- mean(sapply(Vedb, function(x) x[2, 2]), na.rm = TRUE)
  media.corr <- mean(sapply(Vedb, function(x) x[1, 2]), na.rm = TRUE)

  # Ajustar Vedb para casos degenerados
  for (d in 1:D) {
    if (abs(Vedb[[d]][1, 1] - Vedb[[d]][2, 2]) < 1e-20) {
      Vedb[[d]][1, 2] <- Vedb[[d]][2, 1] <- media.corr
      Vedb[[d]][1, 1] <- media.par
      Vedb[[d]][2, 2] <- media.ocu
    }
    if (det(Vedb[[d]]) < 1e-20) {
      Vedb[[d]][1, 2] <- Vedb[[d]][2, 1] <- 0
    }
  }

  # Bucle principal de bootstrap
  while (b < B) {
    b <- b + 1

    # Simular errores de muestreo (edb) y de área (udb)
    for (d in 1:D) {
      edb[d, ] <- mvrnorm(1, medias, Vedb[[d]])
      udb[d, ] <- mvrnorm(1, medias, Vudb)
    }

    # Simular la variable objetivo y calcular mudb
    for (d in 1:D) {
      mudb[, d, b] <- X[[d]] %*% betasb + udb[d, ]
      yb[[d]]      <- mudb[, d, b] + edb[d, ]
    }

    # Calcular proporciones simuladas pidkb y pidkb_ina
    for (d in 1:D) {
      pidkb[, d, b]       <- exp(mudb[, d, b]) / (1 + sum(exp(mudb[, d, b])))
      pidkb_ina[, d, b]   <- 1 - sum(pidkb[, d, b])
    }

    # Ajuste del modelo BFH sobre la muestra bootstrap
    fit2 <- try(
    REML.BFH(Xd, yb, D, Vedb, Vudb, MAXITER = 1e5),
    silent = TRUE
```

```r
    )

      if(class(fit2)=="try-error"){
      # Registrar error y repetir iteración
      excepcion <- c(excepcion, b)
      write.table(
      data.frame(error = class(fit2), D, i, b),
      file = "WARNING.txt", append = TRUE, col.names = FALSE
      )
      b       <- b - 1
      BadTot2 <- BadTot2 + 1
    } else if (fit2[[3]] < 100) {
      # Extraer varianza estimada y calcular betas y u
      theta.gorro.ast[b]    <- fit2[[1]]
      theta.ast             <- theta.gorro.ast[b]
      beta.gorro.ast[[b]]   <- BETA.U.compo(X, yb, D, Vedb, theta.ast)
      betas.gorro.ast       <- as.array(beta.gorro.ast[[b]][[1]])
      u                     <- beta.gorro.ast[[b]][[2]]

      # Estimar mudk.gorro.ast y pidk.gorro.ast
      for (d in 1:D) {
        mudk.gorro.ast[, d, b]     <- X[[d]] %*% betas.gorro.ast + u[[d]]
        pidk.gorro.ast[, d, b]     <- exp(mudk.gorro.ast[, d, b]) /
        (1 + sum(exp(mudk.gorro.ast[, d, b])))
        pidk.gorro.ast_ina[, d, b] <- 1 - sum(pidk.gorro.ast[, d, b])
      }

      # Acumular errores para MSE
      difmu     <- difmu     + (mudk.gorro.ast[, , b] - mudb[, , b])^2
      difpi     <- difpi     + (pidk.gorro.ast[, , b] - pidkb[, , b])^2
      difpi_ina <- difpi_ina + (pidk.gorro.ast_ina[, , b] - pidkb_ina[, , b])
          ^2
      diftp     <- diftp     + (
      (pidk.gorro.ast[2, , b] /
      (pidk.gorro.ast[2, , b] + pidk.gorro.ast[1, , b])) -
      (pidkb[2, , b] / (pidkb[2, , b] + pidkb[1, , b]))
      )^2
    } else {
      # Iteración sin convergencia suficiente, repetir
      b       <- b - 1
      BadTot2 <- BadTot2 + fit2[[4]]
    }
  }

  # Calcular MSE promedio para cada componente
  mse_mudk_ast  <- difmu     / B
  mse_pidk_ast  <- difpi     / B
  mse_pidk_tp   <- diftp     / B
  mse_pidk_ina  <- difpi_ina / B

  # Devolver lista con MSEs
  return(list(mse_mudk_ast, mse_pidk_ast, mse_pidk_tp, mse_pidk_ina))
}
```

Código 8.4: Función `BOOT.compo2d` en R

8.2.2.3. Código R de la función `BETA.U.compo`

Es importante mencionar que esta función es una adaptación de una cedida por Esther López-Vizcaíno, la cual no está disponible en ningún medio on-line y de ahí que se añade al TFM, mas no es propia.

```
BETA.U.compo <- function(X, y, D, Ved, theta.hat) {
  # Preparar variables
  p         <- ncol(X[[1]])            # número de covariables
  Vd.inv    <- vector("list", D)       # inversas de V_d
  Vd.gorro  <- vector("list", D)       # V_d + V_u
  Xd        <- X                       # lista de matrices X_d
  yd        <- y                       # lista de vectores y_d

  # Calcular V_u + V_d y sus inversas
  Vudbeta <- UveU(theta.hat)           # varianza efecto aleatorio
  for (d in 1:D) {
    Vd.gorro[[d]] <- Vudbeta + Ved[[d]]
    Vd.inv[[d]]   <- solve(Vd.gorro[[d]])
  }

  # Ensamblar sumas para beta
  Q.inv <- matrix(0, nrow = p, ncol = p)  # acumulador Q^-1
  XVy   <- numeric(p)                     # acumulador X^t inv(V) y
  for (d in 1:D) {
    Q.inv <- Q.inv + t(Xd[[d]]) %*% Vd.inv[[d]] %*% Xd[[d]]
    XVy   <- XVy   + t(Xd[[d]]) %*% Vd.inv[[d]] %*% yd[[d]]
  }

  # Resolver beta_hat = solve(Qinv) %*% XVy

  Q     <- solve(Q.inv)
  betax <- Q %*% XVy

  # Calcular efectos aleatorios u_hat = V_u %*% solve(V_u+V_d) %*% (y_d - X_d
      %*% beta_hat)

  u <- vector("list", D)
  for (d in 1:D) {
    u[[d]] <- Vudbeta %*% Vd.inv[[d]] %*% (yd[[d]] - X[[d]] %*% betax)
  }

  # Retornar lista: beta y lista de u
  return(list(betax, u))
}
```

Código 8.5: Función `BETA.U.compo` en R

Bibliografía

[1] Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 139-160.

[2] Arima, S., Datta, G. S., & Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. Scandinavian Journal of Statistics, 42(2), 518-529.

[3] Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), 28-36.

[4] Benavent, R., & Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. Computational Statistics & Data Analysis, 94, 372-390.

[5] Berg, E., & Fuller, W. A. (2012). Small area prediction under a Fay-Herriot model with preliminary testing for the presence of random effects. Survey Methodology, 38(1), 1-10.

[6] Best, D. J., & Roberts, D. E. (1975). Algorithm AS 89: the upper tail probabilities of Spearman's rho. Journal of the Royal Statistical Society. Series C (Applied Statistics), 24(3), 377-379.

[7] Boubeta, M., Lombardía, M. J., & Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. TEST, 25(3), 548-569.

[8] Boubeta, M., Lombardía, M. J., Morales, D., Pérez, A., & Santamaría, L. (2016a). Small area estimation of labour force indicators under a multinomial logit mixed model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(2), 535-558.

[9] Boubeta, M., Lombardía, M. J., Morales, D., & Santamaría, L. (2017). Small area estimation of unemployment proportions under a multinomial mixed model with correlated time and area effects. TEST, 26(3), 615-640.

[10] Boubeta, M., Lombardía, M. J., Morales, D., & Santamaría, L. (2023). Generalized mixed models for small area estimation of poverty indicators. Journal of Statistical Planning and Inference, 223, 45-63.

[11] Burgard, J., Do Kim, J., & Lahiri, P. (2020). Measurement error in small area estimation: A review. Journal of Survey Statistics and Methodology, 8(1), 1-27.

[12] Burgard, J. P., Münnich, R., & Schmid, T. (2021). Small area estimation of compositional data using Dirichlet-multinomial models. Computational Statistics & Data Analysis, 157, 107159.

[13] Bugallo, M., Marey-Pérez, M. F., Morales, D., & Esteban, M. D. (2024a) Zero-Inflated Negative Binomial Mixed Models for Predicting Number of Wildfires. Journal of Environmental Management, 328, 116788.

[14] Bugallo, M., González, D. M., Salvati, N., & Francesco, S. S. (2024b). Temporal M-quantile models and robust bias-corrected small area predictors. `arXiv preprint arXiv:2407.09062.`

[15] Chandra, H., Salvati, N., & Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. Spatial Statistics, 20, 30-56.

[16] Csiszár, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, Series A, 8:85-108.

[17] Das, K., Jiang, J., & Rao, J. N. K. (2004). Mean squared error estimation of empirical predictor. Annals of Statistics, 32(2), 818-840.

[18] Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica, 10(3), 613-627.

[19] Datta, G. S., Rao, J. N. K., & Torabi, M. (2011). Pseudo-empirical Bayes estimation of small area means under a nested error linear regression model with functional measurement error. Journal of Statistical Planning and Inference, 141(10), 3388-3396.

[20] Datta, G. S., Hall, P., Mandal, A., & Mukherjee, K. (2018). Model selection using multiple testing with applications to small-area estimation. Statistical Science, 33(3), 356-372.

[21] Demidenko, E. (2013). Mixed models: theory and applications with R. John Wiley & Sons.

[22] Diao, L., Smith, P. W., & Fuller, W. A. (2014). Small area estimation with measurement error. Biometrika, 101(4), 1001-1016.

[23] Diz-Rosales, N., Lombardía, M. J., & Morales, D. (2024). Poverty mapping under area-level random regression coefficient Poisson models. Journal of Survey Statistics and Methodology, 12(2), 404-434.

[24] Dogan, & Kiliç, (2014). A comparative study on variance components estimation methods. Düzce Üniversitesi Salk Bilimleri Enstitüsü Dergisi, 1(2), 9-14.

[25] Erciulescu, A. L., & Fuller, W. A. (2015). Small area prediction of the mean of a binomial random variable. Small area prediction based on unit level models when the covariate mean is measured with error, JSM, Survey Research Methods Section 17.

[26] Esteban, M. D., Morales, D., Pérez, A., & Santamaría, L. (2012a). Small area estimation of poverty proportions under area-level time models. Computational Statistics & Data Analysis, 56(10), 2840-2855.

[27] Esteban, M. D., Morales, D., Pérez, A., & Santamaría, L. (2020). Small area estimation of compositions using a trivariate Fay-Herriot model. TEST, 29(4), 1085-1113.

[28] Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2023). Small area estimation of average compositions under multivariate nested error regression models. TEST, 32(4):1-26.

[29] Faltys, M., Münnich, R., & Schmid, T. (2022). Generalized linear mixed models for small area estimation. Journal of the Royal Statistical Society: Series A (Statistics in Society), 185(1), 152-179.

[30] Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366a), 269-277.

[31] Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. Statistical Science, 9(1), 55-76.

[32] González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Estimation of the mean squared error of predictors based on the Fay-Herriot model. Journal of Statistical Planning and Inference, 138(2), 308-321.

[33] González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2010). Bootstrap mean squared error of a small-area EBLUP. Journal of Statistical Computation and Simulation, 80(5), 449-462.

[34] Hájek, J. (1971). Comment on An Essay on the Logical Foundations of Survey Sampling, Part One. In: Godambe, V. P., & Sprott, D. A. (Eds.), The Foundations of Survey Sampling (pp. 236). Holt, Rinehart, and Winston, New York.

[35] Hall, P., & Maiti, T. (2006a). Nonparametric estimation of mean squared prediction error in small area estimation. Annals of Statistics, 34(5), 2299-2321.

[36] Hall, P., & Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2), 221-238.

[37] Hansen, M. H., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. The Annals of Mathematical Statistics, 14(4), 333-362.

[38] Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics, 9, 226-252.

[39] Henderson, C. R. (1975). Best linear unbiased estimation and prediction under selection model. Biometrics, 31, 423-427.

[40] Hobza, T., & Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. Journal of Official Statistics, 32(3), 661-692.

[41] Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685.

[42] Huang, E. T., & Bell, W. R. (2004). An empirical best prediction approach to small area estimation using survey and satellite data. Journal of Official Statistics, 20(3), 333-350.

[43] Instituto Nacional de Estadística (INE). (2021). Encuesta de Población Activa: Metodología 2021. Madrid: INE. Recuperado de https://www.ine.es/metodologia/t20/t203030471.pdf

[44] Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. The Annals of Statistics, 24(1), 255-286.

[45] Jiang, J. (1997). Wald consistency and the method of sieves in REML estimation. The Annals of Statistics, 25(4), 1781-1803.

[46] Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. Statistica Sinica, 861-885.

[47] Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. TEST, 15, 1-96.

[48] Jiang, J., & Nguyen, T. (2007). Linear and generalized linear mixed models and their applications (Vol. 1). New York: Springer.

[49] Jiang, J., & Tang, C. Y. (2011). The adjusted maximum likelihood method for small-area estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 831-851.

[50] Katz, V. J. (2009). A history of mathematics: An introduction. Addison-Wesley.

[51] Krause, T., Schmid, T., & Münnich, R. (2022b). Penalized multivariate Fay-Herriot models for small area estimation. Statistical Modelling, 22(4), 493-516.

[52] Kubokawa, T. (2011). Prediction in small area estimation: A conditional approach. Journal of the Japanese Statistical Society, 41(1), 69-91.

[53] Lombardía, M. J., Morales, D., & Santamaría, L. (2017). AIC-type criteria for Fay-Herriot model selection. Computational Statistics & Data Analysis, 105, 83-95.

[54] Lombardía, M. J., López-Vizcaíno, E., & Rueda, C. (2018). Selection of Small Area Estimators. Statistics and Applications Volume 16 No. 1, 2018 (New Series), 269-288.

[55] López Vizcaíno, M. E. (2014). Small area estimation: an application to the estimation of the labour market variables in galician counties. Tesis, Universidade de Santiago de Compostela.

[56] López-Vizcaíno, E., Lombardía, M. J., & Morales, D. (2015). Multinomial-based small area estimation of labour force indicators. Statistical Modelling, 15(2), 153-176.

[57] Marchetti, S., Tzavidis, N., & Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. Computational Statistics & Data Analysis, 56(10), 2889-2902.

[58] Marhuenda, Y., Morales, D., & Pardo, M. C. (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics & Data Analysis, 58, 308-325.

[59] Marhuenda, Y., Morales, D., & Pardo, M. C. (2014). Information criteria for Fay-Herriot model selection. Computational Statistics & Data Analysis, 78, 151-165.

[60] Marhuenda, Y., Morales, D., & Pardo, M. C. (2016). On the construction of confidence intervals in small area estimation problems. Statistics and Computing, 26(2), 393-411.

[61] Methodologies, E. (2013). Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys.

[62] Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics, 38(3), 369-385.

[63] Molina, I., Rao, J. N. K., & Datta, G. S. (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects. Survey Methodology, 41(1), 1-19.

[64] Molina, I., & Marhuenda, Y. (2015). sae: An R package for small area estimation.

[65] Morales, D., Pagliarella, M. C., & Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. Statistics and Operations Research Transactions, 39(1), 19-34.

[66] Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). A course on small area estimation and mixed models. Methods, theory and applications in R.

[67] Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika, 58(3), 545-554.

[68] ONS (2006). Model-based estimates of ILO unemployment for LAD/UAs in Great Britain. Guide for users.

[69] Pfeffermann, D. (2002). Small area estimation: New developments and directions. International Statistical Review, 70(1), 125-143.

[70] Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science, 28(1), 40-68.

[71] Polettini, S. (2017). Bayesian hierarchical models for small area estimation: An overview. Statistical Modelling, 17(1-2), 1-24.

[72] Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, 85(409), 163-171.

[73] Pratesi, M. (Ed.). (2016). Analysis of Poverty Data by Small Area Estimation. Wiley, pp. 325-348.

[74] Rao, J. N., & Choudhry, G. H. (1995). Small Area Estimation: Overview and Empirical Study. Business Survey Methods, 527-542.

[75] Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. Survey Methodology, 25, 175-186.

[76] Rao, J. N. K. (2003). Small area estimation. Wiley.

[77] Rao, J. N. K. (2008). Some methods for small area estimation. Rivista internazionale di scienze sociali: 4, 2008, 387-406.

[78] Rao, J. N. K., & Molina, I. (2015). Small area estimation (2nd ed.). Wiley.

[79] R Development Core Team (2024). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

[80] Rubin, D. B. (1993). Statistical disclosure limitation. Journal of official Statistics, 9(2), 461-468.

[81] Särndal, C. E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling. Springer Science & Business Media.

[82] Slud, E. V., & Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(2), 239-257.

[83] Slud, E. V., & Maiti, T. (2011). Small-area estimation based on survey data from a left-censored Fay-Herriot model. Journal of statistical planning and inference, 141(11), 3520-3535.

[84] Sugasawa, S., & Kubokawa, T. (2015). Parametric transformed Fay-Herriot model for small area estimation. Computational Statistics & Data Analysis, 91, 27-40.

[85] Tzavidis, N., Marchetti, S., & Chambers, R. (2008). Robust prediction of small area means and distributions. Australian & New Zealand Journal of Statistics, 50(1), 1-23.

[86] Ybarra, L. M. R., & Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. Biometrika, 95(4), 919-931.

[87] Zhong, H., & Feng, X. (2023). An efficient and fast sparse grid algorithm for high-dimensional numerical integration. Mathematics, 11(19), 4191.