CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17796

Improving Accuracy in Geospatial Information Transfer: A Population Density-Based Approach

Virgilio Pérez, Jose M. Pavía
Applied Economics, University of Valencia, Spain.

How to cite: Pérez, V.; Pavía J.M. 2024. Improving Accuracy in Geospatial Information Transfer: A Population Density-Based Approach. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17796

Abstract

The R package sc2sc offers fundamental tools for transferring information between census sections and postal codes in Spain, based on the cartography of these geographic segmentations. However, certain aspects for improvement have been identified. This document presents a substantial improvement to the package, optimizing the cp2sc function, which facilitates the transfer of information from postal codes to census sections. The introduced improvement considers population density as a corrective factor in the process, resulting in a more accurate and relevant data allocation. Various use cases highlight the improvement of the new methodology, though they also underline the need to work with updated and precise cartography, opening new lines of research and future work.

Keywords: *Spatial Statistics; Longitudinal Data; Census Sections; Postal Codes; Sc2sc; R-Stats; Geospatial Analysis.*

1. Introduction

In Spain, several geographic segmentations coexist. Census sections represent the lowest level of spatial aggregation used by the National Statistics Institute (INE). At this level, the INE collects and distributes demographic and socioeconomic data, including statistics related to population and housing censuses. These geographical units, which generally group between 1,000 and 2,000 people, are crucial for the detailed study of population characteristics, allowing precise segmentation for urban planning and public policy development (Pillet et al., 2013). On the other hand, postal codes, assigned by Correos (the national postal service of Spain), are designed to optimize the postal distribution system and can cover areas ranging from small urban zones to larger regions in non-urban areas (Otero et al., 2019). Postal codes, while considering the spatial distribution of the population across the territory, reflect the organization of the postal network.

Although both geographical systems serve different purposes, they are indispensable for various applications (such as logistics, marketing, and socioeconomic analysis, among others), offering complementary perspectives on the distribution and organization of the population and services in the country (Reina-Usuga et al., 2020). The coexistence of both geographic segmentation systems implies that certain information is available for one level of aggregation but not for the other (Manzanares & Riquelme, 2017). This complicates, if not directly prevents, analysis that includes certain variables, hindering the work of researchers and other social agents.

In this context, having a tool that allows the transfer of information over time between census sections and/or postal codes could transform the way researchers and analysts can understand population and territorial dynamics. Such capability enables a more detailed and accurate analysis of demographic, economic, and social trends over time, facilitating decision-making and the formulation of public policies. For example, by analyzing how the population characteristics of a specific area evolve, urban planners can design more inclusive and sustainable cities, while economists can assess the impact of economic policies in different regions (Ropero et al., 2015; Thompson et al., 2020).

Aiming to offer this analytical capability, the authors of the present document have developed the sc2sc R package (Pérez & Pavía, 2024), a tool that facilitates the transfer of information between census sections and postal codes in Spain, using the cartography (territorial boundaries) of both geographic segmentation methods. This package exports three functions: i) sc2sc, that allows the transfer of statistics between census sections that coincide (totally or partially) at different points in time; ii) sc2cp, that enables the transfer of information from census sections to postal codes; and iii) cp2sc, that facilitates the transfer of information from postal codes to census sections.

The functionalities offered by sc2sc are based on an approach that integrates and analyzes the cartography of census sections, provided by the INE (2024a), and the cartography of postal codes generated by Goerlich (2022). The basic solutions it offers are based on the intersections between polygons representing census sections in consecutive years, or between the polygons of census sections and postal codes from 2019. This approach allows for the transfer of statistics between census sections from different time periods, as well as between census sections and postal codes and vice versa, covering a time range from 2001 to 2023, excluding 2002.

However, we have observed that when transferring information from postal codes to census sections using the basic solution, the distribution is not consistent. Working under the assumption that statistics (variables) are uniformly distributed throughout the territory of a postal code, we are assigning greater importance (weight) to those census sections that are larger, when it is common that the census sections with greater extension are those with less population density.

This document provides a possible solution to this problem by including population density in the mathematical formula that allows for the transfer/imputation of statistics to census sections from postal codes. The rest of the document is structured as follows: In the second section, we present the implemented methodology. In the third section, we demonstrate the opportunity of the new approach by developing some examples. In the fourth section, the work is discussed, and brief conclusions and some ideas for future work are presented.

2. Methodology

A postal code (cp), spatially defined as a polygon, intersects with $n \geq 1$ census sections (sc), also polygons. Thus, if we want to transfer/distribute/impute a value L available in cp among sc, we could leverage in the relationship

$$L = \frac{|sc_1 \cap cp|}{|cp|} \cdot L + \frac{|sc_2 \cap cp|}{|cp|} \cdot L + \cdots + \frac{|sc_n \cap cp|}{|cp|} \cdot L, \quad (1)$$

meeting the condition

$$\frac{|sc_1 \cap cp|}{|cp|} + \frac{|sc_2 \cap cp|}{|cp|} + \cdots + \frac{|sc_n \cap cp|}{|cp|} = 1, \quad (2)$$

where $|A|$ denotes the area of polygon A, to make the transfer.

However, as mentioned in the introduction, this method is not consistent in terms of reversibility for much of the Spanish territory. This is due to the different criteria implemented to configure each territorial segmentation. While census sections group between 1,000 and 2,000 inhabitants, the delimitation of postal codes presents more heterogeneity in terms of resident. To address this imbalance, a significant improvement has been introduced in the methodology: weighting by population density. This adjusts the allocation of information not only based on the area of the census section but also considering its population density.

Let $D_i = P_i/A_i$ be the population density of the target census section, obtained as the quotient between the resident population in that census section, P_i, and the area, A_i, of the polygon. To transfer a value L from cp to sc we can take into account D_i, as it is verified that:

$$L = D_1 \cdot \frac{|sc_1 \cap cp|}{P_{cp}} \cdot L + D_2 \cdot \frac{|sc_2 \cap cp|}{P_{cp}} \cdot L + \cdots + D_n \cdot \frac{|sc_n \cap cp|}{P_{cp}} \cdot L, \quad (3)$$

where $P_{cp} = \sum_{i=1}^{n} D_i \cdot |sc_i \cap cp|$. Meeting the condition that:

$$D_1 \cdot \frac{|sc_1 \cap cp|}{P_{cp}} + D_2 \cdot \frac{|sc_2 \cap cp|}{P_{cp}} + \cdots + D_n \cdot \frac{|sc_n \cap cp|}{P_{cp}} = 1. \quad (4)$$

Levearing on this expressions allows for a more equitable and representative distribution of information, reflecting more accurately the real distribution of the population and its characteristics within the geographic units of analysis.

3. Results

To assess the methodology described in the previous section, several examples have been developed, checking the consistency of the proposed approach based on the number of inhabitants per census section in 2019. Figure 1 depicts the municipality of Alaquàs (Valencia), with postal code 46970 and municipality code 46005. The blue line indicates the boundary of the postal code, while the red lines reflects the boundaries of the 21 census sections that make up the municipality. In this case, the correspondence between postal codes and census sections is 1:n, meaning i) a single postal code encompasses all the census sections of the municipality, and ii) the census sections that make up the represented postal code do not intersect with any other postal code. As can be observed, the extension of the census sections is very unequal, with the section coded 02001 occupying the most territory of the municipality.

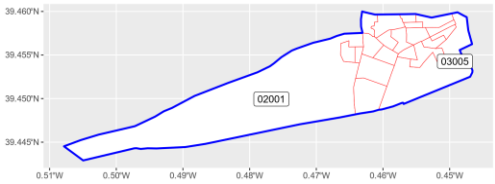

Figure 1. Territorial boundaries (census sections and postal code) of Alaquàs (Valencia) in 2019. Source: own elaboration based on the cartography from INE (2024b) and Goerlich (2022).

In this example, given the structure and composition of the involved polygons, we can assign to postal code 46970 a population equivalent to the sum of the populations of the 21 previously mentioned census sections. With this, we proceed to transfer this information, at the postal code level, to the relevant census sections, using the cp2sc function from the R package sc2sc using version 12, which is based on areal weighting, and using the new latest version, which includes the improvement.

In Table 1, on one hand, the official population of each census section in 2019 is provided (variable POP), as well as the area occupied by each portion of the territory and the ratio to the total (variables AREA and RATIO); and on the other hand, the population estimates calculated, both with the previous method (version 12) and with the method proposed in this document (variables POP_prev and POP_new) alog with the differences, in absolute terms, from the real values (variables diff_prev and diff_new). As can be observed, the previous method, based solely on the proportion of territory occupied by each census section in relation to the extension of the postal code, reports inconsistent values. In contrast, and as would be expected, the new method proposed returns exactly the same population from which it started.

Table 1. Population transfer, by census sections, of the municipality of Alaquàs (Valencia) in 2019.
Source: own elaboration based on the continuous register statistics (INE, 2024b).

CUSEC	POP	AREA	RATIO	POP_prev	diff_prev	POP_new	diff_new
4600501001	1221	0.05	0.012	361	860	1221	0
4600501002	1068	0.08	0.020	588	480	1067	1
4600501003	1349	0.05	0.013	387	962	1348	1
4600501004	1103	0.05	0.013	393	710	1103	0
4600501005	1148	0.04	0.009	278	870	1147	1
4600501006	1238	0.03	0.007	210	1028	1239	1
4600501007	715	0.02	0.004	127	588	715	0
4600501008	723	0.02	0.005	151	572	724	1
4600501010	1934	0.06	0.015	435	1499	1933	1
4600502001	2133	2.34	0.601	17757	15624	2131	2
4600502002	1743	0.06	0.014	423	1320	1744	1
4600502003	1392	0.04	0.010	304	1088	1392	0
4600502004	2515	0.12	0.032	934	1581	2516	1
4600502005	1667	0.05	0.012	352	1315	1667	0
4600502006	1002	0.02	0.004	133	869	1002	0
4600502007	1308	0.15	0.038	1123	185	1307	1
4600503001	1022	0.03	0.007	213	809	1023	1
4600503002	1243	0.03	0.006	192	1051	1242	1
4600503003	1248	0.03	0.008	225	1023	1247	1
4600503004	1701	0.08	0.020	600	1101	1700	1
4600503005	2088	0.58	0.148	4372	2284	2087	1
TOTAL/ERROR	**29561**	**3.90**	**1.000**	**29558**	**35819**	**29555**	**16**

A second example has been developed following the procedure outlined for the previous case. In this instance, the municipality of Alboraia (Valencia), with postal code 46120 and municipality code 46013, was taken as a reference. As can be seen in Figure 2, we again find differences between the two cartographies, primarily in the eastern boundaries of the municipality (beach/coast), generating certain inconsistencies in the results.

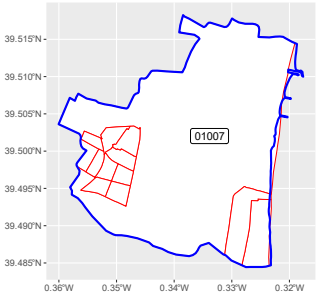

Figure 2. Territorial boundaries (census sections and postal code) of Alboraia (Valencia) in 2019. Source: own elaboration based on the cartography from INE (2024b) and Goerlich (2022).

Table 2. Transfer ot number of voters, residents in Alboraia (Valencia), who participated in the 2019 Valencian Courts elections. Source: own elaboration based on the results of the 2019 Valencian Courts elections (Pérez et al., 2021).

CUSEC	VOTES	AREA	RATIO	VOT_prev	diff_prev	VOT_new	diff_new
4601301001	837	0.09	0.011	152	685	811	26
4601301002	690	0.11	0.013	185	505	662	28
4601301003	908	0.09	0.011	151	757	890	18
4601301004	1509	0.07	0.009	123	1386	1407	102
4601301005	975	0.05	0.006	82	893	1050	75
4601301006	999	0.08	0.009	131	868	1013	14
4601301007	1250	6.89	0.827	11668	10418	1365	115
4601301008	598	0.07	0.008	120	478	557	41
4601301009	937	0.06	0.007	100	837	978	41
4601301010	1565	0.07	0.008	114	1451	1482	83
4601301011	1204	0.09	0.010	150	1054	1120	84
4601301012	1330	0.33	0.040	548	782	1382	52
4601301014	1310	0.35	0.042	588	722	1393	83
4625011033			0.000	0	0	0	0
4625011039			0.000	0	0	0	0
4625011050			0.000	0	0	0	0
TOTAL/ERROR	**14112**	**8.33**	**1.000**	**14112**	**20836**	**14110**	**762**

In this example, we illustrate the differences using a variable of a different kind. In particular, we consider the number of voters who participated in the 2019 Valencian Courts elections. This

case notably elucidates the versatility of sc2sc in addressing analysis in different fields, including the political-electoral sphere.

From these data and the results obtained using the cp2sc function, Table 2 was constructed. Again, the values obtained after applying the methodology proposed in this paper fit more closely with reality (the transfer error is this time more than 27 times smaller). Also in this case, as in the previous one, differences between the estimated values and the actual values (number of voters) are observed, primarily derived from divergences in turnout rates. In addition to the above, and due to cartography issues, values were imputed to three census sections of neighboring municipalities, which, although it did not imply any significant imputation, highlights the relevance of working with consistent geographical data.

4. Discussion and Conclusions

The implementation of the sc2sc package has highlighted the complexity and challenges inherent in the task of transferring information between different levels of geographic aggregation, especially between census sections and postal codes in Spain. The solution proposed in this document, which incorporates population density into the process of transferring information from postal codes to census sections, not only addresses a key limitation of an approach based exclusively on surfaces but also highlights the gains in accuracy that can be obtained in the transferred data, which align more closely with demographic and geographic reality.

The examples presented illustrate the improvements of the proposed methodology in different municipal contexts, demonstrating its ability to more faithfully reflect the distribution of the population within the census sections, even in scenarios of complex correspondence between postal codes and census sections. However, the differences observed between the estimated and actual populations underscore the critical importance of having precise and updated cartographies, as well as the need for additional adjustments to improve the congruence between the geographical data used in the analysis.

Future work should focus on refining these aspects, exploring advanced techniques for handling cartographic discrepancies, and optimizing information transfer algorithms to increase their applicability and accuracy. Likewise, expanding the temporal range covered and including new variables could significantly enrich the analytical capabilities of the sc2sc package, opening new avenues for research and geospatial analysis in Spain.

References

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. doi.org/10.12842/WPIVIE_0322

INE (2024a). Instituto Nacional de Estadística. Cartografía secciones censales y callejero de Censo Electoral. Retrieved from: https://www.ine.es/uc/1dIJtjmE

INE (2024b). Instituto Nacional de Estadística. Estadística del Padrón continuo. Retrieved from: https://www.ine.es/uc/XUXtgIDdi1

Manzanares, A. and Riquelme, P.J. (2017). Análise espacial do desemprego nos mercados locais de traballo españois. *Revista galega de economía: Publicación Interdisciplinar da Facultade de Ciencias Económicas e Empresariais*. *26*(2), 29-42.

Otero, R., García, J., Domínguez, J., and Pérez, A. (2019). Inmigración y dinámicas territoriales en España: crisis y recuperación (2008-2017). In E. Mañé (Ed.), *Anuario CIDOB de la Inmigración (2019)* (pp. 190-217). Barcelona: Ediciones Bellaterra. doi.org/10.24241/AnuarioCIDOBInmi.2019.190

Pérez, V. and Pavía, J.M. (2024). sc2sc: Spatial Transfer of Statistics among Spanish Census Sections. Version 0.0.1-12. https://cran.r-project.org/package=sc2sc

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

Pillet, F., Cañizares, M.C., Ruiz, A.R., Martínez, H.S., Plaza, J., and Santos, J.F. (2013). Los indicadores de la cohesión territorial en el análisis de la escala supramunicipal o subregional. *Ería: Revista cuatrimestral de geografía*, *90*, 91-106.

Reina-Usuga, L., Haro-Giménez, T., and Parra-López, C. (2020). Food governance in Territorial Short Food Supply Chains: Different narratives and strategies from Colombia and Spain. *Journal of Rural Studies*, *75*, 237-247. doi.org/10.1016/j.jrurstud.2020.02.005

Ropero, R.F., Aguilera, P.A., and Rumí, R. (2015). Analysis of the socioecological structure and dynamics of the territory using a hybrid Bayesian network classifier. Ecological Modelling, 311, 73-87. doi.org/10.1016/j.ecolmodel.2015.05.008

Thompson, M., Nowak, V., Southern, A., Davies, J., and Furmedge, P. (2020). Re-grounding the city with Polanyi: From urban entrepreneurialism to entrepreneurial municipalism. *Environment and Planning A: Economy and Space*, *52*(6), 1171-1194. doi.org/10.1177/0308518X19899698