Investigaciones Regionales – Journal of Regional Research
https://doi.org/10.38191/iirr-jorr.24.057

Journal of Regional Research
Investigaciones Regionales

Accepted Article

Automating the Transfer of Data between Census Sections and Postal Codes Areas Over Time. An application to Spain

Virgilio Pérez[*], *Jose M. Pavía*[**]

Received: 12 September 2023
Accepted: 17 September 2024

Abstract:

Cross-section and longitudinal spatial statistics and econometric models rely on spatially and temporally referenced data. Administrative units like cities, counties, and provinces provide stable data sources, enabling models to combine statistics collected at different times. In Spain, census sections serve as the smallest territorial units in which official statistics are delivered. These areas offer valuable statistics, such as population and housing censuses. Providing these statistics at the postcode level is also pertinent for conducting local analyses and surveys. The issue is that boundaries of census sections undergo regular updates, sometimes involving significant reorganization. The R-package sc2sc automates the transfer of variables between different census sections and postal codes. This paper introduces the package and outlines its methodology, which employs areal weighting to transfer counts and rates.

Keywords: Spatial statistics; longitudinal data; census sections; R-stats; sc2sc; geospatial analysis.
JEL Classification: C31; C87; R11; R12.

Automatización de la transferencia de datos entre secciones censales y códigos postales a lo largo del tiempo. Una aplicación para España

Resumen:

Las estadísticas espaciales transversales y longitudinales, así como los modelos econométricos espacio-temporales, se basan en datos referenciados espacial y temporalmente. Las unidades administrativas como ciudades, comarcas o provincias proporcionan fuentes de datos estables, permitiendo que los modelos combinen estadísticas recopiladas en diferentes momentos del tiempo. En España, las secciones censales constituyen las unidades territoriales más pequeñas donde se distribuyen estadísticas oficiales. En estas áreas se ofrecen estadísticas muy valiosas, incluyendo los censos de población y vivienda, cuya disposición a nivel de códigos postales también es relevante para los análisis locales y las encuestas. El problema radica en las constantes actualizaciones que sufren los límites geográficos de las secciones censales, a veces involucrando reorganizaciones significativas, rompiendo la continuidad en las series de datos. Para automatizar el proceso de transferencia de variables entre diferentes secciones censales y códigos postales, hemos desarrollado en R el paquete sc2sc. Este artículo presenta el paquete y describe la metodología, que utiliza ponderación areal, para transferir conteos y tasas.

Palabras clave: Estadística espacial; datos longitudinales; secciones censales; R-stats; sc2sc; análisis geoespacial.
Clasificación JEL: C31; C87; R11; R12.

[*] Departamento de Economía Aplicada. Métodos Cuantitativos para la Economía y la Empresa. Facultad de Economía. Universidad de Valencia. Spain. virgilio.perez@uv.es

[**] Departamento de Economía Aplicada. Métodos Cuantitativos para la Economía y la Empresa. Facultad de Economía Universidad de Valencia, Spain. pavia@uv.es

Corresponding author: virgilio.perez@uv.es

1. Introduction

The size and structure of a population according to its location and its evolution over time constitute highly valuable information that allows us to understand the demographic dynamics of a region. Knowing how the population is distributed and how it moves within a territory is useful for comprehending urbanization patterns, population density, and resource distribution, playing a fundamental role in infrastructure planning and risk management (Liu et al., 2016). Population evolution is also a key indicator for measuring the economic and social progress of a territory. Population growth may be related to increased production and employment, while population decline can be an indicator of economic or social issues in the region (Hopenhayn et al., 2022).

The distribution of the population according to its location also has significant implications for public policy planning and resource allocation (Lenihan et al., 2019). Information about the geographic distribution of the population is used by social planners and public officials to determine which areas require more attention and resources, providing the infrastructure and services that improve the quality of life for citizens (Bandrés & Azón, 2021).

To analyse population dynamics by zones/areas and establish spatiotemporal comparisons, it is necessary to have data that meet certain minimum quality standards, including spatial and temporal referencing (Pérez et al., 2021). However, obtaining this type of data is not straightforward, at least not with a sufficient level of disaggregation and/or within an appropriate temporal framework. Therefore, we believe that it could be highly beneficial, both for the scientific community and for society at large, to have a tool that allows for data movement in space and time among the smallest spatial units used in Spain: census sections. This would facilitate data imputation when changes occur in their spatial boundaries, allowing for the continuity of data series. Additionally, we address the issue of data translation to and from postal codes. Since much relevant information is collected at this level of aggregation (e.g., when surveys are used), we believe that addressing this problem also adds value.

This article aims to address the problem of data translation between census sections and/or divisions into postal codes, offering a solution to the issue of incomplete information (missing data). The remainder of the document is organized as follows. The second section details the tools used to carry out this project, highlighting the difficulties encountered as well as the solutions we have provided to work with these tools. In the third section, we present the implemented methodology, distinguishing between two different mechanisms depending on the type of data we want to translate/impute. The fourth section showcases examples of results obtained, with a particular focus on the tool we have developed: the R package sc2sc (Pérez & Pavía, 2023). An example is developed in section fifth. Finally, the last section is dedicated to discussing the work carried out and presenting some brief conclusions.

2. Spatial objects

For the completion of this research, we utilized files containing georeferenced digitizations of all the census sections in Spain (INE, 2023). These files, in SHP (shapefile) format, are available on the INE website (at the time of writing this article) for the years 2001 to 2023, with the exception of 2002. These datasets/lists, for the year 2023, comprise over 36,000 observations (census sections) and 17 variables, which are presented below:

- Codes for the census sections (CSEC), districts (CDIS), municipalities (CMUN), provinces (CPRO), and autonomous communities (CCA);
- Unified codes for census sections (CUSEC), districts (CUDIS), and municipalities (CUMUN) y (CLAU2);
- Names of municipalities (NMUN), provinces (NPRO), and autonomous communities (NCA);

- Codes corresponding to the 4 levels of statistical territorial units (NUTS): CNUT0, CNUT1, CNUT2, and CNUT3;
- Geometries of the census sections (geometry).

However, the configuration of these objects has not always been consistent. In the older shapefiles (years 2001 to 2010), the cartography is provided in two separate files, one for the Iberian Peninsula and the Balearic Islands, and another for the Canary Islands, implementing two different coordinate projection systems. In the first case, a cartographic projection based on the European ED50 system and UTM zone 30N is used, while in the Canary Islands, the global WGS 84 cartographic projection and UTM zone 28N are utilized. Starting from 2011, the census section data for Spain is offered in a single file (ETRS89 system and UTM zone 30N).

The number and nature of variables included in the datasets have also changed over the years (see Table 1). In the early editions, fundamental information such as municipality names, province, and autonomous community was omitted, limiting the ability to contextualize the data. The absence of a unique identifier is also relevant, making it difficult to link each census section unambiguously, which hinders without the use of spatial tools data comparison and analysis over time. However, in the period from 2001 to 2010, other variables were provided, such as *shape_length* (polygon perimeter in meters) and *shape_area* (polygon area in square meters). These variables were removed from the datasets in 2011, potentially affecting the researcher's ability to conduct specific analyses.

However, the most significant change in the INE shapefiles, without diminishing the importance of the points mentioned earlier, has been the unification of polygons. In the older datasets, each observation or row is associated with a geographic polygon rather than an individual census section. This results in the possibility that multiple observations represent multiple polygons within the same census section. However, since 2011, a more precise structure has been adopted in which each observation corresponds exclusively to a single census section, which has improved the coherence and precision of the geospatial data used.

In addition to the above, the presence of gaps or unassigned areas in the cartography of the census sections in Spain, as is the case in the older shapefiles, poses significant challenges in the management and analysis of geospatial data. These inconsistencies can create issues when attempting to relate or translate geospatial information over time since any unassigned area lacks correspondence and cannot serve as an origin or destination in spatial allocation or analysis processes. To ensure the coherence and accuracy of any study involving census data, it is essential that the census-sectional territory forms a single polygon without internal gaps. This allows for a more reliable and complete representation of the geography and demographics of the region. This becomes especially important in research that requires precise and complete allocation of geospatial data at the sectional level to obtain robust and comparable results over time.

These changes in the structure of the datasets pose significant challenges and important considerations when working with census data over time. It is crucial to take these variations into account when conducting analyses and intertemporal comparisons. To address the issues mentioned and have a homogeneous spatiotemporal database, we have made a series of adjustments, which will be discussed in the following section.

Table 1.
Shapefiles with the cartography of the census sections in Spain (2001-2023). Count of observations and variables.

Year	Number of observations			Number of census sections			Number of variables
	Peninsula and Balearic Islands	Canary Islands	Spain	Peninsula and Balearic Islands	Canary Islands	Spain	
2001	33892	1215	35107*	33042	1209	34251*	9
2003	34527	1249	35776*	33288	1238	34526*	8

TABLE 1. CONT.
Shapefiles with the cartography of the census sections in Spain (2001-2023). Count of observations and variables.

Year	Number of observations			Number of census sections			Number of variables
	Peninsula and Balearic Islands	Canary Islands	Spain	Peninsula and Balearic Islands	Canary Islands	Spain	
2004	34698	1259	35957*	33460	1247	34707*	8
2005	34851	1266	36117*	33609	1260	34869*	8
2006	35321	1275	36596*	33813	1269	35082*	8
2007	35349	1311	36660*	34086	1302	35388*	8
2008	35339	1311	36650*	34352	1305	35657*	8
2009	35523	1317	36840*	34535	1311	35846*	7
2010	35308	1321	36629*	34316	1315	35631*	6
2011			35961			35961	23
2012			35978			35978	19
2013			36071			36071	22
2014			36127			36127	19
2015			36229			36229	19
2016			36215			36215	22
2017			36208			36208	25
2018			36288			36288	17
2019			36317			36317	24
2020			36309			36309	21
2021			36333			36333	19
2022			36382			36382	17
2023			36462			36462	17

Notes: With an asterisk (*), figures obtained by summing the number of observations from the dataset for the Iberic Peninsula (Spanish part) and the Balearic Islands and the dataset for the Canary Islands.
Source: Own compilation based on INE data.

3. METHODOLOGY

The imputation process has been carried out in two clearly differentiated phases: a spatial/cartographic phase, in which the percentage occupied by each census section with respect to another(s) census section(s) in the previous/posterior year was determined, and a mathematical calculation phase, which allows for data imputation.

3.1. SPATIAL MATCHING

The first phase of the work carried out focuses on spatial comparison and matching. The spatial matching process, understood as a data processing technique used to find associations between two or more data sets that share a common spatial dimension (Walter & Fritsch, 1999), becomes an essential tool in data transfer between different temporal moments (and/or spatial partitions) of territorial cartography (Mitxelena-Hoyos & Amaro-Mellado, 2023). In this methodology, the fact that polygons are uniquely geographically identified over time is exploited, allowing for tracking their changes and origins (Pavía &

Cantarino, 2017a; Pavía & Cantarino, 2017b). This approach considers both the spatial and temporal dimensions to achieve precise and reliable correspondence between different frames of reference.

The polygons that make up a territorial cartography, at the same point in time, must adhere to several properties: i) each polygon is identified by a unique code; ii) no polygon overlaps with another; and iii) the union of all polygons configures the surface and boundaries of the entire territory. These properties allow for a clear connection of cartographic representation across different time periods, enabling the identification of polygons that have undergone changes over time and tracking the origins of those changes. By comparing polygons with the same identifier code at different time points, it is possible to determine if they have undergone alterations and, if so, which other polygons have contributed to the creation or modification of the parcel in question.

By analysing polygons with matching identifier codes at different time periods, it is possible to detect the changes that each polygon has undergone. This includes identifying whether a polygon has been divided into smaller fragments, if multiple polygons have been combined to form a new one, or if there have been other types of geometric alterations. Additionally, this approach also allows for tracking the origin relationships of polygons resulting from these transformations, identifying which polygons have contributed to the creation of a particular polygon later.

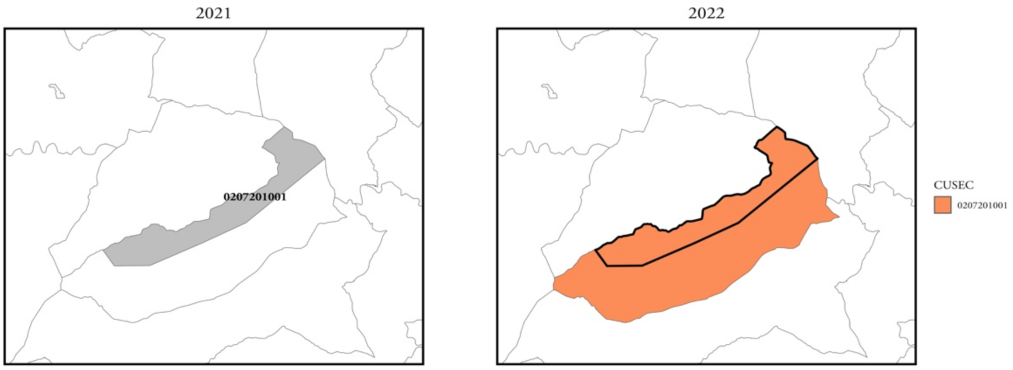
To transfer information spatially and temporally, we compared all the census sections (census polygons) for all available years (pairs of consecutive years). To do this, we had to address certain inconveniences mentioned earlier that are present in the older datasets (2010 and earlier). In Spain, each census section is uniquely identified by a ten-digit code. Similarly, each postal code is uniquely identified by a five-digit code. Thus, we generated, following the INE criteria, the variable CUSEC (unitary section code) by concatenating the province, municipality, district, and census section codes. We used this variable as a unique identifier. Next, we ensured that all observations had a 10-digit CUSEC (2 digits to identify the province, 3 digits for the municipality, 2 digits for the district, and 3 digits for the census section). This verification process allowed us to detect and rectify some errors in the INE datasets. Finally, we grouped observations (polygons) with the same CUSEC using the *st_union()* function from sf R package (Pebesma, 2018).

Having completed the preliminary phase, we have addressed the spatial matching phase. Let's consider that the cartography of a particular territory, denoted as S, is comprised of n polygons. Except for possible changes caused by certain natural phenomena (such as coastal effects or earthquakes, among others), the boundaries of a territory do not vary from one year to the next (Picuno et al., 2019). Thus, the union of the polygons that make up S at a given time t (A) and S at a time $t \pm 1$ (B) coincides, although the polygons that constitute A and B may not necessarily match in number or dimension. To identify the possible changes that a particular polygon (portion of territory) has undergone over time, we overlaid each polygon from A onto the cartography of B. To do this, we used the *st_intersection()* function from sf R package (Pebesma, 2018). This allowed us to determine i) which polygons from B make up each polygon from A and ii) what proportion of each polygon from B is included within each polygon from A.

In the case at hand, after aligning, for each dataset, the number of observations and the number of CUSEC, we compared the geographical boundaries of census sections from two consecutive years to detect which polygons have changed. This comparison allowed us to identify which census sections have undergone any changes compared to the previous/posterior year. Consequently, we could determine the combination of census sections from $t-1$ that contributed to a specific census section at t, and the combination of census sections from t that emerged from a particular census section at $t-1$.

Let's illustrate the methodology used with a simple example. In the left panel of Figure 1, the surface of the census section with CUSEC 0207201001 for 2021 has been represented. This portion of territory intersects entirely with the census section of the same code for 2022 (see the right panel of Figure 1).

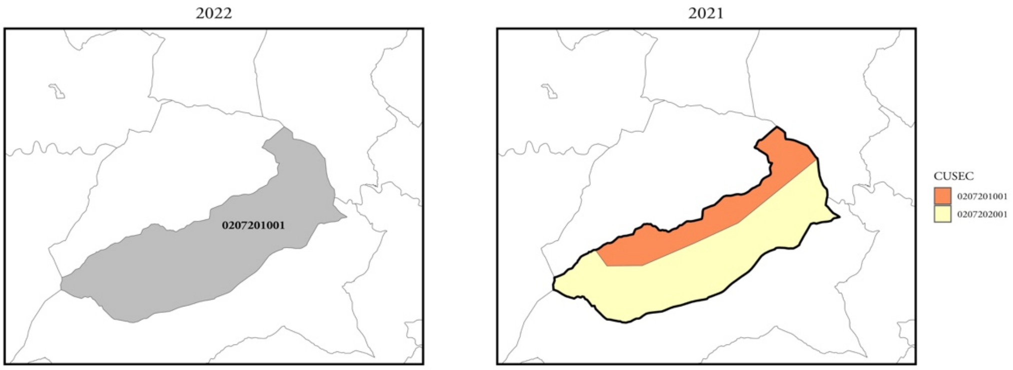
FIGURE 1.
Spatial matching in which it can be observed how the census section with CUSEC 0207201001 for 2021 (shaded area in the left panel) becomes part of a single census section for 2022 (right panel)



Source: Own compilation based on INE data.

If we analyse which census sections from 2021 have contributed to the census section with CUSEC 0207201001 for 2022, we can see that there are two: those with CUSEC 0207201001 and 0207202001 (see Figure 2).

FIGURE 2.
Spatial matching in which it can be observed how the census section with CUSEC 0207201001 for 2022 (shaded area in the left panel) originates from 2 census sections for 2021 (right panel)



Source: Own compilation based on INE data.

From this spatial matching, we can quantify the proportions of territory, both forward and backward in time. To determine what proportion of territory from 2022 (section B) has originated from a particular section in 2021 (section A), we use the following expression:

$$ratio = \frac{area(sscc_B \cap sscc_A)}{area(sscc_B)} \quad (1)$$

For the example shown in Figure 2, the results obtained are presented in Table 2. It can be observed that 31.27% of the section 0207201001 for 2022 corresponds to section 0207201001 for 2021, while the remaining 68.73% corresponds to section 0207202001 for 2021.

Table 2.
Results of the 2022-2021 spatial matching. Census section for 2022 with unitary section code (CUSEC) 0207201001

CUSEC 2022	$area(sscc_B)$	CUSEC 2021	$area(sscc_B \cap sscc_A)$	ratio
0207201001	138,528,122.0	0207201001	43,321,345.0	0.312726
		0207202001	95,206,777.0	0.687274
			138,528,122.0	1.000000

Source: Own compilation based on INE data.

Similarly, we calculate the percentage occupied by each census section A (in year 2021) with respect to a specific census section B (in year 2022) using the following expression:

$$ratio = \frac{area(sscc_A \cap sscc_B)}{area(sscc_A)} \tag{2}$$

For the previous example (see Figure 1), the results are shown in Table 3. As you can see, the entire territory occupied by the sections with CUSEC 0207201001 and 0207202001 is located in the census section 0207201001 for 2022.

Table 3.
Results of the 2021-2022 spatial matching. Census sections for 2021 with unitary section codes (CUSEC) 0207201001 and 0207202001

CUSEC 2021	$area(sscc_A)$	CUSEC 2022	$area(sscc_A \cap sscc_B)$	ratio
0207201001	43,321,345.0	0207201001	43,321,345.0	1
0207202001	95,206,777.0		95,206,777.0	1

Source: Own compilation based on INE data.

With this method, which considers both spatial and temporal dimensions, we construct (can construct) a coherent system of correspondences. This facilitates the transfer of information and values between different time periods, allowing for effective comparison and precise data imputation. By establishing these correspondences, a solid frame of reference is created for statistical analysis and the generation of models that require consistent and temporally and spatially referenced data.

3.2. Values Imputation

Once the spatial and temporal correspondences are established, a value imputation methodology is implemented to transfer and impute counts and rates between different territorial units. Recall that the main objective of this work is to have data for specific places and moments in time that enable us to develop statistical and econometric models. This process is based on territorial weighting, ensuring that values are redistributed in a proportional and coherent manner based on the characteristics of each unit and its relevance in the overall context. Since we have worked with two types of values (counts and rates), two different imputation mechanisms are distinguished.

3.2.1. Counts

We refer to count imputation when we assign numerical values (such as the number of inhabitants, the number of businesses, or other discrete quantities) to a census section B, based on the available information in a census section A (or a set $A_1, A_2, \ldots, A_i, \ldots, A_n$ of census sections) sharing the same physical space but located in different moments in time (in our case, as a starting point, two consecutive years). For example, if we have the number of inhabitants residing in each census section for a specific year (population), we can obtain (calculate an estimate) the number of inhabitants per census section corresponding to another year, following the criterion of geographical proportionality (equitable distribution throughout the territory).

This calculation is performed using the following expression:

$$result = \sum_{i=1}^{n} product_i = \sum_{i=1}^{n} population_i \cdot ratio_i \tag{3}$$

Where:

$$ratio_i = \frac{area\left(sscc_{A_i} \cap sscc_B\right)}{area\left(sscc_{A_i}\right)} \tag{4}$$

Since the spatial matching process allows us to transfer information from a partition to a different partition of previous or subsequent year, it would suffice to repeat the process between another pair of consecutive years until reaching the desired year. To automate this operation, this functionality has been incorporated into the sc2sc package.

3.2.2. Averages

When working with summary data, the imputation process is done differently. In this case, to transfer rates or averages (such as per capita income, unemployment rate, or other continuous indicators), the weighting ratio is obtained using the following expression:

$$ratio_i = \frac{area\left(sscc_{A_i} \cap sscc_B\right)}{area(sscc_B)} \tag{5}$$

This method allows values to be assigned considering the relative contribution of each donor area based on its intersection area with the target census section, ensuring a precise distribution of rates at the census section level. It also facilitates comparative analyses over time and in different geographic areas with a high degree of reliability.

4. sc2sc R package

The application of the methodology described in the previous section generates 42 files containing correspondences between pairs of (consecutive) years, both forward and backward (2001 and 2003-2023). In these files, data is organized in a structured table format as follows: the first column presents the CUSEC of the source sections; in the following columns, the CUSEC of the census sections for the target year and the proportion it represents with respect to the source section are presented in an alternating manner. Thus, if the territory assigned to a specific census section in the source year is entirely within a census section in the target year, the number of columns for that observation will be 3 (5 columns for two sections, 7 for three sections, and so on), so that for each observation, the sum of all "ratio_x" columns results in 1.

To give the reader a better understanding, we return to the example discussed in previous sections and present the numbers in Tables 4 and 5. The first table shows that 100% of the territory located in the census section with CUSEC 0207201001 in 2021 is assigned to the census section in 2022 with the same code. This does not necessarily mean that the source and destination sections are identical. In fact, Table 5 shows that the territory located in the census section with CUSEC 0207201001 in 2022 is assigned to two census sections in 2021, each with a different weighting (ratio).

Table 4.
Example of correspondence between census sections for consecutive pairs of years. Origin year: 2021; destination year: 2022

sscc_A	sscc_B_1	ratio_1
0207201001	0207202001	1

Source: Own compilation based on INE data.

Table 5.
Example of correspondence between census sections for consecutive pairs of years. Origin year: 2022; destination year: 2021

sscc_B	sscc_A_1	ratio_1	sscc_A_2	ratio_2
0207201001	0207202001	0,687274	0207201001	0,312726

Source: Own compilation based on INE data.

This way of saving the results obtained by applying the proposed methodology offers two main advantages. On one hand, it avoids performing geographical calculations every time an imputation is needed, significantly speeding up the processes. On the other hand, it allows for a significant amount of space savings, making it possible for all the relevant information from the shapefiles to be contained in a few files. This organization also enables the easy detection of correspondences forward and backward in time with a highly visual presentation in tabular data format. It facilitates the calculation of the data to be imputed, as this format allows filtering by source sections, avoiding unnecessary traversal of the entire dataset.

Despite the space-saving advantage compared to the original files, this representation still poses a problem: the excessive size of the output (tabulated data file) due to a large number of blanks (NAs). While it is true that in most cases, the source and target census sections are coincident, there are census sections that are reconfigured into up to 41 new codes, generating observations with 83 columns (most of which only need 3). To address this, we have converted the tables (dataframes) into R objects of type list, which are subsequently saved in files with RDS (R Data Serialization) extension. This has significantly reduced the sizes of the objects and made them available in an R package with its space limitations (5MB, at the time of producing the package).

The main goal of the authors is not to quantify the spatial matching (obtain the ratios) for all census sections in Spain over the last 20 years, but to automate the process of imputing certain socioeconomic and demographic variables, allowing for the transfer of information between census sections over time.

To achieve this, the authors of this document have created sc2sc, an R package that enables the imputation of rates and counts for any year using the spatial information mentioned earlier. This package is available on the Comprehensive R Archive Network (CRAN) and provides the ability to work with both census sections and postal codes. The sc2sc package offers researchers three functions: *sc2sc()*, *sc2cp()*, and *cp2sc()*.

The first of these functions allows for the transfer of statistics between different partitions based on census sections (either wholly or partially), in accordance with what has been discussed in this document. To do this, the researcher must provide the years of origin and destination, the actual values for the census sections of the origin year, and the type of data they are working with (counts or averages). Upon executing this function, a set of values paired with the codes of census sections (CUSEC) in the destination year is obtained.

However, the authors have considered it would be beneficial to enrich the sc2sc package by incorporating information obtained from another spatial division that has raised controversy in Spain: postal codes (Goerlich, 2022). In Spain, postal codes are used in many areas, such as logistic or transportation. These are five-digit codes, where the first two digits refer to the province (or autonomous cities), ranging from 01 to 52. If these codes are treated as numerical values, four-digit postal codes (and nine-digit census section codes) may be incorrectly generated. Therefore, the sc2sc package automatically converts numeric codes to character strings, padding with leading zeros to achieve five digits for postal codes and ten for census sections, ensuring consistency in the input dataset.

Although it is common for a postal code to be associated with a single municipality, there are cases where the same postal code covers multiple localities or even parts of a city. This is because postal code assignment is done with the aim of optimizing mail distribution and logistics, prioritizing efficiency in mail delivery. As a result, it is frequent to encounter areas where a postal code encompasses different municipal areas, which can lead to some confusion, especially in regions with a high population density or in metropolitan areas. This highlights the importance of having spatial correspondence not only between census sections at different points in time but also between census sections and postal codes. To facilitate data transfer and imputation, combining both delineations, we propose the function *sc2cp()* to impute data to postal codes based on census sections and the function *cp2sc()* for the reverse case.

5. An empirical application

The transfer of statistics between census section partitions is useful for solving a multitude of problems. Typical examples are found in electoral studies, where past electoral outcomes play a pivotal role to many analysts. Among other uses, they are utilized by the media and political party teams to complete quick assessments of results, by political scholars, sociologists, and electoral geographers to perform detailed scrutinizes, and by pollsters and forecasters to predict electoral results. For instance, they are used in ecological inference models to estimate voter transitions between elections (Pavía and Romero, 2024) or in election night forecasting models to produce educated guesses of the final results (Pavía-Miralles, 2005). These models exploit the structures of covariations between the votes recorded for the different political options in both elections, together with sociodemographic data, to learn about the changes. Typically, the vote shares (or their logit transformations) gained by each party in two consecutive elections is linearly related.

In this section, we demonstrate the usefulness of the sc2sc package by (i) temporally translating the proportion of votes recorded at the census section level in the 2019 local election of the Spanish city of Valencia to 2023 and (ii) graphically comparing the transferred proportions with those actually recorded in the 2023 census sections.

For this example, we used data sourced from the Spanish Electoral Archive Database (Pérez et al., 2021) at the level of census section for the city of Valencia and corresponding to the last two local election held. This resulted in two datasets: one for 2019 (590 observations/census sections) and another for 2023 (591 observations). Although the number of observations/rows in each dataset differs only by one, a comparison of the unique codes of the census sections shows that there has been some reconfiguration of the Valencian municipal sections. Specifically, 5 out of the 590 census sections from 2019 are not present in 2023, and 6 census sections from 2023 are not in 2019. This creates breaks in the longitudinal series of votes, complicating comparisons. To address this issue, we implemented the function *sc2sc()*, which facilitates the transfer of information from one year to another. Table 6 provides the R code (script) to replicate this empirical example.

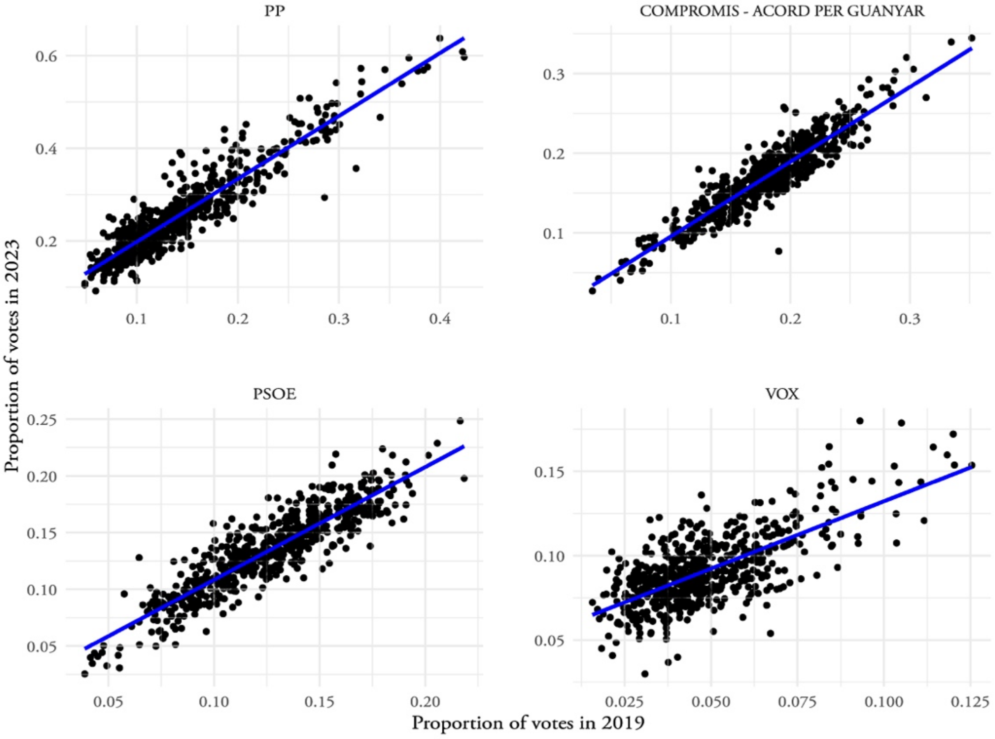
Figure 3 illustrates the proportion of votes transferred from 2019 to 2023 (x-axis) and the proportion of votes recorded in 2023 (y-axis) for the four main political parties in both elections. Each panel of the graph represents one of the four political parties that received the most support in the 2023 elections: People's Party (PP), Acord Per Guanyar (formerly known as Compromís in 2019), Spanish Socialist Workers' Party (PSOE), and VOX. In each panel, the graphical representation demonstrates the expected patterns of change in electoral support over time, without outliers. This supports the notion of a proper transfer of vote proportions with the implemented methodology.

TABLE 6.
R code for replicate the empirical example about temporal transfer of vote proportions between 2019 and 2023 (local elections in Valencia) using sc2sc R package

```
# Library
install.packages("sc2sc")
library(sc2sc)

# Loading the data
Vlc_19 <- read.csv("https://www.uv.es/pavia/VLC_2019.csv", sep= ";")
Vlc_23 <- read.csv("https://www.uv.es/pavia/VLC_2023.csv", sep= ";")

# Transfer of proportions of votes
Vlc_19[, -1] = Vlc_19[, -1]/rowSums(Vlc_19[, -1])
Vlc19_23 <- sc2sc(Vlc_19,
                  year.sscc.origin= 2019,
                  year.sscc.dest= 2023,
                  data.type = "averages")$df

# Comparisons
Vlc_23[, -1] <- Vlc_23[, -1]/rowSums(Vlc_23[, -1])

# Example of plot
plot(Vlc19_23$PP, Vlc_23$PP, xlab="PP 2019", ylab="PP 2023")
```

Source: Own compilation based on SEA Database data (Pavía et al., 2020; Pavía et al., 2024).

FIGURE 3.
Graphical representation of the proportion of votes recorded at the census section level in the city of Valencia during the local elections of 2023 (y-axis) and 2019 (x-axis)



Notes. The values of 2019 have been temporally translated to 2023 using the function sc2sc(). Political parties that obtained a higher percentage of votes in the 2023 municipal elections in Valencia are represented. Blue lines represent linear regressions.
Source: Own compilation based on Ministry of Interior data.

6. Discussion and Conclusions

The approach presented in this article provides a valuable solution to the challenge of dealing with changing territorial units in statistical and econometric models. The automation provided by the R package sc2sc simplifies the process of transferring and imputing missing data, improving the consistency and usefulness of models. This methodology can have applications beyond Spain and can be adapted to other contexts with changing territorial units, such as postal codes. Overall, this study highlights the importance of addressing the spatial and temporal dimensions in building robust models and the utility of automated approaches to achieve this. It is important to note, however, that our package does not include functionalities related to geographical representation of data. This would require incorporating shape files of the different partitions into the package, making it too large to be freely available on CRAN. We refer to interested readers to the official providers of the shape files and to use specialized packages for graphical representation.

Through an example, involving the transfer of electoral results from 2019 to 2023 in the locality of Valencia, we have been able to demonstrate the functionality of the sc2sc R package. This practical application exemplifies how the package can be effectively used to adjust data across reconfigured territorial units, ensuring that comparisons and analyses remain relevant despite changes in census sections. This underscores the sc2sc value in providing researchers with robust tools for dealing with dynamic geographical data landscapes.

It is evident that the data obtained by researchers after applying the functions provided by the sc2sc package do not include certain necessary statistical adjustments in specific situations or for certain types of indicators. However, they can serve as a good starting point for applying other mathematical/statistical operations to further approximate the proposed figures to reality.

For example, by using only the surface area involved and not considering how the population is distributed within the territory, some imputations, especially counts, may be significantly improved. In this regard, it would be interesting in the future to enhance the weighting structures used by employing dasymetric techniques, including additional layers such as cadastre data in the analysis. Similarly, developing an alternative weighting structure for transferring rates using spatial interpolation techniques could be valuable, taking advantage of the spatial autocorrelation structures exhibited by many socioeconomic variables (see, for example, Larraz et al., 2013).

References

Bandrés, E. & Azón, V. (2021). *La despoblación de la España interior.* Funcas.

Goerlich, F. (2022). Elaboración de un mapa de Códigos Postales de España con recursos libres: cómo evitar pagar por disponer de información de referencia. (IVIE working papers, 2022:03). https://doi.org/10.12842/WPIVIE_0322

Hopenhayn, H., Neira, J., & Singhania, R. (2022). From Population Growth to Firm Demographics: Implications for Concentration, Entrepreneurship, and the Labor Share. *Econometrica*, *90*(4), 1879-1914. https://doi.org/10.3982/ECTA18012

INE (2023). Sitio web del Instituto Nacional de Estadística (España). Retrieved from: https://www.ine.es/uc/1dIJtjmE

Larraz, B., Pavía, J.M., & and Ferrari, G. (2013). Weighting Elementary Prices in Consumer Price Index Construction Using Spatial Autocorrelation. *Communications in Statistics – Theory and Methods*, *42*, 4460-4475. https://doi.org/10.1080/03610926.2011.648793

Lenihan, H., McGuirk, H., & Murphy, K. R. (2019). Driving innovation: Public policy and human capital. *Research Policy*, *48*(9), 103791. https://doi.org/10.1016/j.respol.2019.04.015

Liu, Z., He, C., & Wu, J. (2016). General Spatiotemporal Patterns of Urbanization: An Examination of 16 World Cities. *Sustainability*, *8*, 41. https://doi.org/10.3390/su8010041

Mitxelena-Hoyos, O., & Amaro-Mellado, J.L. (2023). A Comparison of Cartographic and Toponymic Databases in a Multilingual Environment: A Methodology for Detecting Redundancies Using ETL and GIS Tools. *ISPRS International Journal of Geo-Information*, *12*, 70. https://doi.org/10.3390/ijgi12020070

Pavía, J. M., & Cantarino, I. (2017a). Can Dasymetric Mapping Significantly Improve Population Data Reallocation in a Dense Urban Area? *Geographical Analysis*, *49*(2), 155-174. https://doi.org/10.1111/gean.12112

Pavía, J. M., & Cantarino, I. (2017b). Dasymetric distribution of votes in a dense city. *Applied Geography*, *86*, 22-31. https://doi.org/10.1016/j.apgeog.2017.06.021

Pavía J. M., Aybar, C., & Pérez, V. (2020). Elecciones Municipales 2019, Valencia. Bases Electorales GIPEyOP. (http://sea.uv.es/gipeyop/sea.html).

Pavía J. M., Aybar, C., & Pérez, V. (2024). Elecciones Municipales 2023, Valencia. Bases Electorales GIPEyOP. (http://sea.uv.es/gipeyop/sea.html).

Pavía, J. M., & Romero, R. (2024). Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming. *Sociological Methods and Research, 187*(4), 919-943. https://doi.org/10.1177/00491241221092725

Pavía-Miralles, J. M. (2005). Forecasts from non-random samples: The election night case. *Journal of the American Statistical Association*, *100*, 1113-1122. https://www.jstor.org/stable/27590658

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439-446. https://doi.org/10.32614/RJ-2018-009

Pérez, V., Aybar, C., & Pavía, J. M. (2021). Spanish electoral archive. SEA database. *Scientific Data*, *8*(1), 193. https://doi.org/10.1038/s41597-021-00975-y

Pérez, V., & Pavía, J. M. (2023). sc2sc: Spatial Transfer of Statistics among Spanish Census Sections. Version 0.0.1-7. https://cran.r-project.org/package=sc2sc

Picuno, P., Cillis, G., & Statuto, D. (2019). Investigating the time evolution of a rural landscape: How historical maps may provide environmental information when processed using a GIS. *Ecological Engineering*, *139*, 105580. https://doi.org/10.1016/j.ecoleng.2019.08.010

Walter, V., & Fritsch, D. (1999). Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, *13*(5), 445-473. https://doi.org/10.1080/136588199241157

ORCID

Virgilio Pérez	https://orcid.org/0000-0002-7628-2855
Jose M. Pavía	https://orcid.org/0000-0002-0129-726X