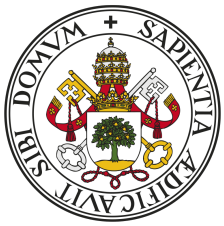

Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

SecCyL: Aplicación Shiny para la identificación de secciones censales similares en Castilla y León

Autora: **Estrella Santos Santos**

Tutores:
Agustín Mayo Íscar
Jorge Galván Del Río

2025

AGRADECIMIENTOS

A mis tutores, Agustín Mayo Íscar, por su paciencia y ayuda cada día de estos últimos meses, y Jorge Galván del Río, por enseñarme con tanta amabilidad cómo funciona el INE desde dentro y el importantísimo trabajo que realizan.
Gracias también a Raquel Álvarez Esteban, por acogerme como alumna de prácticas en la Delegación Provincial del INE de Valladolid.

A todos mis amigos que he conocido en Valladolid, en Bolonia y, por supuesto, a mis amigos de León. Gracias por vuestra confianza en mi.

A mis compañeros y amigos de Estadística e InDat, sin vosotros estos increíbles 4 años no hubieran sido lo mismo.

A mi familia, en especial a mis padres, por ayudarme y apoyarme en todos los momentos en los que lo he necesitado.

Gracias a todos.

Resumen

En las delegaciones provinciales del Instituto Nacional de Estadística, hay ocasiones en las que surge la necesidad de sustituir secciones censales por razones de seguridad, accesibilidad o cambios territoriales. Para dar soporte a este proceso, se ha diseñado una herramienta que, aplicando la distancia de Mahalanobis robusta (MCD) sobre indicadores transformados y estandarizados del Censo 2021, identifica de inmediato la sección más similar a la seleccionada. De este modo, se agiliza la toma de decisiones en la gestión de muestras y se garantiza la coherencia y representatividad del diseño muestral.

Palabras clave: Sección censal, similaridad, MCD (Minimum Covariance Determinant), Distancia de Mahalanobis, aplicación, Shiny.

Abstract

In the provincial delegations of the National Statistics Institute (INE), there are situations where it becomes necessary to replace census sections due to security, accessibility, or territorial changes. To support this process, a tool has been designed that, by applying the Robust Mahalanobis distance (MCD) to transformed and standardized indicators from the 2021 Census, immediately identifies the most similar section to the selected one. In this way, the decision-making process in sample management is streamlined, ensuring the coherence and representativeness of the sampling design.

Key words: census section, similarity, MCD (Minimum Covariance Determinant), Mahalanobis distance, application, Shiny.

Índice general

Índice de figuras

Capítulo 1

Introducción

Este proyecto se realiza por la necesidad de contar con una herramienta capaz de identificar, de forma rápida y fiable, la sección censal más parecida a una sección de referencia. Para ello, se ha desarrollado una aplicación que permite al usuario seleccionar cualquier sección de Castilla y León y, calculando la distancia de Mahalanobis sobre variables previamente transformadas y estandarizadas, obtiene al instante la sección más similar. Gracias al uso del estimador MCD (*Minimum Covariance Determinant*), las distancias no se ven distorsionadas por posibles valores atípicos en los datos. De esta manera, el personal de las delegaciones provinciales del INE puede sustituir secciones censales sin necesidad de escribir código, agilizando la toma de decisiones en la gestión de muestras y la sustitución de secciones censales cuando sea necesario.

1.1. Secciones censales

En España, una sección censal es un territorio que se encuentra dentro de un municipio. Todos los municipios están divididos en uno o más distritos, que a su vez se dividen en una o más secciones censales. Las secciones censales tienen su origen en las "secciones electorales" definidas por la Ley Orgánica 5/1985, de 19 de junio, del Régimen Electoral General, cuyo artículo 23 establece que cada sección debe agrupar entre 500 y 2.000 electores (Boletín Oficial del Estado, 1985). Aunque esa cifra es teórica –en la práctica se admiten ciertas variaciones; en Castilla y León se ha detectado al menos una sección con más de 3.000 habitantes–, sirve de referencia para el diseño muestral. Posteriormente, la Resolución de 17 de febrero de 2020, publicada en el BOE el 29 de abril de 2020, adaptó este límite para hablar de habitantes en lugar de electores, fijando un máximo de 2.500 habitantes por sección (Boletín Oficial del Estado, 2020a). Sin embargo, sigue tratándose de una recomendación y no de un límite estricto, de modo que las secciones pueden superar ligeramente ese umbral cuando las circunstancias así lo requieran. En algunas zonas rurales existe la posibilidad de que haya secciones con una población menor, en caso de que el municipio al que pertenezcan tenga menos de 500 habitantes.

Las secciones censales no son estructuras fijas. En muchas ocasiones, es necesario que el Ayuntamiento o la Delegación Provincial del INE lleve a cabo una reestructuración de sus secciones. Por ejemplo, creando una sección nueva, uniendo dos secciones existentes o modificando sus límites, por la necesidad de cumplir con los requisitos establecidos en cuanto a los habitantes/electores por sección censal. En el Censo de 2021, el cual se va a tomar como referencia en este trabajo, España estaba dividida en más de 36.000 secciones censales, de las que más de 3.500 pertenecen a Castilla y León. Son la unidad territorial mínima utilizada en la recogida de datos para las operaciones estadísticas como los Censos de Población y Viviendas, la Encuesta de Población Activa (EPA) o la Encuesta de Presupuestos Familiares (EPF).
Existe una excepción en el ámbito rural, ya que el *Nomenclátor* del INE contempla dos subdivisiones inframunicipales complementarias a la sección censal, que son las entidades singulares y los núcleos diseminados. De este modo, mientras que en entornos urbanos la sección censal es la unidad más pequeña, en las zonas rurales se pueden distinguir estos dos tipos de unidad para reflejar con más detalle la dispersión de la población. No obstante, en las encuestas oficiales se continúa utilizando únicamente las secciones censales como unidad de muestreo.

Cada sección censal se identifica con un código formado por 10 números que pertenecen a elementos imprescindibles: 2 dígitos para el código de provincia (CPRO), 3 dígitos para el municipio (CMUN), 2 dígitos para el distrito (CDIS) y, por último, 3 dígitos para la sección (CSEC), por lo que el resultado sería un código similar al siguiente: 4718607014, sección a la que pertenece el Campus Universitario Miguel Delibes.

Para consultar las secciones censales de la provincia de Valladolid, se puede acceder al portal GIS-Valladolid, en el apartado de distritos y secciones censales. [1]

Censo de Población y Viviendas 2021

En España, los censos de población y viviendas se realizan cada diez años. El último, correspondiente al año 2021, representa un gran cambio en el INE, por ser el primer censo basado mayoritariamente en el uso de registros administrativos, además de complementarse con una encuesta denominada Encuesta sobre Características Esenciales de la Población y las Viviendas (ECEPOV-2021), que solamente se realizó aproximadamente a un 1 % de la población (tamaño muy inferior al de los censos anteriores). Poder realizar este censo sin necesidad de realizar encuestas a gran parte de la población fue posible gracias a que en España se dispone de registros administrativos de calidad y base legal clara y suficiente para acceder a ellos.

La idea después de un censo tan novedoso es que a partir de 2021, se publiquen de manera no tan detallada censos de población anuales, como se menciona en la nota de prensa del INE titulada “Censo de población. 1 de enero de 2023”, publicada el 13 de diciembre de 2023. :

[1] https://www10.ava.es/cartografia/ficheros_pdf.html

> "Gracias a estos nuevos censos, se podrá disponer de información actualizada cada año, en el caso de la población, y cada tres o cuatro años en el caso de las viviendas, en lugar de cada 10, como sucedía con los censos decenales que se venían elaborando hasta ahora." (Instituto Nacional de Estadística, 2023)

No obstante, para este proyecto se ha decidido utilizar los indicadores del Censo de Población y Viviendas de 2021. Esta elección se debe a ciertas razones metodológicas. Por una parte, el Censo de 2021 es la fuente más completa, homogénea y coherente disponible a nivel de sección censal, con todas sus variables recopiladas con la misma fecha (1 de enero de 2021). Por otra parte, a pesar de la utilidad de disponer de datos actualizados, estos están distribuidos en diferentes fuentes y no todas las variables se encuentran disponibles para un mismo año o con el mismo nivel de desagregación. Además, dado que los límites de las secciones censales pueden variar con el tiempo, trabajar con un único momento de referencia para la mayor parte de los datos permite garantizar la consistencia territorial y evitar incompatibilidades en los análisis espaciales y cartográficos.
Es importante recalcar que la información a nivel de sección censal es relativamente consistente y comparable a lo largo del tiempo ya que las características territoriales de las secciones censales no cambian radicalmente de un año a otro. Esto permite que los datos de 2021 sean perfectamente válidos y útiles para este análisis.

Los ficheros de datos con los que se va a trabajar son registros estadísticos de población elaborados a partir del padrón y de su cruce con otras fuentes de registros administrativos, como el SEPE (Servicio Público de Empleo Estatal) o la TGSS (Tesorería General de la Seguridad Social), que constituyen la base poblacional sobre la que se construye el Censo de Población y Viviendas de 2021.

1.2. Motivación y objetivos del estudio

Este proyecto surge a partir de una propuesta planteada durante el periodo de prácticas externas realizado en la Delegación Provincial del Instituto Nacional de Estadística en Valladolid. En este contexto, se planteó la necesidad de desarrollar una herramienta que identificase la sección censal más parecida a una dada, en función de un conjunto de variables que representan las distintas características de la población y los hogares, como la actividad económica, la formación o la estructura del hogar. Además, se ha generado un indicador complementario que representa el contexto geográfico en el que se sitúan las secciones.

El interés por encontrar secciones censales similares se debe a las necesidades prácticas en la realización de determinadas encuestas oficiales, como la Encuesta de Población Activa (EPA) o la Encuesta de Presupuestos Familiares (EPF), que supone el desplazamiento del encuestador a los domicilios incluidos en la muestra. En determinadas circunstancias, algunas secciones pueden resultar problemáticas para ser visitadas, bien por encontrarse en **zonas consideradas conflictivas** o bien porque algún encuestador haya manifestado sentirse en riesgo durante visitas anteriores. Además, existen otras

situaciones que requieren la sustitución de una sección censal:

- Cambios en las **probabilidades muestrales**, que provocan la salida de ciertas secciones de la muestra.
- **Agotamiento de la sección**, en el momento en el que no quedan unidades de población (viviendas) disponibles para ser entrevistadas dentro de una sección censal.
- **Variaciones estructurales**, como cambios en los límites geográficos de una sección censal, ya sea por quedar por debajo o por encima de los límites de tamaño establecidos para las secciones.
- Secciones **ya incluidas en varias operaciones estadísticas**, lo que puede llevar a la saturación de los informantes. Cuando una sección está incluida en múltiples encuestas, existe el riesgo de entrevistar repetidamente a las mismas viviendas, por lo que conviene sustituirla para no sobrecargar a los entrevistados.

En estos casos, las delegaciones provinciales del INE pueden proponer a los Servicios Centrales de Madrid una sección alternativa, siempre que sea parecida a la original para no comprometer la representatividad y fiabilidad del diseño muestral inicial.

En este estudio se va a identificar la sección censal más similar a una dada, utilizando variables seleccionadas del Censo de 2021 y aplicando criterios estadísticos que justifican dicha similitud. El objetivo principal es proporcionar una herramienta que facilite a las delegaciones provinciales del INE identificar una sección alternativa adecuada, asegurando que no se vea afectada la representatividad ni la precisión del diseño muestral original.

Capítulo 2

Base de datos

En este capítulo se describe el tratamiento y el conjunto de datos utilizado para el análisis. Dado que el objetivo del trabajo consiste en el análisis de las secciones de Castilla y León, se trabaja con datos a nivel de sección censal. El hecho de que los datos pertenezcan a diferentes fuentes de datos hace necesaria la depuración y el acondicionamiento de los datos para su posterior análisis, con el fin de evitar información redundante.

2.1. Fuentes de información

Todos los datos utilizados en este proyecto se caracterizan por ser datos públicos de libre acceso, obtenidos a través de la página web del Instituto Nacional de Estadística. Estos datos se dividen en dos partes principales:

1. **Datos cartográficos:** Esta parte proviene de la cartografía digitalizada de secciones censales. Estos datos, disponibles en formato *shapefile*, contienen los contornos georreferenciados de todas las secciones censales a nivel nacional, según las coordenadas del sistema UTM (*Universal Transverse Mercator*) huso 28, 29, 30 y 31. Estos datos son fundamentales para facilitar las visualizaciones geográficas de las secciones censales y poder proporcionar una aplicación intuitiva y accesible para los usuarios que necesiten acceder a ella.

   Los datos de la cartografía digital del INE están compuestos por geometrías, que corresponden a los contornos georreferenciados de las secciones censales. Los datos asociados a estas geometrías proporcionan una tabla de correspondencia entre los códigos y los nombres asignados a cada región. Alguna de las variables más relevantes son las siguientes:

   - CPRO: código de dos dígitos que representa cada provincia
   - CUMUN: código que representa cada municipio, compuesto por cinco dígitos. Los dos primeros hacen referencia al código de la provincia (CPRO) y los tres restantes al municipio de la misma (CMUN)

- CCA: código de cada comunidad autónoma
- NMUN: nombre de cada municipio
- CUSEC: código que representa cada sección censal.

2. **Indicadores estadísticos:** La segunda parte de los datos incluye las variables relevantes para el análisis, que provienen de dos fuentes de datos principales:

   - Censo de Población y Viviendas 2021: De esta fuente se obtienen los indicadores para secciones censales, seleccionando los más relevantes para el objetivo buscado, la comparación entre secciones censales. Son fundamentales para definir las características de las secciones en el análisis.
   - Atlas de Distribución de Rentas (ADRH): Este conjunto de datos pertenece a una operación estructural de periodicidad anual, basada en la explotación de datos obtenidos a partir de registros administrativos. El fichero de datos con el que se va a trabajar pertenece a la Serie de los años 2015-2022, más concretamente al año 2021. Se ha decidido escoger los datos correspondientes a la renta neta media por unidad de consumo, ya que esta medida permite comparar el nivel de vida real entre los hogares al ajustarse a las diferentes estructuras familiares, además de ser la medida que se utiliza internacionalmente con el fin de obtener una mejor comparación de los ingresos individuales según distintos tipos de hogar.

   Esta medida se obtiene dividiendo los ingresos totales del hogar entre el número de unidades de consumo de este, cuyo valor se obtiene sumando los pesos de la escala modificada de la OCDE (Organización para la Cooperación y el Desarrollo Económicos): 1 para el sustentador principal, 0.5 para los siguientes adultos (mayores de 13 años) y 0.3 para los menores de 14 años.

   Es importante tener en cuenta que por secreto estadístico existe una acotación de los valores extremos de los indicadores de renta. Se calculan unos topes superiores e inferiores de forma que cuando el valor esté por encima de la cota superior o por debajo de la cota inferior, se censurará dicho valor por la cota, calculada a partir del percentil 0,1 –cota inferior– y el percentil 99,5 –cota superior– de cada sección censal.

Con los datos ya organizados, el siguiente paso consiste en la depuración y adecuación de los mismos. Esto permitirá asegurar que la información sea homogénea y esté preparada para su análisis, lo que facilitará su posterior uso en el desarrollo de la herramienta propuesta.

2.2. Manejo, depuración y adecuación de los datos

Una vez obtenidos los datos de las diferentes fuentes, es necesario llevar a cabo un proceso cuidadoso de depuración y adecuación. Dado que el enfoque principal de este estudio es el análisis de las secciones censales, el objetivo es construir una base de datos sólida y estructurada, donde cada fila corresponda a una sección censal de Castilla y

León, asegurando que los datos sean aptos para el análisis y su integración en la aplicación.

El primer paso consistió en filtrar los datos para seleccionar únicamente aquellos correspondientes a Castilla y León. Para ello, en los datos cartográficos, se ha utilizado la variable *CCA* (código 07); en los datos de renta, se extrajo la provincia a partir de los dos primeros dígitos del código *CUSEC*, lo que ha permitido filtrar las 9 provincias de Castilla y León. Finalmente, en los datos de Población y Viviendas se ha utilizado la variable *CCAA* para completar los datos.

Con los datos de Castilla y León, el siguiente paso ha consistido en explorar los indicadores del Censo de 2021, creando una nueva variable llamada *CUSEC*, correspondiente al código único de cada sección censal. Esta variable ha sido esencial para unir los tres conjuntos de datos (Censo, datos de renta y cartografía) mediante esta columna en común. Gracias a esta variable se ha podido establecer una correspondencia precisa entre las secciones censales de las distintas fuentes de datos, facilitando su integración para el análisis.

El Censo de 2021 cuenta con más de 20 variables, pero para facilitar el análisis se han seleccionado únicamente aquellas de mayor relevancia y coherencia con la información auxiliar del diseño muestral de la Encuesta de Presupuestos Familiares. Más concretamente:

- Total de habitantes.
- Porcentajes de población: menores de 16 años; entre 16 y 64 años; mayores de 64 años.
- Porcentaje de población extranjera.
- Porcentaje con estudios superiores (sobre población de 16+).
- Tasa de paro, tasa de empleo y tasa de actividad[1].
- Porcentaje de población estudiante (16+).
- Número de personas por hogar (1, 2, 3, 4, 5 o más personas), a partir del cual se ha creado la variable del tamaño medio de hogar.

Al combinar ambos conjuntos de datos estadísticos, han aparecido algunos problemas debido a la variabilidad en las secciones censales, pese a que todos los datos estuvieran referidos a 1 de enero de 2021. Para resolverlos, se ha tomado como referencia el listado de secciones que aparecen en el Censo de 2021 (que coincide con la cartografía), por ser de donde procede la mayor parte de la información. En los casos en los que el dato de Renta estuviera ausente, se ha imputado reemplazándolo por la media de la renta de

[1] Expresadas en porcentaje.

todas las secciones del distrito al que pertenecen.

Resulta importante destacar que, teniendo en cuenta el objetivo del trabajo, se ha tomado la decisión de prescindir de las secciones con menos de 100 habitantes. Esta decisión se ha tomado por dos motivos fundamentales. En primer lugar, la normativa sobre secreto estadístico establecida en el artículo 13.1 de la Ley de la Función Estadística Pública (LFEP) (Boletín Oficial del Estado, 1989), prohíbe la difusión de datos que puedan llevar a la identificación directa o indirecta de personas o entidades y, por tanto, debido a esta restricción, no se publica la información que sería necesaria para las secciones con una población inferior a 100 habitantes.

Además, este tipo de secciones tienen bajas probabilidades de ser seleccionadas en la muestra por el diseño muestral seguido tanto en EPA como en EPF.
Es un diseño muestral bietápico con estratificación, técnica utilizada para asegurarse de que las muestras obtenidas sean representativas de toda la población, dividiendo el proceso de selección en dos etapas. En la primera etapa las unidades de muestra son las secciones censales, que se agrupan en estratos en función del tamaño de la población a la que pertenecen.
En concreto, el estrato 6 agrupa a los municipios con menos de 10.000 habitantes, dentro de los cuales pueden encontrarse secciones censales con menos de 100 habitantes. Este tipo de secciones tienen una probabilidad extremadamente baja de ser seleccionadas debido a que, en Castilla y León, la mayor parte de los municipios tienen una población inferior a 10.000 habitantes, lo que implica que hay muchos más municipios entre los que se puede seleccionar la muestra.
Asimismo, dado que la selección se realiza de manera proporcional al tamaño poblacional, las secciones con menos habitantes tienen una menor probabilidad de ser incluidas en la muestra.

2.3. Creación de variables

Una vez construida la base de datos, con todo lo necesario obtenido a través de la página web del INE, fue necesario crear nuevas variables, tanto para facilitar el análisis, como para hacerlo más concreto, sintetizando información dispersa o generando nuevos indicadores territoriales con ayuda de los datos cartográficos disponibles.

Algunas de estas variables han sido diseñadas para reunir, en una única medida, información que en los indicadores del Censo estaba originalmente recogida en más variables. Para empezar, se ha creado la variable *t23*, que representa el número medio de personas por hogar en cada sección censal. Para ello, se ha utilizado la información disponible sobre el número de hogares según su tamaño, recogida en las variables *t22_1* a *t22_5*, que indican cuántos hogares hay con una, dos, tres, cuatro o cinco o más personas, respectivamente. Se ha optado por calcular una media ponderada de dichas variables en lugar de analizar cada grupo por separado, con el fin de tener un valor único que refleje el tamaño medio de los hogares en cada sección.

A todos los hogares se les asigna su número exacto, a excepción de los hogares de 5 o más personas (*t22_5*), a los que se les ha asignado un valor de 6 personas, como estimación conservadora.

Posteriormente, con el objetivo de crear un indicador de ruralidad, se ha creado la variable *t24*, medida en metros, que refleja la posición relativa de cada sección censal dentro de su contexto geográfico. Esta variable busca capturar un aspecto imprescindible para el análisis, el grado de "centralidad" o "aislamiento" de una sección. Dado que las secciones censales de Castilla y León presentan características diversas, no tiene sentido aplicar una única medida de distancia para todas ellas, por lo que se han creado tres variables intermedias.

- *dist_cap*: distancia desde cada sección al centro de la capital de provincia, calculada únicamente para las 9 capitales de provincia de Castilla y León.
- *dist_municip*: distancia desde cada sección al centro de su propio municipio, utilizada en municipios con más de 10.000 habitantes, sin ser capital de provincia.
- *dist_pueblo*: distancia desde secciones situadas en municipios pequeños (menos de 10.000 habitantes) al municipio más cercano con más de 10.000 habitantes.

Para medir la distancia a lo que se llama centro en *dist_cap* y *dist_municip* se ha tomado como referencia la sección 01001 de cada municipio.

Estas tres variables han sido unificadas en una sola, *t24*, que toma el valor que le corresponde en función del tipo de sección. De esta forma, se obtiene una variable que facilita la comparación entre las secciones, ya que se tiene una medida de distancia adaptada al contexto de cada una.

A continuación, se enumeran las variables finales incluidas en el análisis, que se pueden dividir en dos grupos principales, variables demográficas y variables socioeconómicas. Estas variables son fundamentales para identificar la situación de cada sección censal, ya que recogen aspectos clave de la situación laboral, económica y demográfica de cada sección.

- **t1** - número total de personas en la sección
- **t4** - porcentaje de población menor de 16 años
- **t4_2** - porcentaje de personas entre 16 y 64 años (ambos incluidos)
- **t4_3** - porcentaje de personas con más de 64 años
- **t5** - porcentaje de población extranjera
- **t9** - porcentaje de población con estudios superiores sobre población de 16 y más
- **t10** - porcentaje de población parada sobre población activa (tasa de paro)

- **t11** - porcentaje de población ocupada sobre población de 16 y más (tasa de empleo)
- **t12** - porcentaje de población activa sobre población de 16 y más (tasa de actividad)
- **t16** - porcentaje de población estudiante sobre población de 16 y más
- **t23** - tamaño medio del hogar
- **t24** - distancia en función del tipo de entorno (capital de provincia, municipio grande o municipio pequeño)
- **Renta** - Renta neta media por unidad de consumo

Con las variables finales ya ordenadas, el siguiente paso es realizar un análisis descriptivo, que permitirá explorar el comportamiento general de las variables.

Capítulo 3

Análisis exploratorio de los datos

Teniendo ya seleccionadas las variables relevantes para el análisis, se ha llevado a cabo un análisis descriptivo inicial con el objetivo de entender el comportamiento general de las variables. Es un paso fundamental, ya que hay muchos factores a tener en cuenta a la hora de realizar un estudio, como distribuciones muy asimétricas o valores extremos, que pueden condicionar los resultados posteriores.

Para apreciar de forma visual la forma de cada distribución, se incluyen los histogramas correspondientes en el Apéndice A.

3.1. Matriz de características y transformaciones

Como punto de partida, se ha construido una matriz de características que recoge una serie de medidas estadísticas básicas para cada variable: valor mínimo, máximo, media, mediana, desviación típica y asimetría. Con la siguiente tabla se dispone de una visión general del rango de cada variable y a su vez permite identificar aquellas variables que necesiten transformaciones debido a su distribución.

Los indicadores de asimetría indican si los valores de la distribución se disponen simétricamente alrededor de la media o no. Se puede interpretar de la siguiente manera:

- Asimetría > 0: distribución asimétrica positiva, los datos se decantan en mayor medida hacia la derecha.
- Asimetría $= 0$: distribución simétrica.
- Asimetría < 0: distribución asimétrica negativa, los datos se decantan en mayor medida hacia la izquierda.

En cuanto a la curtosis, mide si los datos están más o menos dispersos, por lo que se comparan con una distribución normal. Al coeficiente de curtosis se le resta 3 ya que, por definición, una distribución normal tiene curtosis igual a 3.

- Curtosis > 0: distribución más apuntada que la normal, datos menos dispersos (distribución leptocúrtica).

- Curtosis = 0: distribución igual que la normal (distribución mesocúrtica).
- Curtosis < 0: distribución menos apuntada que la normal, datos más dispersos (platicúrtica).

A continuación, se presenta la tabla resumen con las estadísticas mencionadas:

Variable	**Min**	**Max**	**Media**	**Mediana**	**SD**	**Asimetría**	**Curtosis**
t1	101	3191	830.23	767	604.98	0.67	-0.29
t4	0.00	31.01	10.03	9.73	5.18	0.78	0.97
t4_2	32.04	81.97	59.51	59.83	6.82	-0.35	0.05
t4_3	1.24	65.61	30.46	31.13	10.66	-0.10	-0.14
t5	0.00	55.92	5.60	4.20	5.03	2.10	8.22
t9	2.31	69.43	28.03	25.28	11.99	0.95	0.43
t10	0.00	66.51	14.68	13.89	5.86	1.39	6.06
t11	19.81	84.63	43.54	42.72	8.60	0.80	1.63
t12	25.33	89.80	50.92	50.30	8.67	0.69	1.31
t16	0.00	17.31	5.67	5.50	2.73	0.69	1.20
t23	1.45	3.19	2.27	2.26	0.24	0.26	0.72
t24	0.00	87160.37	15022.87	9176.21	16434.54	1.10	0.60
Renta	4687.00	25562.00	13384.78	12944.00	2460.20	1.24	2.68

Tabla 3.1: Estadísticos descriptivos de las variables

En términos generales, la tabla muestra que la mayor parte de las variables demográficas y económicas (*t1*, *t4*, *t9*, *t11*, *t12*, *t16*, *t23*) presentan asimetrías y exceso de curtosis moderadas (valores absolutos menores que 1), mientras que variables como *t5* (porcentaje de extranjeros), *t10* (tasa de paro) o *Renta* destacan por su elevada asimetría o curtosis con valores como 2,10, 1,39 y 1,24 de asimetría y 8,22, 6,06 y 2,68 de curtosis. Además, *t24* (distancia al centro según contexto) presenta la mayor desviación, cosa que ocurre debido a que la variable toma valores muy extremos (muy cercanos a 0 en núcleos urbanos y muy altos en zonas aisladas). Es importante observar las variables en las que la desviación típica supere a la media, ya que la dispersión de los datos es muy alta y, por tanto, la media puede no ser representativa.

La asimetría tan pronunciada se debe a la heterogeneidad territorial. Por un lado, para Renta hay determinadas secciones –probablemente en áreas urbanas– con ingresos muy elevados, mientras que zonas periféricas o rurales presentan rentas mucho más bajas. Algo muy similar ocurre con *t5*, la proporción de población extranjera, y *t10*, la tasa de paro, que varían drásticamente entre secciones, con algunas zonas casi libres de extranjeros o desempleados y otras con porcentajes muy altos.
Además de la asimetría, las distribuciones de *t5*, *t10* y Renta presentan colas muy pesadas (coeficientes de curtosis altos), lo que se traduce en una frecuencia elevada de valores extremos. Esto refuerza la necesidad de comprimir los extremos mediante transformaciones, ya que estas características generan una cola derecha muy alargada

y hacen que la media deje de ser un buen resumen para la distribución.

Con el objetivo de reducir el peso de los valores atípicos y uniformizar las escalas –dado que se están combinando variables con rangos muy dispares– se han aplicado:

1. **Transformación logit** a todas las proporciones (*t4*, *t4_2*, *t4_3*, *t5*, *t9*, *t10*, *t11*, *t12* y *t16*):

$$\text{logit}(p) \;=\; \ln\!\Big(\frac{p}{1-p}\Big)$$

   Para evitar los valores infinitos que genera la transformación logit en los casos $p = 0$ o $p = 1$, se introdujo un pequeño ajuste $\varepsilon = 10^{-3}$. Definimos

$$p^* = \begin{cases} \varepsilon, & \text{si } p = 0, \\ 1-\varepsilon, & \text{si } p = 1, \\ p, & \text{en cualquier otro caso.} \end{cases}$$

   De este modo, al aplicar la transformación

$$\text{logit}(p^*) \;=\; \ln\!\Big(\tfrac{p^*}{1-p^*}\Big)$$

   se evita la aparición de $-\infty$ o $+\infty$, garantizando que todos los valores queden en un rango finito y usable en el análisis.

2. **Transformación logarítmica** a *Renta*, *t1* y *t24*, aplicando:

$$X' \;=\; \ln(X)$$

   y, en el caso de *t24*, usando $\ln(X+1)$ para evitar casos en los que $X = 0$.

Una vez se tiene la transformación hecha, es conveniente verificar su efecto sobre la distribución de las variables.

Variable	**Min**	**Max**	**Media**	**Mediana**	**SD**	**Asimetría**	**Curtosis**
logit_t4	-5.33	-0.80	-2.35	-2.23	0.67	-0.87	1.24
logit_t42	-0.75	1.51	0.39	0.40	0.29	-0.21	0.07
logit_t43	-4.38	0.65	-0.90	-0.79	0.60	-1.16	2.67
logit_t5	-6.50	0.24	-3.19	-3.13	0.97	-0.25	-0.18
logit_t9	-3.74	0.82	-1.01	-1.08	0.60	0.40	0.13
logit_t10	-4.60	0.69	-1.84	-1.82	0.48	-0.28	2.10
logit_t11	-1.40	1.71	-0.27	-0.29	0.36	0.78	2.21
logit_t12	-1.08	2.18	0.04	0.01	0.37	0.96	2.79
logit_t16	-5.62	-1.56	-2.95	-2.84	0.60	-1.04	1.56
log_t1	4.62	8.07	6.38	6.64	0.92	-0.41	-1.12
log_t24	0.00	11.38	8.18	9.12	2.66	-1.62	2.58
log_Renta	8.45	10.15	9.49	9.47	0.17	0.45	1.50

Tabla 3.2: Estadísticos descriptivos de las variables transformadas

Se aprecia que, tras la transformación, *logit_t5* y *logit_t10* han reducido drásticamente sus valores absolutos de asimetría (cerca de cero). Para *log_Renta*, la asimetría pasa a ser moderada (0.28) y la curtosis baja a 1.50, lo que indica colas todavía un poco pesadas, pero mucho más manejables que en la escala original. Todas las variables muestran desviaciones típicas muy próximas a 1, asegurándose de que ninguna variable influya en la métrica por diferencias de escala. Estas mejoras confirman que las transformaciones han equilibrado las distribuciones.
Además, todas las variables han sido estandarizadas para situarlas en la misma escala. Esta preparación es esencial para el Análisis de Componentes Principales – que asume una matriz de covarianzas aproximadamente elíptica y cercana a la normal – y para el cálculo de distancias de Mahalanobis de forma coherente.

Como complemento visual, se incluyen en el Apéndice A los histogramas de las variables ya transformadas.

A continuación, para ver cómo se comportan las variables ya transformadas y estandarizadas, se presenta un diagrama de correlaciones de Pearson. El diagrama de correlaciones de Spearman se incluye en el Anexo, donde se observa que, aún sin asumir linealidad, las correlaciones son prácticamente iguales.

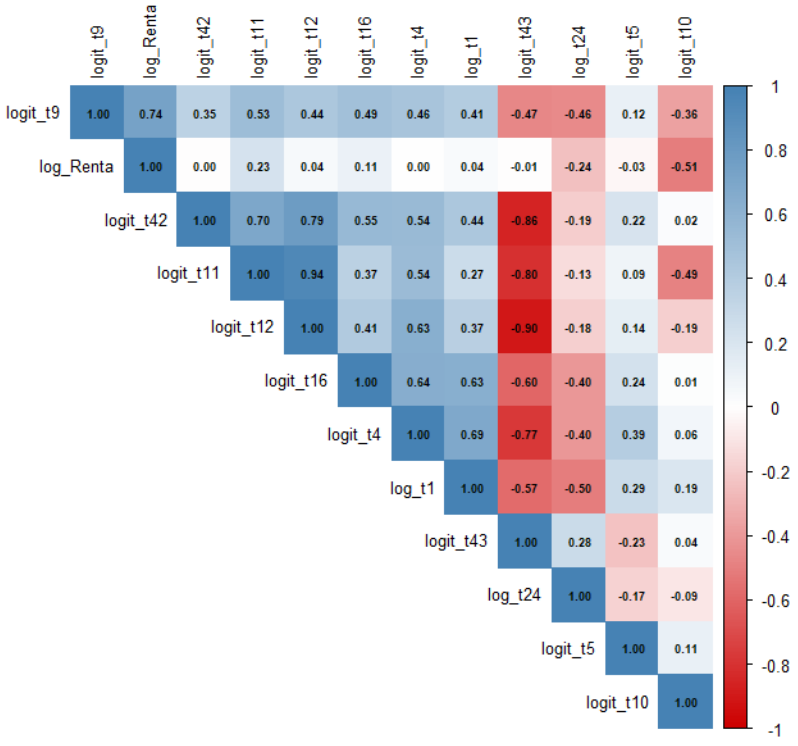

Figura 3.1: Diagrama de correlaciones de Pearson de las variables finales.

El gráfico resume las correlaciones de Pearson entre las 13 variables finales del análisis. Para sacar conclusiones, se pueden identificar cuatro bloques principales de relaciones.

1. **Educación e ingresos.** Se observa una fuerte correlación positiva entre la renta (*log_Renta*) y el nivel educativo (*logit_t9*) ($\approx +0{,}74$), lo que indica que las secciones con mayor proporción de población con estudios superiores tienden a registrar también mayores ingresos. Además, *logit_t9* y *logit_t10* muestran correlación negativa ($\approx -0{,}36$), de modo que a mayor nivel educativo, menor tasa de paro, y de forma complementaria, la renta y la tasa de paro (*log_Renta* vs. *logit_t10*) se correlacionan negativamente ($\approx -0{,}51$).

2. **Edad y empleo.** La proporción de población en edad activa (`logit_t42`) se asocia muy positivamente con la tasa de empleo (*logit_t11*) ($\approx +0{,}94$) y con la tasa de actividad (*logit_t12*) ($\approx +0{,}80$). Las secciones con más población menor de 16 años (*logit_t4*) presentan también un mayor porcentaje de estudiantes (*logit_t16*), con correlación $\approx +0{,}63$.Por otro lado, la proporción de mayores de 64 años (*logit_t43*) se asocia negativamente con empleo y actividad ($\approx -0{,}77$ y $\approx -0{,}90$, respectivamente).

3. **Población y tamaño de los hogares.** La población total (*log_t1*) se relaciona positivamente con el tamaño medio del hogar (*logit_t23*) ($\approx +0{,}57$). No obstante, dado que en las zonas urbanas las secciones censales típicamente oscilan entre 500 y 2 500 habitantes, la variación de *log_t1* es reducida. Por el contrario, son las secciones rurales de menos de 500 habitantes las que explican la correlación con *log_t23*, ya que en estos entornos los hogares suelen ser más pequeños. De forma análoga, *log_t1* presenta asociaciones positivas con *logit_t42* ($\approx +0{,}69$) y *logit_t12* ($\approx +0{,}48$), subrayando que las secciones de menor población –típicamente rurales– presentan menor proporción de activos y menor actividad económica.

4. **Presencia de población extranjera.** La proporción de extranjeros (*logit_t5*) se asocia negativamente con el envejecimiento (*logit_t43*) ($\approx -0{,}46$), reflejando una menor presencia de inmigración en secciones con más población mayor. Su correlación con la renta (*log_Renta*) es prácticamente nula ($\approx -0{,}03$), lo que sugiere que la renta media no es un factor determinante de la distribución de población extranjera.

5. **Ruralidad (distancia al "centro").** La variable de aislamiento (*log_t24*) muestra un vínculo negativo con la renta ($\approx -0{,}18$) y la educación ($\approx -0{,}74$), indicando que a mayor aislamiento, disminuyen ingresos y nivel formativo. También está correlacionada positivamente con el envejecimiento (*logit_t43*$\approx +0{,}36$), lo cual confirma un perfil demográfico más mayor en entornos rurales. Además, se asocia negativamente con empleo y actividad ($\approx -0{,}44$ y $\approx -0{,}43$), destacando la menor intensidad económica de las áreas más alejadas.

3.2. Análisis de Componentes Principales

Con las correlaciones examinadas y las variables adecuadamente transformadas y estandarizadas, se procede al Análisis de Componentes Principales, cuyo objetivo es:

- Confirmar que la nube de puntos en el espacio de componentes principales adopta una forma aproximadamente elíptica, requisito clave para la validez de la distancia de Mahalanobis.
- Determinar el porcentaje de varianza explicada por las primeras componentes, para comprobar si la mayor parte de la variabilidad multivariante se concentra en unos pocos ejes.

A continuación se presenta el ACP aplicado a las 13 variables finales.

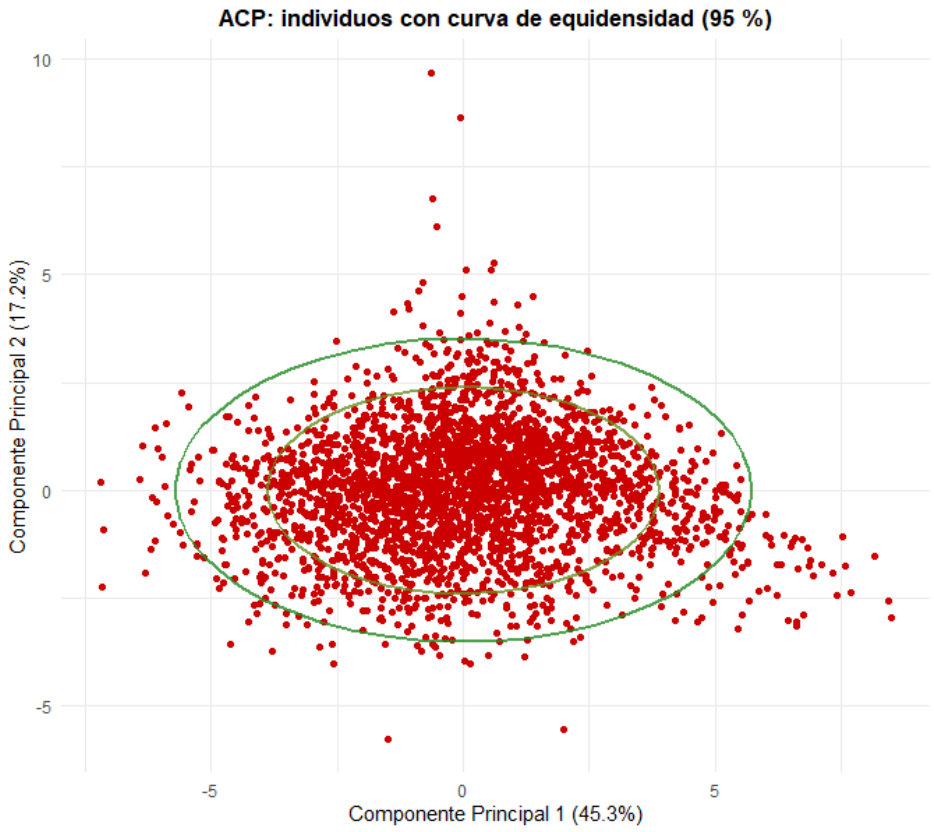

Figura 3.2: Diagrama de individuos en 2 dimensiones con curva de equidensidad al 95 % y al 75 %.

El diagrama muestra que los puntos, donde cada uno es una sección censal, se distribuyen de forma concentrada en un núcleo aproximadamente elíptico, sin grandes agrupamientos aislados. El nivel de equidensidad al 95 % (trazado como curva continua en verde más oscuro) ajusta muy bien ese núcleo, lo que confirma gráficamente la hipótesis de dispersión elipsoidal.
Se observa que la Dimensión 1 (45,3 %) y la Dimensión 2 (17,2 %) capturan conjuntamente el 62,5 % de la variabilidad total. Este elevado porcentaje, junto con la forma casi elíptica de la nube y sus curvas de equidensidad, respalda la aplicación de la distancia de Mahalanobis (robusta) como métrica de similitud multivariante.

Capítulo 4

Metodología. MCD

Para garantizar que el cálculo de similitud entre secciones censales no se vea afectado por valores atípicos, se va a utilizar un estimador robusto de la media y la matriz de covarianzas conocido como **MCD** (Minimum Covariance Determinant), propuesto por Peter J. Rousseeuw en 1984 (Rousseeuw, 1984).

MCD es un método utilizado para estimar el vector de medias μ y la matriz de covarianzas Σ de tal forma que se minimice la influencia de posibles valores atípicos. Es un método muy interesante ya que es el estimador máximo verosímil recortado utilizado para eliminar una cantidad fija de datos.
La idea consiste en calcular estas estimaciones a partir de un subconjunto de datos cuya dispersión sea más reducida, con el fin de excluir aquellos valores atípicos que se encuentren más alejados del resto de la muestra.

Para ello, MCD utiliza como criterio de selección el determinante de la matriz de covarianzas. Lo que busca es minimizar el determinante $\det(\Sigma)$ de la matriz de covarianzas, ya que dicho determinante es proporcional al volumen del elipsoide que describe la dispersión multivariante.
El MCD alcanza un *breakdown point* de hasta el 50 %, lo que quiere decir que puede tolerar que hasta la mitad de observaciones sean outliers sin que las estimaciones de μ y Σ se vean afectadas, lo que hace que esté entre los estimadores más robustos.

Para elegir el porcentaje de datos a seleccionar, se puede fijar un α, que determina el subconjunto robusto. Con un $\alpha = 0.50$ se garantiza máxima robustez, pero a su vez se descarta un número alto de observaciones, lo que implica menor eficiencia estadística y pérdida de representatividad de las características de los datos. Por tanto, es típico escoger valores de α de aproximadamente el 75 % de los datos.
Si no se especifica, se emplea por defecto:

$$\alpha \;=\; \frac{n+p+1}{2\,n},$$

donde n es el número de observaciones –secciones censales– y p el número de variables, garantizando así que el subconjunto seleccionado sea lo suficientemente grande para

que la matriz de covarianzas resultante sea invertible.

Para implementar el MCD de forma eficiente se utiliza el algoritmo FAST–MCD propuesto por Rousseeuw y Van Driessen, 1999, que consiste en generar múltiples submuestras "semilla" de tamaño $p+1$ y, para cada una, se calcula la media y la covarianza clásicas (T_0, S_0). Después, se ejecuta un proceso iterativo de *C–pasos*:

1. Para cada estimador (T_k, S_k), se computan las distancias de Mahalanobis de las n observaciones:
$$D_i^2 = (x_i - T_k)^\top S_k^{-1} (x_i - T_k).$$
2. Se ordenan las distancias D_i^2 en orden creciente y se seleccionan las $n(1-\alpha)$ observaciones con valores más pequeños, conformando el nuevo subconjunto A_{k+1}.
3. Sobre A_{k+1} se recalculan la media y la covarianza clásicas:
$$T_{k+1} = \frac{1}{n*(1-\alpha)} \sum_{i \in A_{k+1}} x_i, \quad S_{k+1} = \frac{1}{n*(1-\alpha)} \sum_{i \in A_{k+1}} (x_i - T_{k+1})(x_i - T_{k+1})^\top.$$
4. Se repiten estos pasos hasta que el determinante $\det(S_{k+1})$ deje de disminuir de manera significativa, en cuyo punto se asume que el algoritmo ha llegado a una solución estable.

De todas las iteraciones se escoge la que consiga el menor valor final de $\det(S)$, obteniéndose así las estimaciones robustas $\mu_{\text{MCD}} = T^*$ y $\Sigma_{\text{MCD}} = S^*$.

En R, este algoritmo se implementa en la función `covMcd()` del paquete `robustbase`. Al no especificar α, la función mencionada utiliza el α especificado por Rousseeuw y Van Driessen. A partir del objeto devuelto por `covMcd()` se extrae el vector de medias robustas mediante su componente `$center` y la matriz de covarianzas robusta con `$cov`. Estas estimaciones se utilizan posteriormente en la fórmula de la distancia de Mahalanobis para calcular, de manera insensible a outliers, la similitud entre cada sección censal y la sección de referencia.

Comparación con la matriz clásica

Para evaluar de manera objetiva cómo mejora la robustez con MCD y si realmente es necesario, se ha comparado su matriz de covarianzas Σ_{MCD} con la covarianza clásica Σ_{cl} calculada sobre todas las observaciones. Se va a evaluar de la siguiente manera:

1. **Norma de Frobenius:**
$$\|\Sigma_{\text{cl}} - \Sigma_{\text{MCD}}\|_F = \sqrt{\sum_{i,j} \left(\Sigma_{\text{cl,ij}} - \Sigma_{\text{MCD,ij}}\right)^2},$$
resume en un único valor cuánto difieren ambas matrices en todos sus elementos.

2. **Comparación de direcciones:** se representa en un gráfico los autovalores de ambas matrices, para apreciar cómo cambia la varianza explicada en cada dirección principal.

3. **Comparación de distancias Mahalanobis:** se representa en un diagrama de dispersión cada par de distancias $\{d_{\text{cl}}(x_i),\, d_{\text{MCD}}(x_i)\}$ para todas las secciones censales.

Estos análisis permiten cuantificar y visualizar en qué medida la versión realizada con MCD corrige la influencia de valores atípicos.

Comparación de matrices de covarianzas: $\alpha = 0{,}501$ **y** $\alpha = 0{,}75$

Si además se quiere observar el impacto del parámetro α en la estimación del MCD, hace falta comparar Σ_{cl} con las dos versiones MCD obtenidas para $\alpha = \text{default} \approx 0{,}501$ y $\alpha = 0{,}75$. Aplicando la norma de Frobenius en R, se obtiene:

$$D_F(\text{default}) \;=\; 3.1188, \qquad D_F(0,75) \;=\; 2.4057.$$

- Con $\alpha \approx 0{,}501$, la distancia que se obtiene es $D_F \approx 3{,}12$, ya que descarta la mitad de las observaciones y elimina gran parte de la influencia de outliers.

- Con $\alpha = 0{,}75$ (25 % de los datos), la distancia es menor ($D_F \approx 2{,}41$), lo que indica que sigue habiendo robustez frente a valores extremos, pero conservando mejor la estructura original.

Ya que $\alpha = 0{,}75$ reduce notablemente el efecto de los outliers, pero sin alterar en exceso la relación entre las variables, se va a tomar como valor óptimo para la estimación MCD en este trabajo.

Comparación de direcciones

En la siguiente figura se muestran dos biplots con el círculo unitario: a la izquierda, las cargas de las variables sobre las dos primeras componentes principales calculadas con la covarianza clásica (flechas en azul); a la derecha, las correspondientes a la matriz robusta MCD ($\alpha = 0{,}75$) (flechas en rojo).

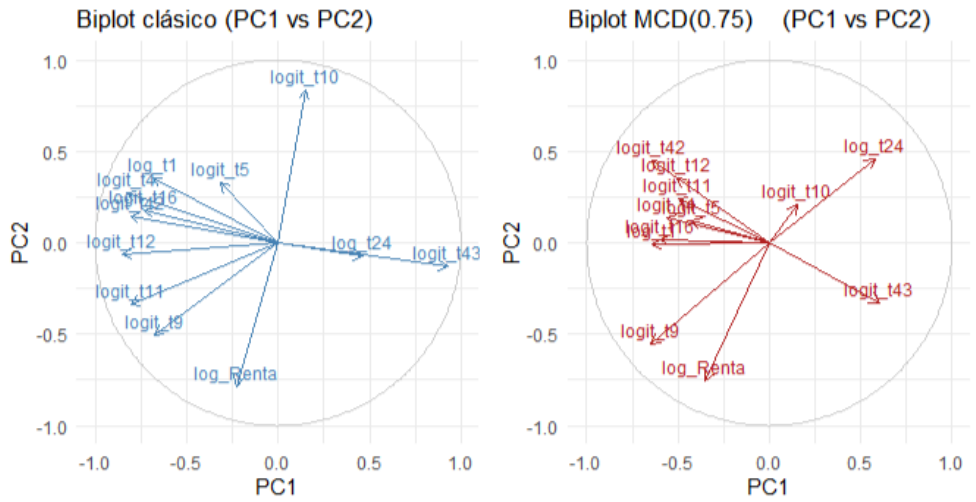

Figura 4.1: Biplots de las dos primeras componentes principales: Covarianza clásica vs. Covarianza robusta MCD (0.75)

- La orientación de la mayoría de las flechas apenas varía entre los dos biplots, lo que indica que la eliminación de outliers no altera de forma significativa las variables que definen los ejes principales.
- Las longitudes cambian ligeramente, algunas flechas rojas son más cortas, mostrando que MCD (0.75) ha reducido la varianza explicada en esas variables al disminuir el efecto de outliers.
- En conjunto, la forma global del "círculo de correlaciones" se conserva, demostrando así que MCD (0.75) protege frente a valores atípicos sin distorsionar la estructura multivariante subyacente.

Comparación de distancias de Mahalanobis

Para completar la evaluación del MCD, se comparan las distancias de Mahalanobis clásicas y las robustas para cada sección censal. En ambos gráficos, el eje horizontal representa la distancia clásica $d_{\text{cl}}(x_i)$, y el vertical la distancia robusta $d_{\text{MCD}}(x_i)$.

En la Figura 4.2 se observa que las distancias clásicas raramente superan los 400, mientras que las robustas alcanzan valores de hasta 4.000. Esto ocurre porque al excluir el 25 % más extremo de las observaciones al calcular Σ_{MCD}, la dispersión de los datos se reduce y, por tanto, incluso pequeñas diferencias con el centro generan distancias más altas.

La Figura 4.3 muestra únicamente las secciones con $d_{\text{cl}}(x_i) < 10$, es decir, el "núcleo" de la distribución. Se puede apreciar que:

- La gran mayoría de puntos quedan por encima de la diagonal $y = x$, lo que indica que $d_{\text{MCD}}(x_i) \geq d_{\text{cl}}(x_i)$. Al eliminar outliers, la dispersión disminuye, y hasta pequeñas diferencias generan distancias mayores.

- Las distancias más pequeñas –aquellas que realmente importan para encontrar sustituciones similares– son prácticamente idénticas bajo la métrica clásica y la robusta, confirmando que MCD(0.75) no altera la relación de similitud en el núcleo de las secciones.

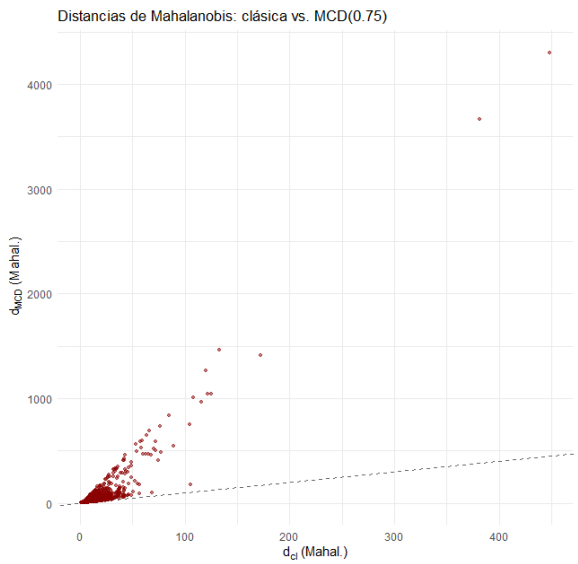

Figura 4.2: Comparación distancias de Mahalanobis

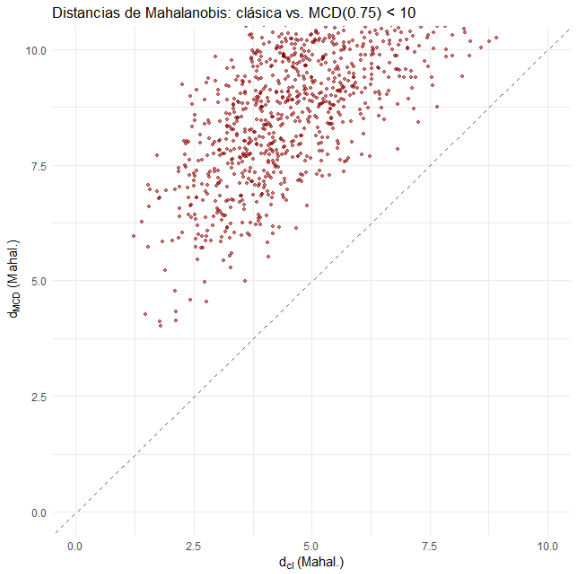

Figura 4.3: Comparación distancias de Mahalanobis < 10

Además, los resultados numéricos –norma de Frobenius– y visuales –biplots– coinciden en que la estimación MCD con $\alpha = 0{,}75$:

1. Mantiene las principales direcciones de las variables del conjunto de datos.

2. Consigue un equilibrio entre la resistencia a valores atípicos y la preservación de la estructura de covarianzas original.

Por tanto, el uso de la matriz Σ_{MCD} obtenida con `covMcd(..., alpha = 0.75)` está justificado en el cálculo de distancias de Mahalanobis robustas, ya que mantiene las direcciones principales y la forma elipsoidal de los datos, a la vez que reduce la influencia de valores atípicos.

Aunque para el objetivo concreto de esta herramienta –identificar la sección más parecida a una dada– el MCD no aporta grandes diferencias frente a la covarianza clásica, es recomendable incluirlo siempre en el proceso. De este modo se pueden prevenir posibles distorsiones debidas a outliers en futuros escenarios o conjuntos de datos distintos y así se puede garantizar la mayor robustez del método de selección de secciones censales.

4.1. Distancias de Mahalanobis

Se utiliza la distancia de Mahalanobis como métrica de comparación entre secciones censales, ya que, a diferencia de la distancia euclídea, esta medida incorpora la información de la matriz de covarianzas de las variables transformadas y estandarizadas, penalizando en mayor proporción los indicadores que están muy relacionados debido a que aportan información similar. De este modo, la distancia de Mahalanobis permite identificar secciones "más parecidas" considerando las correlaciones de los indicadores, garantizando una sustitución coherente en el diseño muestral.

La distancia de Mahalanobis entre un punto x_i y el centro de la distribución μ, que corresponde al vector de la sección de referencia x_{ref}, se define para cada sección i de la siguiente forma:

$$d_{\text{MCD}}(x_i,\, x_{\text{ref}}) \;=\; \sqrt{(x_i - x_{\text{ref}})^{\top}\, \Sigma_{\text{MCD}}^{-1}\, (x_i - x_{\text{ref}})},$$

Estas distancias permiten ordenar las secciones censales según su grado de similitud con una sección de referencia, garantizando que la comparación sea resistente a valores atípicos.

Capítulo 5

Aplicación Shiny

Con el objetivo de poner en práctica el método desarrollado, se ha creado una aplicación interactiva en Shiny que, seleccionando cualquier sección censal de Castilla y León, calcula su sección más parecida en función de las variables mencionadas anteriormente.

Para ello, se ha utilizado la librería Shiny, desarrollada por Rstudio, que permite construir aplicaciones web interactivas a partir de código R. Más concretamente, se ha utilizado un `shinydashboard` para proporcionar una interfaz con un diseño más atractivo para el usuario, basado en una cabecera, barra lateral y pestañas, contando también con la ayuda del paquete `shinycssloaders` para incorporar indicadores de carga visuales.
Gracias al sistema de reactividad de Shiny, cualquier cambio en los controles de entrada ("inputs") produce la recalculación de las salidas ("outputs") automáticamente. Así, al seleccionar otro municipio o sección, el mapa se actualiza al instante sin tener que recargar toda la página.

5.1. Estructura general

Las aplicaciones Shiny consisten en dos partes principales:

- **Interfaz de usuario (UI)**: define los elementos visuales y de entrada con los que interacciona el usuario. Controla el diseño final y la apariencia de la aplicación Shiny.

- **Lógica del servidor (Server)**: contiene la parte reactiva y los procesos que actualizan y analizan la información según las interacciones del usuario.

La App puede estructurarse en un único archivo `app.R` que combine ambas partes, o bien en dos archivos separados: `ui.R` y `server.R`. A continuación, se describen ambos componentes.

5.1.1. Interfaz de usuario (UI)

La función `ui` construye la estructura visual de la aplicación con los siguientes elementos:

- `fluidPage()` o `dashboardPage()`: esquemas de diseño que adaptan automáticamente la disposición de los elementos al tamaño de la ventana. Con `fluidPage()` se tiene un diseño básico y sencillo, donde los elementos se distribuyen automáticamente de arriba a abajo y se ajustan al ancho. En este trabajo se ha optado por `dashboardPage()`, ya que permite un acceso rápido a los controles en el menú lateral y la presentación simultánea de mapas, indicadores y resultados con un diseño más cuidado y estructurado.
- `inputControls`: se utilizan `menuItem()` y `menuSubItem()` para organizar la estructura de las pestañas del menú lateral, y `selectInput()` para elegir la provincia, el municipio o la sección censal de partida.
- `absolutePanel()`: panel flotante que posiciona elementos en coordenadas fijas sobre el mapa y los hace arrastrables.
- `outputPlaceholders`: los espacios reservados, definidos en la ui, para los resultados que genera después el servidor. El cuerpo principal (`dashboardBody`) se organiza mediante `tabItems()`, cada `tabItem()` incluye elementos como `leafletOutput()` para mapas interactivos o `plotOutput()` para gráficos, a los que se les puede añadir una animación con `withSpinner()` mientras se procesan los datos.
- `dashboardPage()`: esquema de la aplicación que integra un encabezado (`dashboardHeader`), un panel lateral (`dashboardSidebar`) con los controles, y un cuerpo principal (`dashboardBody`) con los resultados, ofreciendo un diseño limpio y profesional.

En resumen, la parte de interfaz de usuario actúa como el "panel de control"de la aplicación, donde el usuario puede elegir la provincia, el municipio y la sección censal. Pulsa un botón para seleccionar lo buscado y, a continuación, visualiza de forma clara el mapa con ambas secciones –la original y la más parecida– junto a determinados indicadores numéricos que resumen sus características. Esto se debe a que las salidas se definen en el Server y en la UI se configura su visualización.

5.1.2. Lógica del servidor (Server)

La función `server` define la parte reactiva de la aplicación, es decir, los procesos que se ejecutan cada vez que el usuario modifica una de las entradas. Sus elementos clave son:

- `renderUI()`: genera dinámicamente los controles de selección de municipio y sección una vez que el usuario elige la provincia y el municipio, respectivamente.

- `observeEvent()`: detecta cuándo el usuario elige una sección (ya sea en la búsqueda general o en las pestañas de capitales) y, en ese momento, llama a la función que calcula cuál es la sección más parecida. A continuación, le indica a Shiny que actualice el mapa.
- `renderLeaflet()` y `leafletProxy()`: crean y modifican los mapas interactivos, dibujando las fronteras de las secciones y resaltando la sección seleccionada y la más similar.

En la aplicación desarrollada en concreto:

1. `renderUI()` construye los `selectInput()` para municipios y secciones censales en función de la provincia seleccionada.
2. `observeEvent(input$seccion_ref)` llama a `buscar_mas_parecida()`[1], obtiene la sección más parecida y ejecuta `renderLeaflet()` para dibujar el mapa de búsqueda.
3. Para cada capital de provincia, otro `observeEvent(input$cusec_<capital>)` actualiza el mapa correspondiente mediante `leafletProxy()`, resaltando la sección origen y la sección hallada.

En conjunto, el funcionamiento interno de Shiny se basa en un sistema reactivo que enlaza directamente los controles de la interfaz de usuario con los procesos definidos en el servidor.

5.2. Manual de uso de la aplicación creada

A continuación, se presenta la herramienta desarrollada, *SecCyL*, detallando su estructura y funcionalidades principales. La pantalla de inicio actúa como página de bienvenida y referencia, con los siguientes elementos principales:

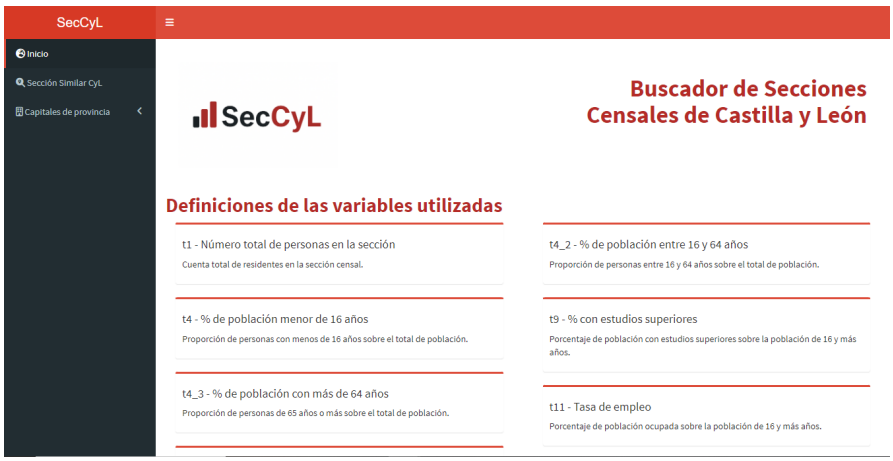

Figura 5.1: Pantalla de inicio de la aplicación.

[1]Función creada en R para buscar la sección más parecida

- **Panel izquierdo de navegación:** desde este menú lateral se puede acceder en cualquier momento a las tres secciones principales de la aplicación:
  - "Sección Similar CyL"
  - "Capitales de provincia"
  - "Indicadores"
- **Definiciones de variables:** al abrir la aplicación se despliega un listado de las variables utilizadas para el cálculo de similitud, cada una con su código (por ejemplo, *t1*, *t23*, ...) y su descripción breve. Esto facilita que el personal del INE entienda al instante cómo se ha buscado la sección más parecida y qué representa cada indicador antes de iniciar la búsqueda de secciones.

De este modo, la pestaña "Inicio" sirve como referencia rápida y orienta al usuario sobre las herramientas disponibles en el menú lateral antes de pasar al proceso de búsqueda y visualización de resultados.

Sección Similar CyL

Al seleccionar la pestaña **"Sección Similar CyL"** aparece la siguiente pantalla:

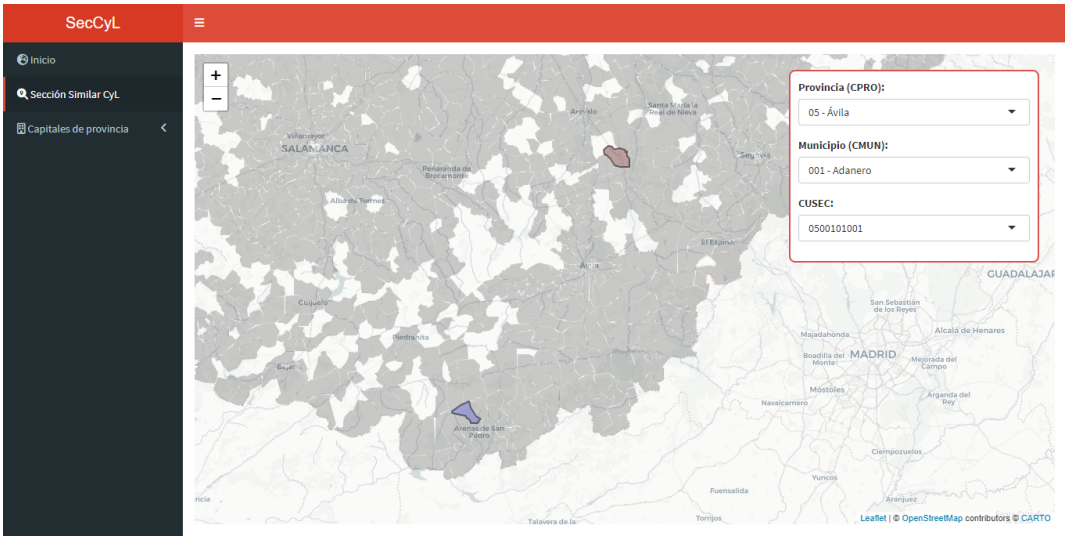

Figura 5.2: Sección Similar CyL. Caso 1.

En la parte derecha, el usuario dispone de un panel flotante con tres desplegables encadenados:

1. **Provincia**: lista de las nueve provincias de Castilla y León con sus códigos.
2. **Municipio**: una vez elegida la provincia, se cargan automáticamente los municipios correspondientes.

3. **Sección censal**: tras elegir el municipio, aparece el listado de secciones (CUSEC) disponibles.

Al cambiar cualquiera de estas selecciones, la aplicación recalcula al instante la sección más parecida usando la distancia de Mahalanobis. El mapa central muestra:

- En **rojo oscuro**, la sección seleccionada.
- En **azul**, la sección más parecida.

Además, al pasar el cursor sobre cada polígono, aparece un pequeño panel emergente con información relevante (renta media/ud de consumo, tasa de paro, porcentaje de extranjeros y tamaño medio del hogar).

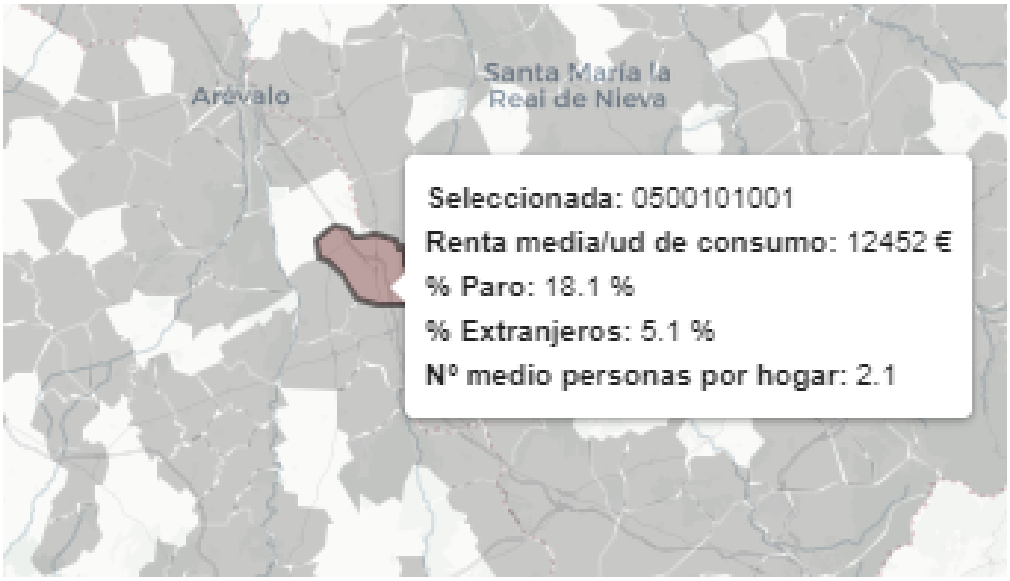

Figura 5.3: Panel emergente informativo.

En esta pestaña la búsqueda se adapta automáticamente a la situación del municipio seleccionado. En el primer caso, cuando el municipio seleccionado solo cuenta con una sección, la aplicación amplía automáticamente el ámbito de búsqueda a toda la provincia y muestra una notificación en la esquina inferior derecha indicando dicho cambio. En el segundo caso, si el municipio cuenta con varias secciones censales, la búsqueda de la sección más parecida se realiza dentro del municipio seleccionado.

En la Figura 5.2 se puede observar el primer caso. Aquí, el CUSEC `0500101001`, seleccionado por defecto al abrir la pestaña, corresponde al municipio de Adanero (Ávila), que contaba con 196 habitantes a fecha 1 de enero de 2021 y una única sección censal (caso de "sección rural"). En este caso, la función `buscar_mas_parecida()` busca la sección en toda la provincia.

En cambio, si seleccionamos el CUSEC `4717501001`, correspondiente a Tudela de Duero (Valladolid), que cuenta con 6 secciones censales, la función `buscar_mas_parecida()` busca la sección en el mismo municipio.

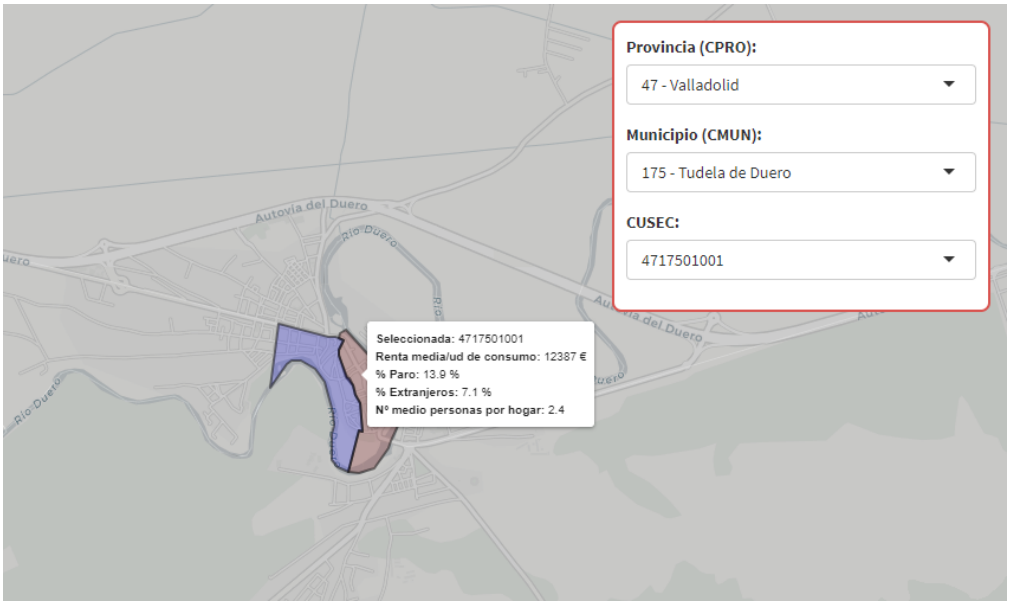

Figura 5.4: Sección Similar CyL. Caso 2.

Resulta importante destacar que las zonas en blanco del mapa no aparecen coloreadas ni interactivas ya que corresponden a las secciones con menos de 100 habitantes que se han excluido del análisis (por secreto estadístico y baja probabilidad de selección).

Capitales de Provincia

La sección **"Capitales de Provincia"** permite analizar cada una de las nueve capitales de Castilla y León de forma individual. Al desplegar este menú, el usuario encuentra un subitem para cada capital (Ávila, Burgos, León, Palencia, Salamanca, Segovia, Soria, Valladolid, Zamora)

Al elegir una capital, se carga una nueva página con un desplegable con las secciones censales de esa capital. Una vez seleccionada, el mapa muestra la sección de referencia, y automáticamente resalta su homóloga más parecida en el mismo estilo que la búsqueda general.

Estos apartados simplifican la búsqueda cuando el usuario necesita trabajar específicamente con las capitales de provincia, sin tener que filtrar manualmente municipios y secciones en el buscador general. Es importante tenerlos ya que las delegaciones suelen necesitar encontrar la sección más parecida dentro de la propia capital.

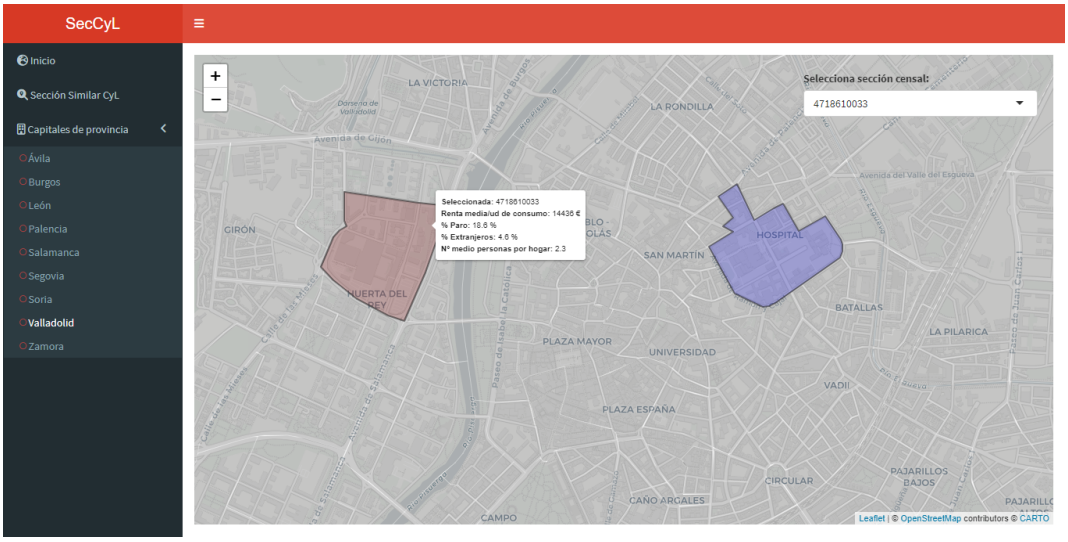

Figura 5.5: Sección similar capital de provincia.

En la Figura 5.5 se muestra la pantalla correspondiente al análisis de la capital de Valladolid, seleccionando el CUSEC `4718610033`, sección a la que pertenece la Delegación Provincial del Instituto Nacional de Estadística. Al seleccionar "Valladolid" en el menú lateral, aparece un desplegable con las secciones censales de la capital y, al elegir una de ellas, el mapa resalta en color **rojo oscuro**, la sección seleccionada y en **azul**, la sección más parecida.

Los desplegables que aparecen al pasar el cursor por encima son los siguientes:

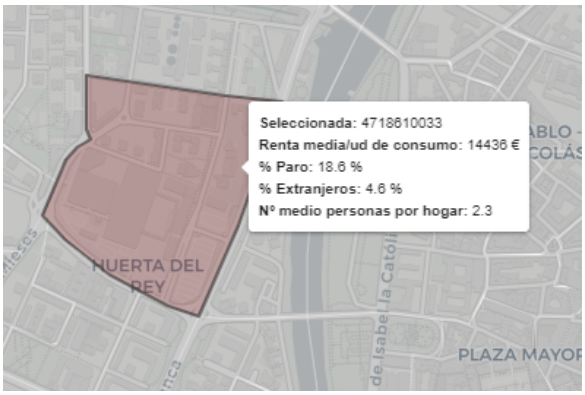

Figura 5.6: Sección seleccionada

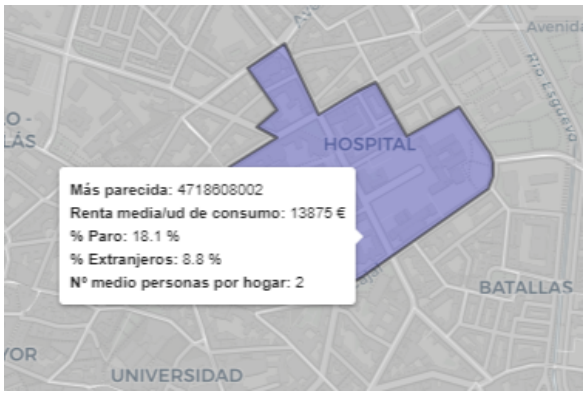

Figura 5.7: Sección más parecida

Esta opción se encuentra disponible para todas las capitales de provincia de Castilla y León.

Para facilitar el acceso a esta herramienta, se ha publicado en la plataforma de ShinyApps, donde puede consultarse directamente desde cualquier navegador y dispositivo:

https://estrellasantos.shinyapps.io/SecCyL/

Capítulo 6

Conclusión y trabajo futuro

En este Trabajo de Fin de Grado se ha presentado SecCyl, una aplicación interactiva que permite identificar la sección censal más parecida a una sección de referencia en Castilla y León mediante una métrica multivariante robusta. Se ha validado el uso de distancias de Mahalanobis –gracias al Análisis de Componentes Principales–, se ha implementado un estimador robusto de la matriz de covarianzas y se ha desarrollado una interfaz Shiny para obtener una herramienta intuitiva para el personal de las delegaciones provinciales del INE de Castilla y León.
En conjunto, la aplicación desarrollada, SecCyL, proporciona una solución práctica y fiable para la sustitución de secciones censales, reforzada por métodos estadísticos robustos.

A partir de los resultados obtenidos, se plantean algunas ideas para el desarrollo futuro:

- **Expansión territorial.** Adaptar la aplicación para cubrir la totalidad de España, cargando de forma dinámica los datos censales de cada provincia y permitiendo su uso en todas las delegaciones provinciales del INE.
- **Nuevas funcionalidades.** Incorporar filtros avanzados (p. ej. por rango de población o indicadores concretos) y permitir al usuario seleccionar las variables con las que quiera que se busque la sección similar. Se podría añadir también la opción de que aparezcan, no solo la sección más parecida, sino las 2 o 3 secciones más parecidas.
- **Actualización automática.** Conectar la aplicación a ficheros abiertos del INE para actualizar periódicamente los indicadores (por ejemplo, el Censo Anual de Población) sin necesidad de añadirlos manualmente.
- **Mejora del uso de la aplicación.** Mejorar la experiencia de usuario con tutoriales integrados, opciones de exportación de mapas y resultados, y accesibilidad para dispositivos móviles.
- **Incorporación de secciones con menos de 100 habitantes.** Actualmente se excluyen para cumplir requisitos de secreto estadístico, pero sería interesante

explorar estrategias de imputación que permitan incluir también estas unidades de pequeño tamaño.

- **Gestión de datos faltantes mediante CellMCD.** Cuando existan valores perdidos en las variables censales, se podría aplicar el método *CellMCD* de Rousseeuw & Raymaekers, que ofrece una estimación robusta de media y covarianza a nivel de celdas, permitiendo conservar observaciones parciales sin necesidad de eliminarlas por completo.

Apéndice A

Histogramas de variables

Aquí se presentan los histogramas de las variables iniciales utilizadas en el análisis descriptivo.

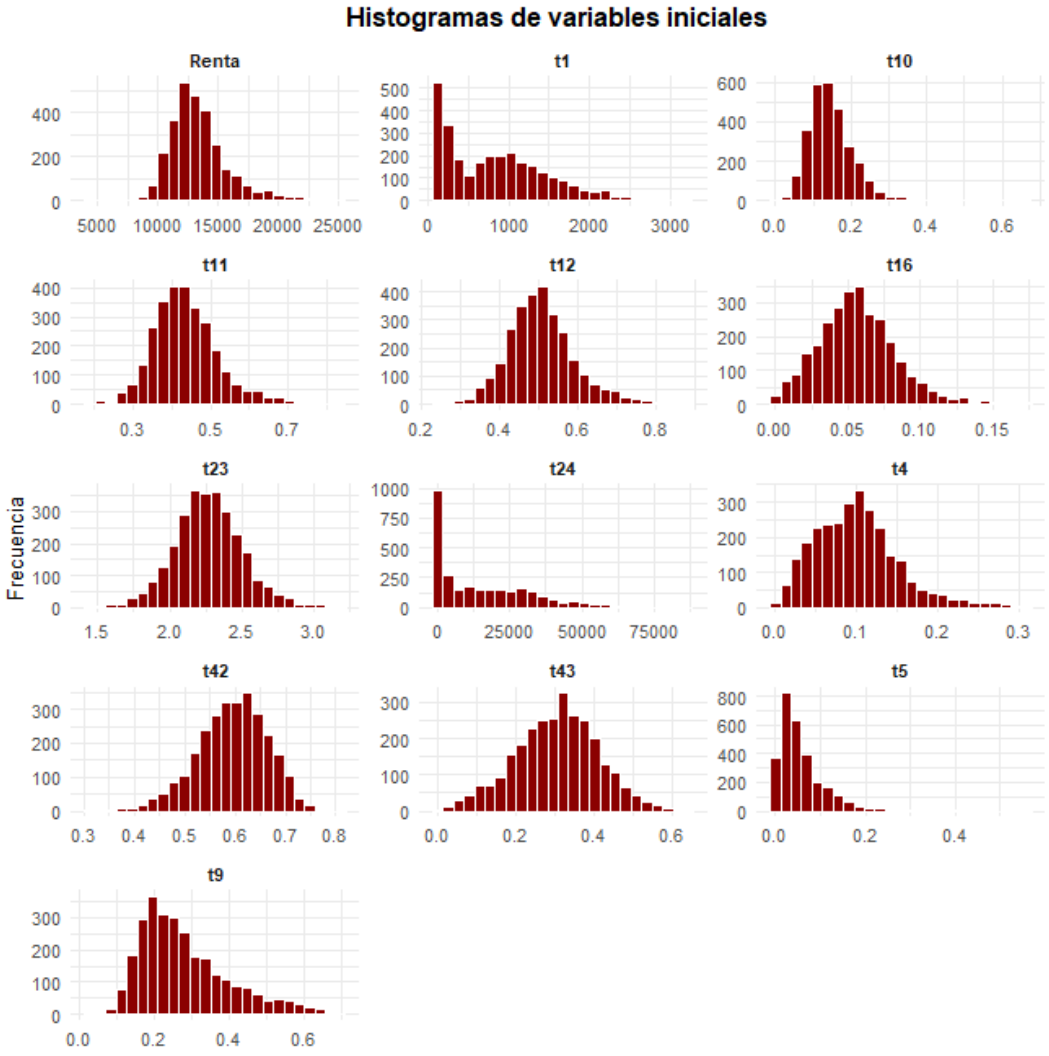

Figura A.1: Histogramas de las variables iniciales.

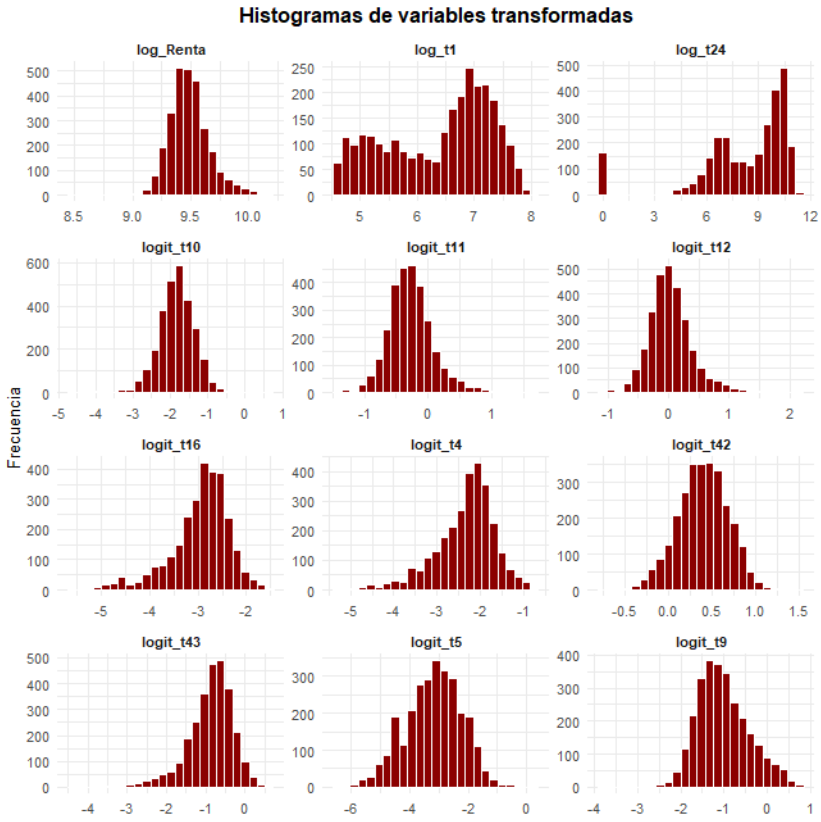

Figura A.2: Histogramas de las variables transformadas.

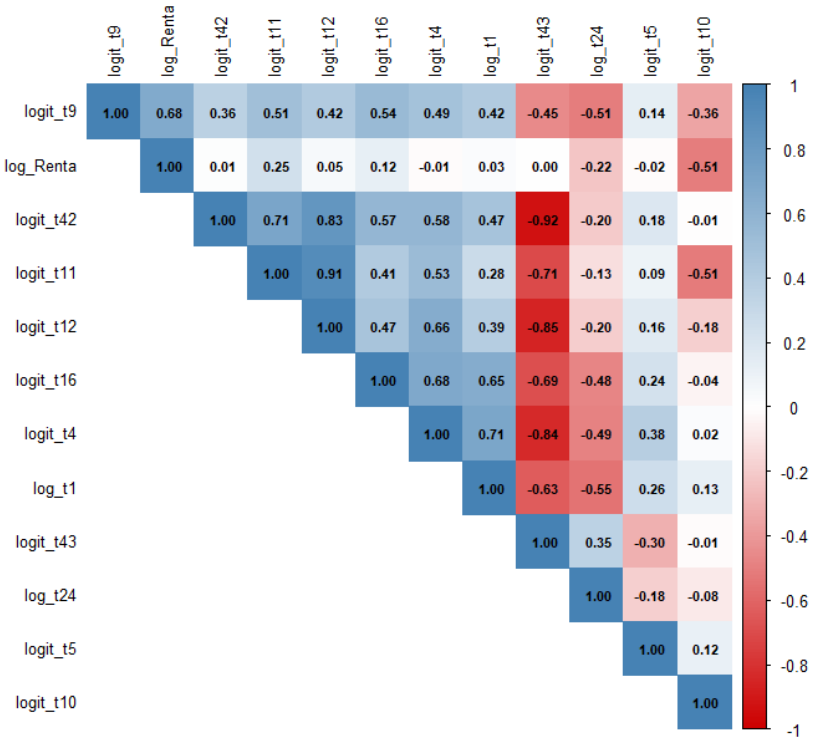

Figura A.3: Diagrama de correlaciones de Spearman de las variables finales.

Apéndice B

Código R

El código completo de la aplicación SecCyL, junto con los datos utilizados, está disponible en GitHub:

https://github.com/estreesantos/SecCyL-TFG

Bibliografía

Amor Pulido, R., Aguilar Peña, C., & Morales Luque, A. (2014). *Estadística Aplicada* (2ª). Grupo Editorial Universitario.

Boletín Oficial del Estado. (1985). *Ley Orgánica 5/1985, de 19 de junio, del Régimen Electoral General, art. 23* [Accedido: 14-05-2025]. https://www.boe.es/buscar/doc.php?id=BOE-A-1985-11672

Boletín Oficial del Estado. (1989, mayo). *Ley 12/1989, de 9 de mayo, de la Función Estadística Pública* [Accedido: 11-06-2025]. https://www.boe.es/buscar/doc.php?id=BOE-A-1989-10767

Boletín Oficial del Estado. (1997, julio). *Orden de 11 de julio de 1997 sobre comunicaciones electrónicas entre las Administraciones públicas referentes a la información de los Padrones municipales* [Accedido: 26-05-2025]. https://www.boe.es/buscar/doc.php?id=BOE-A-1997-15814

Boletín Oficial del Estado. (2020a, febrero). *Resolución de 17 de febrero de 2020, de la Presidencia del Instituto Nacional de Estadística y de la Dirección General de Cooperación Autonómica y Local, por la que se dictan instrucciones técnicas a los Ayuntamientos sobre la gestión del Padrón municipal* [Accedido: 26-05-2025]. https://www.boe.es/buscar/act.php?id=BOE-A-2020-4784

Boletín Oficial del Estado. (2020b, diciembre). *Real Decreto 1110/2020, de 15 de diciembre, por el que se regula el Censo de Población y Viviendas y el Censo de Población continuo* [Accedido: 18-06-2025]. https://www.boe.es/eli/es/rd/2020/12/15/1110

CDR Book. (2025). *Shiny* [Accedido: 30-04-2025]. https://cdr-book.github.io/shiny.html

de la Plaza, A. G. (2023). *TennisBI: Aplicación Shiny para la visualización de datos de tenis* [Trabajo Fin de Grado]. Universidad de Valladolid [Tutor: María Teresa González Arteaga].

Epi-R Handbook. (2023). *Shiny Basics* [Accedido: 08-05-2025]. https://www.epirhandbook.com/es/new_pages/shiny_basics.es.html

Font Awesome. (2025). *Icons* [Accedido: 22-06-2025]. https://fontawesome.com/icons

Instituto Nacional de Estadística. (2020a). *Metodología del Censo de Población Continuo* [Accedido: 18-06-2025]. https://www.ine.es/dyngs/INEbase/operacion.htm?c=Estadistica_C&cid=1254736177108&menu=metodologia&idp=1254735572981

Instituto Nacional de Estadística. (2020b). *Resultados del Censo de Población Continuo* [Accedido: 18-06-2025]. https://www.ine.es/dyngs/INEbase/operacion.htm?

c=Estadistica_C%5C&cid=1254736177108%5C&menu=resultados%5C&idp=1254735572981

Instituto Nacional de Estadística. (2023, diciembre). *Censo de población. 1 de enero de 2023* [Accedido: 27-05-2025]. https://www.ine.es/prensa/censo_2022_2023.pdf

Instituto Nacional de Estadística. (2025a). *Definición de concepto: Sección censal (código 5228)* [Accedido: 30-04-2025]. https://www.ine.es/DEFIne/concepto.htm?c=5228

Instituto Nacional de Estadística. (2025b). *Metadatos: Respuesta de Datos (Operación 30325)* [Accedido: 04-05-2025]. https://www.ine.es/dynt3/metadatos/es/RespuestaDatos.html?oe=30325

Instituto Nacional de Estadística. (2025c). *Padrones y Padrón Continuo: Métodos y resultados* [Accedido: 19-06-2025]. https://www.ine.es/dynt3/inebase/index.htm?padre=12385%5C&capsel=12384

Instituto Nacional de Estadística. (2025d). *Productos y Servicios: Información estadística* [Accedido: 02-05-2025]. https://www.ine.es/ss/Satellite?L=es_ES%5C&c=Page%5C&cid=1259952026632%5C&p=1259952026632%5C&pagename=ProductosYServicios%5C%2FPYSLayout

Instituto Nacional de Estadística. (2025e). *Relación de municipios y sus códigos por provincia* [Accedido: 02-05-2025]. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177031&menu=ultiDatos&idp=1254734710990

Instituto Nacional de Estadística. (2025f). *Últimos datos disponibles del Censo de Población y Viviendas* [Accedido: 18-06-2025]. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C%5C&cid=1254736177031%5C&menu=ultiDatos%5C&idp=1254734710990

Instituto Nacional de Estadística. (2025g, abril). *Metodología del Censo Anual de Población* [Accedido: 27-05-2025]. https://www.ine.es/dyngs/INEbase/operacion.htm?c=Estadistica_C&cid=1254736177095&menu=metodologia&idp=1254735572981

Lorente, J. (2020). *Capítulo 11: Análisis de Componentes Principales* (inf. téc.) (Accedido: 03-06-2025). Universidad de Granada. http://www.ugr.es/~lorente/APUNTESCN/capitulo11web.pdf

Rousseeuw, P. J. (1984). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, *8*, 283-297.

Rousseeuw, P. J., & Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator [Accedido: 18-06-2025]. *Technometrics*, *41*(3), 212-223. https://doi.org/10.2307/1270566

RStudio. (2023). *shinydashboard: Structure of a DashboardPage* [Accedido: 06-05-2025]. https://rstudio.github.io/shinydashboard/structure.html

Stack Exchange. (2019). *Intuitive explanation of Minimum Covariance Determinant (MCD)* [Accedido: 17-06-2025]. https://stats.stackexchange.com/questions/475636/intuitive-explanation-of-minimum-covariance-determinant-mcd