Journal of Internet Services and Applications, 2026, 17:1, doi: 10.5753/jisa.2026.6795

Restoring Continuity: An Aggregative, Open-Source Methodology for Harmonizing Disparate Census Tracts Across Time

Pedro Henrique Rezende Mendonça [**University of São Paulo** | ***pedrorezende@ime.usp.br***]
Fabio Kon [**University of São Paulo** | ***kon@ime.usp.br***]

Institute of Mathematics, Statistics and Computer Science at the University of São Paulo, São Paulo, Brazil.

Received: 05 November 2025 • **Accepted:** 30 March 2026 • **Published:** 25 April 2026

Abstract
Urban analysis often relies on census or census-like survey data, but the redesign of census tract layers often breaks historical comparability, thereby hampering longitudinal studies at the local scale. In this article, we propose an automated methodology for making census tracts comparable through the construction of a graph, using spatial join operations and non-spatial id comparisons. We adopt an aggregative heuristic to minimize the impact of the matching process on analytical possibilities. To validate the proposed methodology, we deploy it to match the most recent Brazilian censuses for the São Paulo, Vitória, and Recife metropolitan regions, as well as for the Brazilian Federal District. We also test our implementation for New York City and Buenos Aires. We then evaluate the results against available gold standard data and alternative metrics, concluding with a discussion of potential improvements. Our implementation produces comparability files efficiently and has the potential to enable studies that would otherwise be impractical.

Keywords: Census, Spatial data, Data Science, Urban Computing

1 Introduction

Studies on urban processes and urban planning policies often integrate multiple data sources with census data, as these constitute an official source with extensive thematic coverage, nationwide coverage, and temporal regularity. Census spatial layers also often subsume the spatial basis for other official surveys, which, in turn, expand the thematic scope of public statistics. However, urban analytics that rely on historical data series at the local scale commonly face a methodological challenge: the comparability of datasets across different years.

The consistency of longitudinal studies based on surveys such as the census requires (I) that the methodology of data acquisition and aggregation is historically consistent (equivalence of scope), and (II) that the boundaries of the spatial areal units used for data aggregation correspond to the same geographic regions (equivalence of spatial structure). The first condition can be asserted through the technical documentation of each survey, as discussed in Barbosa [2014], and is not the focus of this article. The equivalence of spatial structure, however, depends on the geographic scale and the level of aggregation of the published data and is the focus of our work.

From the municipal scale to the national scale, spatial structure equivalence is often easy to satisfy – e.g., census data aggregated at the municipal level are generally comparable across different censuses, as municipal boundaries rarely change. At a finer scale, however – that of census tracts – the correspondence requires a particular type of data integration. Unlike the implementation of spatial join operations in spatial database management systems or Geographic Information Systems software, the integration of census tract datasets depends not only on the set of geometries but also, simultaneously, on non-spatial indexes. We elaborate on this problem in Section 3.

In this article, we present a method to make local-scale boundary layers comparable over time. We adopt the Brazilian Institute of Geography and Statistics (IBGE) censuses as our case study, as they pose significant incompatibilities between boundary layers. Our method relies on both census tract ids and polygons, through an aggregative heuristic (using sum and union operations). To this end, we resort to the comparability logic adopted in related works. We found that the most common approach in the literature relies on a multiplicative heuristic (using weighted interpolations), implying a certain level of homogeneity within each areal unit. We introduce an efficient algorithm that avoids interpolation, dismissing the local-scale homogeneity assumption and thus enabling direct aggregation of raw census data. Our proposed approach may be particularly valuable for studies of phenomena with high spatial sensitivity – such as the assessment of urban and housing policies and studies in demography or epidemiology – where assuming local homogeneity can lead to inaccurate conclusions and compromise validity.

We implement our method and apply it to a set of cities; then we test it using available gold-standard data and discuss the results to evaluate potential use cases and future improvements. A Python implementation extending our previous work [Mendonça and Kon, 2025] is available as open-source software at `https://github.com/peredrozende/compat_censos`. Beyond scientific impact, our results may serve as an additional tool in helping census bureaus develop and validate longitudinal datasets.

2 Related Work

The challenges of longitudinal spatial comparability between incompatible zoning systems have been addressed by the Spatial Analysis literature for decades [Goodchild *et al.*, 1993]. The problem fundamentally derives from the division of the same geographic region into different, overlapping sets of zones. This renders data attributed to each zoning system not directly comparable – e.g. demographic data from two different censuses –, and specific procedures to achieve comparability are required.

A long trend in comparability methods follows the areal interpolation heuristics [Flowerdew and Openshaw, 1987]. The interpolative approach resorts to a weighted redistribution of data from one zoning system to another. Such weight functions can be calculated in many ways [Goodchild *et al.*, 1993; Comber and Zeng, 2019], sometimes using auxiliary dasymetric data to increase precision. Ruther *et al.* [2015] and Markley *et al.* [2022] converge that target density weighting with dasymetric refinement outperforms many of the most common interpolative techniques, but is sensitive to fast changes in population. Many works deploy methods to measure accuracy or the degree of error introduced by interpolation [Cockings *et al.*, 1997; Gregory, 2002]. Nevertheless, although errors can be estimated, they cannot be eliminated, as they are inherited from the method's assumptions of total or partial inner-zone homogeneity.

Since data aggregation in zoning systems comes from smaller spatial units within each zone – e.g. household plots for census data –, the issue can be traced back to the Modifiable Areal Unit Problem (MAUP), which has also been extensively known and discussed for nearly a century [Gehlke and Biehl, 1934; Openshaw, 1979]. The MAUP states that geographic data can be zoned into an infinite number of polygons of varying size and shape, creating an infinite number of new arrangements for data aggregation [Buzzelli, 2020]. Such data regrouping can produce drastically different spatial relationships. Since interpolative methods rely on two possible zoning systems, their results are equally susceptible to MAUP effects.

If one had access to the original geographic point of data collection, MAUP could be avoided by reaggregating the original point data, and no areal-interpolation assumptions would be needed. Recent work has highlighted the growing use of demographic microdata as an alternative to census zones [Breen and Feehan, 2025], and in some cases, even census data is now available at finer spatial granularity [Junior *et al.*, 2025]. But privacy protection constrains how many variables can be collected or published in point-scale aggregation, and some evidence shows that increasing the level of data exposure might not bring an equivalent increase in analytic quality [Norman *et al.*, 2023]. The problem of longitudinal spatial comparability cannot be solved by higher data precision; nonetheless, as older datasets will still be limited to lower precision areal aggregation.

In this work, we propose an aggregative method to achieve longitudinal comparability while avoiding MAUP and the effects of areal interpolation. More specifically, we focus on the comparability of local-scale census data. The nature of census data breaks some assumptions of areal interpolation: (i) a pair of census zone systems is not completely independent, as census tracts are usually derived from previous censuses, and (ii) a census is spatially organized in a hierarchical fashion, which reduces the possibility of valid polygon overlaps. Our methodology benefits from these specifics by relying on census tract indexing, drawing precision, and hierarchical restrictions.

2.1 Specifics of Census Data

The comparability between census areas is fundamentally tied to the hierarchy of spatial entities in each national census, and related work, in general, is produced by technicians and researchers in close contact with States and their institutions. Our literature review has shown strong adherence to their respective national censuses, which might lead to blind spots in topicality due to language barriers. However, some common issues and trends are recognizable.

Areal interpolation is clearly the dominant method across older and contemporary studies. The availability of auxiliary data and overall technical advancements have advanced interpolation and reduced its limitations Ruther *et al.* [2015]. Studies with an empirical, data-focused approach are far more abundant than those that deal with longitudinal comparability as a topic per se. We have found works with censuses from Spain [Pérez and Pavía, 2024], Australia [Nguyen *et al.*, 2023], Canada [Allen and Taylor, 2018; Dias and Silver, 2021], Argentina [Rodríguez, 2021], the United States [Dias and Silver, 2021; Markley *et al.*, 2022; Jurjevich *et al.*, 2025; US Census Bureau, 2024a], and Brazil [Lobo, 2009; Reis, 2013; Hirye *et al.*, 2016; Yamaguchi, 2017]. These works rely on an interpolative heuristic.

Allen and Taylor [2018] published a longitudinal census tract database for Canada using dasymetric areal interpolation and conflation procedures to reduce the impact of polygon translation. Conflation is particularly useful when tract geometry details are reasonably consolidated across censuses. Many longitudinal databases are available for the US [Schroeder, 2007; GeoLitycs, 2007; Logan *et al.*, 2014; Schroeder *et al.*, 2025] using similar interpolative methods. Australia also has an equivalent longitudinal dataset in which comparability is established across a sampled group of individuals [Nguyen *et al.*, 2023]. The existence of ready-to-use longitudinal datasets facilitates several applied studies in Spatial Analysis and should be pursued in countries with long time series of census data.

In our work, we focus on the parallels between the Brazilian census and the structure of the censuses of the United States and Argentina, which also use geographic entities equivalent to census tracts (*setores censitários* in Brazil, *radios censales* in Argentina) and have a territorial extent of the same order of magnitude as Brazil. The Brazilian censuses constitute a particular challenge since no longitudinal database is available and tract-level changes are significant between censuses.

In the United States, spatial structure equivalence at the local scale can be directly ensured using the "relationship files" between tracts, which are available from the US Census Bureau [2024b]. These files allow, for example, the creation of cartographic visualizations of population variation (see Figure 1). The construction of relationship files relies solely on the spatial overlap of polygons, meaning that any minor inconsistency in polygon tracing could result in a mistaken relation. However, US census tracts are defined as groups of census blocks, which in turn are bounded by natural delimiters (streets, rivers, administrative boundaries). The risk of identifying false census tract redesign operations is low, as these natural delimiters constitute a base reference layer that rarely changes.

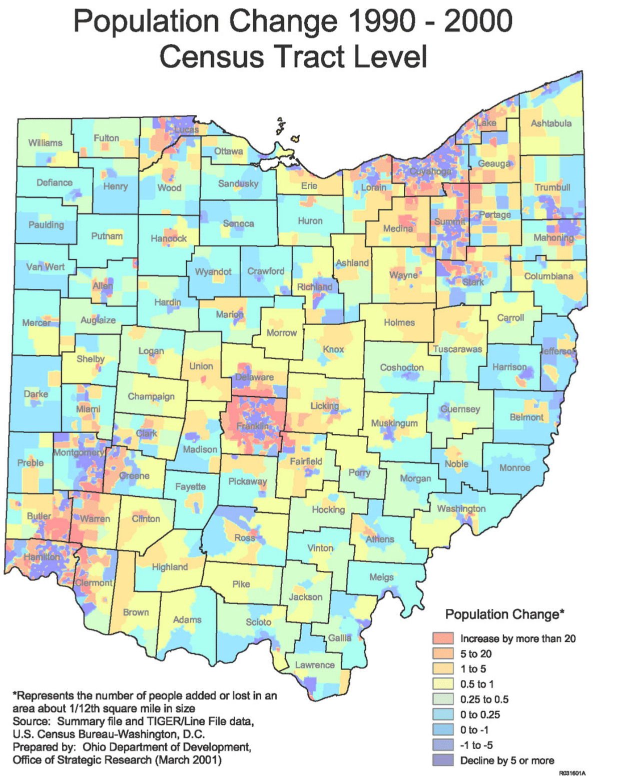

Figure 1. Demographic changes in Ohio mapped using census tract comparable areas [US Census Bureau, 2024a].

For the Argentine census, Rodríguez [2021] proposes and applies a method of tract "correction" across the 1991, 2001, and 2010 censuses through spatial union operations. The process aims to adjust inaccuracies in tract boundaries by using fixed layers of natural delimiters (see Figure 2). The method is tailored for each census, employing validation criteria for corrections, such as the distance between the centroids of two tracts or the visual, human-assisted overlap of polygon borders. The methodology was applied to the whole extent of Argentina, meaning that intensive human work was required to check and edit tract geometry. All software used in the process (ArcGIS, Excel, SPSS) is proprietary. The result is available as an official file by the National Scientific and Technical Research Council (Conicet), enabling straightforward creation of compatibility tables by polygon overlay.

In the Brazilian context, the major difficulties for comparability are the geographic imprecision in polygon tracing and the absence of complete spatial coverage in digital files for the 1991, 2000, 2010, and 2022 censuses [Hirye *et al.*, 2016; Lobo, 2009; Reis, 2013]. Yamaguchi [2017] categorizes studies on the comparability of Brazilian census tracts according to the use of auxiliary dasymetric data, such as satellite imagery or socioeconomic indicators. The works

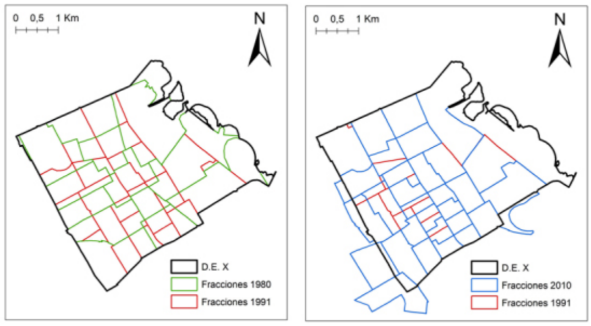

Figure 2. Breaks in retrospective comparability between 1980-1991 and 1991-2010 censuses in Distrito Escolar X, Buenos Aires, according to [Rodríguez, 2021]

reported by the author rely on areal interpolation to reaggregate data onto a new regular grid, or use *compatibility tables* (equivalent to the US Census Bureau's relationship files) at larger geographic scales.

Compatibility tables link redesign operations to each of the census units involved, enabling the identification of corresponding areas across censuses and the reconstruction of spatial structure equivalence. Yamaguchi cites the work of the Institute of Applied Economic Research (IPEA), which uses *Minimum Comparable Areas* (MCAs) derived from compatibility tables of municipal grids from 1872 to 2000 [Reis *et al.*, 2008], which have been employed in many publications of the institute.

The principle of MCAs is to ensure that data are reaggregated into a geographic polygon layer consistent with the methodologies employed in the historical series. By performing only sum operations on quantitative variables and polygon merging, the methodology of data collection and aggregation remains unaltered. Interpolative methodologies, on the other hand, add new layers of assumptions – such as the homogeneity of a variable's distribution within a geographic entity – to calculate interpolation weights between the original grids and the new regular grid.

In terms of implementation, interpolative compatibility requires operations between geometries (polygons, points) and, depending on the use of auxiliary data, manual checks of the resulting features, as is the case in Rodríguez's method for the Argentine census. IPEA's implementation of MCAs is based on simple relational algebra operations but requires the additional step of creating compatibility tables.

In this work, we seek to enable the application of the aggregative heuristic of MCAs at the census tract scale. This heuristic offers greater conceptual simplicity, circumventing the need for the additional assumptions inherent in an interpolative approach. In other words, our approach avoids methodology-induced errors in data aggregation, allowing clean demographic rates at local scale. Beyond making comparability achievable with little human effort, it makes it possible to assess comparability for finely-traced tracts such as those from the US Census or the resulting layers of Rodríguez work in Argentina. Furthermore, unlike the Argentine solution, we provide an open-source implementation. In doing so, we facilitate reproduction, reuse and future improvements.

3 Brazilian Census Background

Modern census data in Brazil is acquired exclusively through door-to-door in-person interviews. IBGE adopted census tracts for the first time in the 1940 Census, with the original purpose of organizing data collection by enumerators. An urban census tract represents an area that aggregates a set of households that can be surveyed by the same enumerator. Starting with the 2000 Census, the geographic boundary layer of tracts began to be publicized in digital format, along with data aggregated by census tract, enabling the broad use of census data in geospatial databases and demographic analytics at the local scale (see Figure 3). Since then, beyond its original operational function, the census tract has also acquired an analytical function for several areas of knowledge.

In recent censuses, the adoption and improvement of digital mapping technologies improved tract tracing but made census tract geometries incompatible across censuses (Figure 4). For the 2022 Census, IBGE developed a decentralized redesign system of the census grid using high-resolution imagery, which allowed for more precise tract boundaries but further accentuated the differences compared to older census grids.

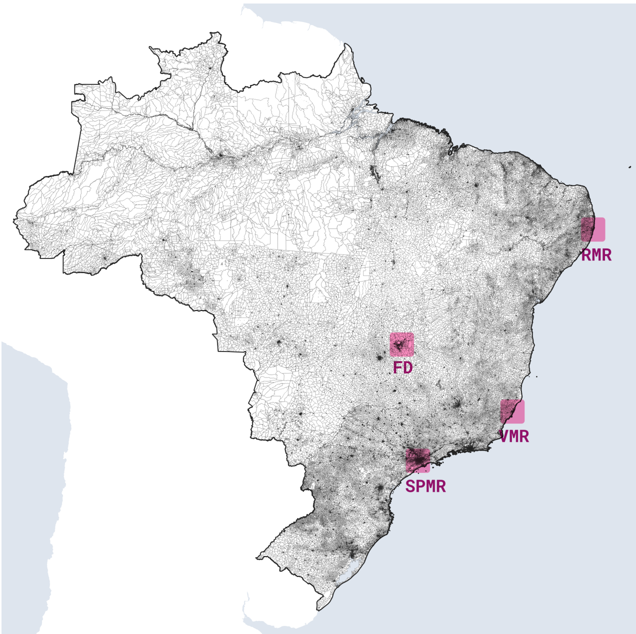

Figure 3. Brazilian census tracts for the 2022 Census. Study cases chosen for our work are pointed in red: São Paulo Metropolitan Region (SPMR), Recife Metropolitan Region (RMR), Vitória Metropolitan Region (VMR) and the Federal District (FD).

The redesign of census tracts has been a recurring topic in IBGE's technical notes [IBGE, 1980], and it is also present in publications of the US Census [US Census Bureau, 2024b] and Argentina's census [Rodriguez and de Grande, 2024]. The manuals point to a set of logical operations resulting from the redesign of a census tract, which we systematize as:

- **Split**: when a census tract is divided into several $(1 : n, n > 1)$
- **Merge**: when more than one census tract is joined $(n : 1, n > 1)$
- **Maintenance**: when the tract is maintained $(1 : 1)$
- **Reconfiguration**: when there is no single unambiguous relation of split, merge, or maintenance $(m : n$, with $m, n > 1)$

Figure 5 illustrates how these operations can be summarized into MCAs using a comparability graph. In the IBGE censuses, the technical manuals also introduce another variable related to redesign, associated with tract type (urban or rural). In this article, we disregard qualitative aspects of the census tract. We consider the census tract solely as a geographic entity composed of a polygon and an identifier code (id).

The id of a census tract is a structured numeric string, as shown in Figure 6, where the first digits determine the tract's membership to higher-level geographic entities. Since the coding of parent entities is usually consistent across censuses, the redesign of census tracts generates correspondences of the form $r = (I, J)$, where I and J are sets of census tracts from censuses C_I and C_J, respectively, such that every tract $s \in (I \cup J)$ in the same relation r is a member of the same higher-level geographic entity.

Isolating tract ids from their polygons, a redesign operation between censuses produces the following possible id relations:

- **Exclusion**: an id ceases to exist;
- **Creation**: an id is created;
- **Maintenance**: an id is maintained and identifies the same census tract;
- **Disassociation**: an id exists in both censuses but refers to different geographic entities. This is a particular case of two unrelated tracts that underwent exclusion and creation.

Census tracts are validated by prior recognition of localities by enumerators, who may modify them according to operational needs. Situations of polygon reconfiguration or id disassociation are not indicated as best practices in the technical manuals, but they do occur. The maintenance of tracts, in turn, assumes that both the polygon and id of tracts across two censuses represent the exact same geographic entity. From a computational perspective, identifying compatible groups of census tracts across two censuses can be divided into two fundamental problems: (i) identifying maintenance cases, ensuring that they do not include tracts with id disassociation, and (ii) correctly identifying the subsets of tracts involved in each redesign operation.

4 Methodology

Based on the literature review, we propose an automated methodology for the aggregative harmonization of census tracts, using their polygons and ids to generate Census Tract MCAs and comparability tables. The algorithm that implements this methodology contains three subroutines – identification, division identification, and forced overlap – that, combined, construct an undirected graph which we call the *comparability graph*. A more detailed description in pseudocode is available at section 7, along with a visual description of our implementation.

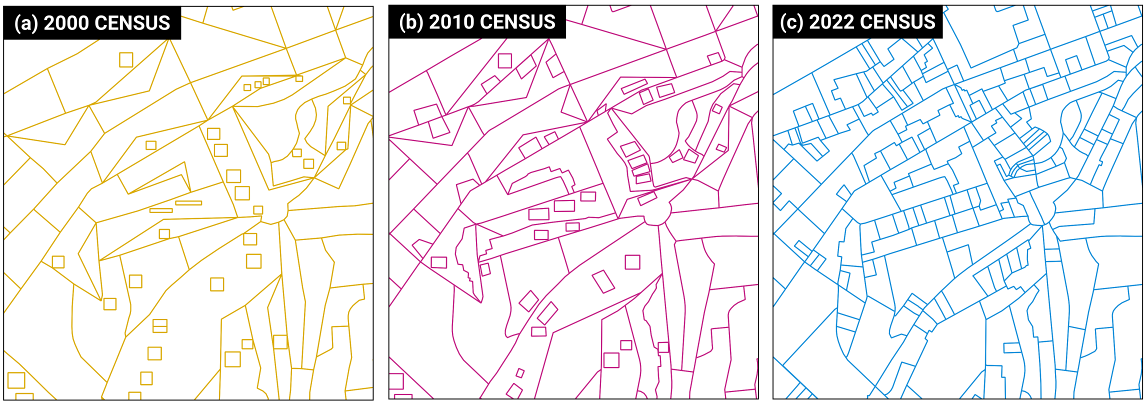

Figure 4. Census tracts of the same geographic extent in the Bela Vista neighborhood, central São Paulo, in the 2000, 2010, and 2022 censuses. Figure from [Mendonça and Kon, 2025].

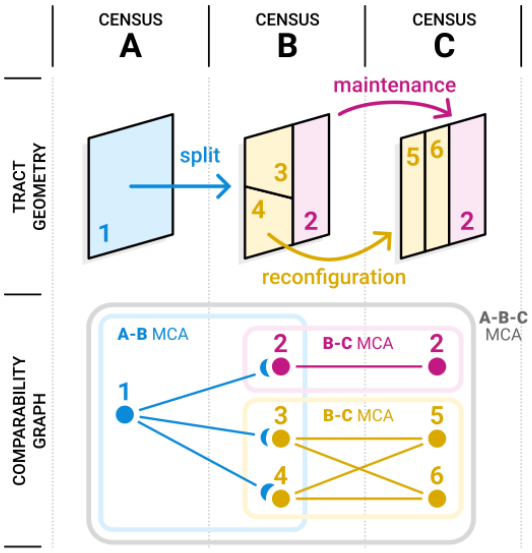

Figure 5. Conceptual diagram of tract redesign operations across census A, B and C, and their abstraction into a comparability graph. Relationship between pairs of censuses form A-B and B-C census MCAs, and the whole period forms a single A-B-C MCA.

The order in which these procedures are applied should preserve as much as possible the original definition of ids and polygon shapes, creating a "funnel-like" flow that starts without geometric manipulations, used to identify maintenance operations. The following procedures employ increasingly complex geometric manipulations — polygon intersections for division identification, and then buffer-based intersections for forced overlap — but only among subsets of tracts defined by the logical structure of the redesign operation. Tracts involved in maintenance operations, for instance, should not be included in division identification.

At the beginning of the algorithm's execution, the comparability graph is empty, containing only vertices that represent census tracts from two censuses. During execution, relationships between tracts that do not belong to the same higher-level geographic entity are considered inconsistent and, therefore, discarded.

All the subroutines of the $MakeComparable$ algorithm are implemented as methods for an inherited `networkx.Graph` class named `comparability_graph`, as available at `https://github.com/peredrozende/compat_censos/blob/main/utils.py`.

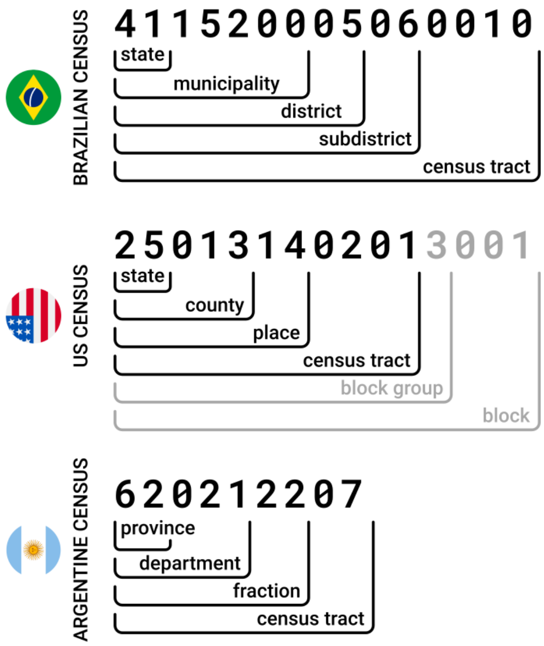

Figure 6. Structure of the census tract ids.

4.1 Identifying Maintenance Operations

The goal of the maintenance step is to uniquely identify pairs of tracts t_I, t_J that represent exactly the same geographic area in both censuses. To assert this condition, we apply several criteria. We first verify the absolute preservation of location, that is, whether two tracts with the same id are not geographically distant. We then check the relative location, that is, whether it is possible to identify neighboring tracts that were also preserved between censuses. Finally, we evaluate whether the polygons were preserved, that is, whether t_I and t_J did not undergo geometric redesign to represent a different geographic extent. Once these criteria are met, we add the edge $t_I t_J$ to G, representing a 1:1 redesign operation (Figure 13, item 1). The tracts classified as such are spared by default from the subsequent steps, since they should not receive new correspondence links.

There are cases of distinct tracts that represent the same geographic extent but do not share the same id. Although they are, conceptually, a maintenance case, it is not possible to establish an unambiguous relationship. The correspondence between these tracts will be identified in later steps.

4.2 Identifying Splitting Operations

In this stage, we aim to identify relationships between census tracts whose codes were created, deleted, or disassociated, based on the spatial overlap of their polygons, to detect cases where a tract was split. Tract splitting implies a redesign operation of type $1 : n$, in which the area of the resulting tracts is fully contained by the original tract. We also allow occurrences where two disassociated tracts with the same id overlap spatially, creating a split operation with $n = 1$. In such cases, we label the split operation as *disassociated maintenance* and, by default, do not spare these tracts from other stages. Nevertheless, these cases are accounted as maintenance operations when assessing the results.

There might be significant differences in the precision of polygon tracing between two census tract boundary layers. Because of that, it is necessary to reduce the sensitivity to spatial overlap, as we cannot expect split tracts to be fully contained within their predecessors. Figure 13, item 2, illustrates this procedure. The first criterion we used is the removal of residual polygons based on the area-to-perimeter (A/P) ratio. Next, we compute the ratio between the area of each intersected polygon and the area of the tract from the most recent census to which it originally belongs. For intersection pairs where this ratio is above a parameterized threshold L_{min} – that is, when the intersected area represents at least a L_{min} fraction of the area of a tract from the newer census – we add an edge between the respective vertices of these tracts in the comparability graph.

4.3 Forced Overlay

The filters applied in the previous steps exclude exceptional cases of redesign operations, usually associated with reconfiguration operations ($m : n$). In these cases, we perform a forced spatial overlay. We select the isolated nodes from the comparability graph and apply a negative buffer parameterized to retract the polygon boundaries by B meters. Then, we group the polygons by census and intersect them with the complete census tract boundary layer of the complementary census, removing residual polygons using the same criterion as before (the A/P ratio). This process is repeated symmetrically for both censuses. The remaining intersection pairs are added to the graph as forced correspondence edges (Figure 14, items 3a and 3b).

4.4 Encoding of MCAs

The steps described above yield a graph whose components correspond to bipartite sets of census tracts from the two censuses. From the members of each component, it is possible to create comparability tables with relations $r_C = (t, K)$, where K is the identifier of a component and t is a census tract from a census C such that $t \in vertices(K)$. Two tables are created, one for each census. We use the table to generate a new polygonal geographic layer of census tract MCAs by dissolving the polygons of tracts belonging to the same component K (Figure 14, item 4). The components are classified according to the redesign operation they represent ($1 : 1$, $1 : n$, $n : 1$, or $m : n$).

5 Implementation and Testing

We implemented the comparability algorithm in Python 3, as described in Figures 13 and 14. Each step of the algorithm was implemented as a method of an inherited graph object from `networkx` library. We used `geopandas` to read and handle the spatial databases from both censuses, and `shapely` for geometric operations on polygons. To standardize units and reduce the use of computational resources in the following steps, census tract layers are first pre-processed to identify neighbors and perform geographic reprojection into a UTM metric coordinate system.

To identify absolute location preservation in maintenance operations, a pair of tracts is considered to be in maintenance only if the distance between their centroids is at most half the sum of the diagonals of their bounding boxes. This ensures that the smaller the tract (that is, the more detailed the local tracing of the census tract boundary layer), the lower the tolerance for spatial displacements. The criterion of relative preservation was adopted by verifying the adjacent polygons of each tract, requiring that at least one neighboring tract be maintained.

We ran tests for six regions in three national censuses:

- **Brazilian Census**: São Paulo Metropolitan Region (SPMR), Recife Metropolitan Region (RMR), Vitória Metropolitan Region (VMR) and the Federal District (FD), all for 2010-2022 censuses, with data available at IBGE GeoFTP service;
- **Argentine Census**: Buenos Aires Province and City, for 2010-2022 censuses, with data available at Conicet repositories;
- **US Census**: New York City (NYC), for 2010-2020 censuses, with data available at US Census Bureau FTP service.

The higher-level geographic entity was set based on the existence of a higher-level polygon redesign during the target period. The subdistrict (11 digits) was adopted for SPMR and RMRec, the district (9 digits) for DF and RMVitoria, the place (7 digits) for NYC, and the province (2 digits) for Buenos Aires.

To assess test results, a different gold standard source was applied for each national census (Figure 14, item (5) on edge assessment). For the Brazilian Census, IBGE recently released a comparability table for the 2010 and 2022 censuses. Although a visual check of the relationships in this table revealed some inconsistent census tract pairs, this is the only publicly available reference. For the Argentine Census, we rely on Rodriguez and de Grande [2024] tracing adjustments to create a comparability table via a simple spatial overlay of tract polygons. For the US Census, we use the relationship table published by the US Census Bureau.

Edges in the comparability graph are classified as false and true positives, and missing edges are accounted for as false negatives. We then calculate recall, precision, and F1-score. For parameter sensibility test, we ran A/P ratio values ranging from 0.1 to 2.0 and L_{min} ranging from 0.1 to 1.0, using 0.1 steps. Figure 7 exemplifies the resulting graph for a small portion of the SPMR.

In addition to the regular implementation, which aims to make a pair of censuses comparable, we also implemented a proof of concept for recursive comparability using the zone polygon layers of São Paulo Metropolitan Region Origin-Destination Survey (OD Survey). The OD Survey was conducted with full SPMR coverage in 1987, 1997, 2007, 2017, and 2023, and spatial zoning files are available for every edition. A recursive implementation consists of running the comparability algorithm using the result of a previous run paired with a third polygon layer. To make such implementation possible, MCAs are encoded with the same leading set of digits of its higher-level geographic entity. Since newer polygon layers tend to have more accurate tracing, recursive comparability should preferably run retrospectively.

To assess each step of the algorithm separately, we implemented them as separate methods of a `comparability_graph` class that inherits from `networkx.Graph`. This implementation enables flexible chain operations when building the comparability graph. The default chain implementation follows the "funnel" pipeline: first identify maintenances, then identify splittings, then force overlays. To ensure that no census tracts remained disconnected, we apply a series of forced overlay operations with three different B values (-20, -10, and 0 meters). We did not discard tracts classified as maintenance in the last iteration.

5.1 Alternative Metrics

As official ground-truth comparability tables are not available for any given pair of censuses, an alternative way of assessing results is needed. We propose the following metrics for general result assessment:

- **Inconsistent maintenance**: Number of maintenance edges (assumed $1 : 1$) involved in redesign operations other than $1 : 1$;
- **Inconsistent splitting**: Number of division edges (assumed $1 : n$) involved in $m : n$ redesign operations;
- **Unconnected polygons**: Number of tracts that remained disconnected at the end of the harmonization process;
- **Persistent unconnected**: Number of census tracts that did not receive any edges for any buffer $B < 0$;
- **Large redesigns**: Number of $m : n$ redesign operations where both m and n are greater than 10.

In an ideal execution, there should be no inconsistent maintenances or unconnected polygons, and the occurrences of persistent unconnected polygons, inconsistent splittings, and large redesigns should be low.

6 Results and Discussion

Figure 8 shows the results of the proposed metrics as we varied the A/P ratio and the L_{min} parameter for each region. Regarding the A/P ratio variations, they exhibit region-dependent behavior (e.g., increasing inconsistent maintenance in SPMR and decreasing it in VMR), resulting in a weak overall correlation with the assessed metrics. Fine-traced tract layers, such as those of NYC, Buenos Aires, and FD, further weaken this correlation, as overlay operations do not produce residual geometries. Since SPMR persistent unconnected polygons were the most sensitive to A/P ratio variations, we chose to adopt a low value (0.3) as the default, though the overall results call for improvements in residual geometries filtering strategies.

The L_{min} parameter variations result in a more legible pattern. Metrics such as precision show a parabola-like curve peaking at $L_{min} \approx 0.5$, that is, at a 50% threshold for area overlays in splitting. But recall is minimum at the same threshold, requiring caution. Precision measures true positives among retrieved elements, where false positives mean unnecessary aggregation of census tracts. Recall, on the other hand, considers false negatives, which means insufficient census tract aggregation. Since lacking aggregation is more severe than oversimplification under an aggregative heuristic, recall is a preferable metric over precision. Thus, we adopted a default 0.8 value for L_{min}, where recall starts to stabilize while precision has not yet plummeted. Results point out that L_{min} also has a residual effect on fine-traced tract layers, since polygon overlay in such cases is rarely ambiguous. Nevertheless, the best precision and recall values for each region range from around 90 to 100%, indicating that the algorithm produces a successful, automated comparability graph.

Table 1 presents the final results and metrics from this implementation, and Figure 11 shows the MCAs classified by type of operation. Maintenance operations prevail over the others, and reconfiguration and union operations are uncommon, as expected. The order of steps in the algorithm corresponds to the most frequent redesign operations (maintenance, division, and reconfiguration), fulfilling the proposed "funnel-style" classification flow.

The correlograms in Figures 9 and 10 support the adoption of our five proposed metrics, as they show moderate to high negative Spearman correlations with the gold-standard recall and precision. We use Spearman correlation since some metrics vary non-linearly, as shown in Figure 8. The five proposed metrics also show moderate to high positive cross-correlation, suggesting that different types of comparability inconsistencies might not be independent. This is not unexpected, since all inconsistencies stem from the same phenomenon: manual census tract redesign. Overall, our metrics serve as a proxy for the gold standard and may thus be deployed as an alternative assessment strategy when no gold-standard data is available.

Inconsistent splittings are particularly high in SPRM, RMR, and VMR, when compared to the other cities. NYC tracts are geometrically comparable by default; Buenos Aires tracts have been made comparable by Conicet; and FD tracts have regular shapes because Brasilia is a planned city. The greatest inconsistencies occur where the regularity of polygons across censuses has been violated to either improve geographic precision or to accommodate changes in the urban grid, both of which are more common in most Brazilian cities, such as SPRM, RMR, and VMR. Nevertheless, these values illustrate how specific local knowledge is key for assessing comparability results.

The gold standard assessment is higher for fine-traced census layers. For instance, the gold-standard metrics for NYC and Buenos Aires are high and remain unaffected by parame-

Table 1. Summary of compatibility results by region

	NYC	**Buenos Aires**	**FD**	**RMR**	**SPMR**	**VMR**
Census Layers						
Tracts in 2010	2,168	23,130	4,454	4,667	30,815	2,867
Tracts in 2020/2022	2,327	27,721	5,418	7,365	47,184	3,957
Relationships (Edges)						
Total relationships	100.00% (2,426)	100.00% (27,762)	100.00% (5,676)	100.00% (8,106)	100.00% (49,844)	100.00% (4,241)
Maintenance	12.82% (311)	59.31% (16,465)	31.50% (1,788)	46.39% (3,760)	45.34% (22,600)	45.25% (1,919)
Splittings	4.04% (98)	0.10% (28)	0.37% (21)	1.09% (88)	2.87% (1,430)	1.70% (72)
Disassociated maintenance	0.00% (0)	0.68% (190)	0.02% (1)	0.06% (5)	0.01% (6)	0.09% (4)
Forced overlay ($B = -20$ m)	82.89% (2,011)	38.52% (10,694)	53.95% (3,062)	31.78% (2,576)	36.37% (18,128)	36.69% (1,556)
Forced overlay ($B = -10$ m)	0.25% (6)	1.39% (385)	14.11% (801)	20.40% (1,654)	14.28% (7,116)	15.44% (655)
Forced overlay ($B = 0$ m)	0.00% (0)	0.00% (0)	0.05% (3)	0.28% (23)	1.13% (564)	0.83% (35)
Redesign Operations (MCAs)						
Total MCAs	100.00% (2,167)	100.00% (23,110)	100.00% (4,264)	100.00% (4,140)	100.00% (27,990)	100.00% (2,611)
Maintenance ($1:1$)	93.95% (2,036)	83.78% (19,362)	90.01% (3,838)	64.25% (2,660)	66.93% (18,733)	77.25% (2,017)
Splittings ($1:n$)	6.00% (130)	16.15% (3,733)	7.95% (339)	30.89% (1,279)	28.22% (7,898)	19.57% (511)
Merging ($n:1$)	0.00% (0)	0.01% (2)	0.54% (23)	0.53% (22)	0.51% (143)	0.27% (7)
Reconfiguration ($m:n$)	0.05% (1)	0.06% (13)	1.50% (64)	4.32% (179)	4.34% (1,216)	2.91% (76)
Metrics						
Inconsistent maintenance	0.00% (0)	0.00% (0)	0.03% (2)	0.19% (10)	0.39% (140)	0.90% (28)
Inconsistent splitting	0.32% (1)	0.15% (24)	14.49% (259)	23.80% (895)	20.17% (4,559)	17.40% (334)
Unconnected polygons	0.00% (0)	0.00% (0)	0.09% (9)	0.04% (5)	0.00% (2)	0.41% (28)
Persistent unconnected	0.00% (0)	0.00% (0)	0.03% (3)	0.19% (23)	0.72% (564)	0.52% (35)
Large redesigns	0.00% (0)	0.00% (0)	0.02% (1)	0.19% (8)	0.09% (26)	0.19% (5)
Gold standard assessment						
True positives	2,328	27,744	5,473	7,423	46,173	3,878
False positives	0	0	195	651	4,512	299
False negatives	98	28	21	88	1,430	72
Recall	95.96%	99.90%	99.62%	98.83%	97.00%	98.18%
Precision	100.00%	100.00%	96.56%	91.94%	91.10%	92.84%
F1-score	97.94%	99.95%	98.06%	95.26%	93.95%	95.43%

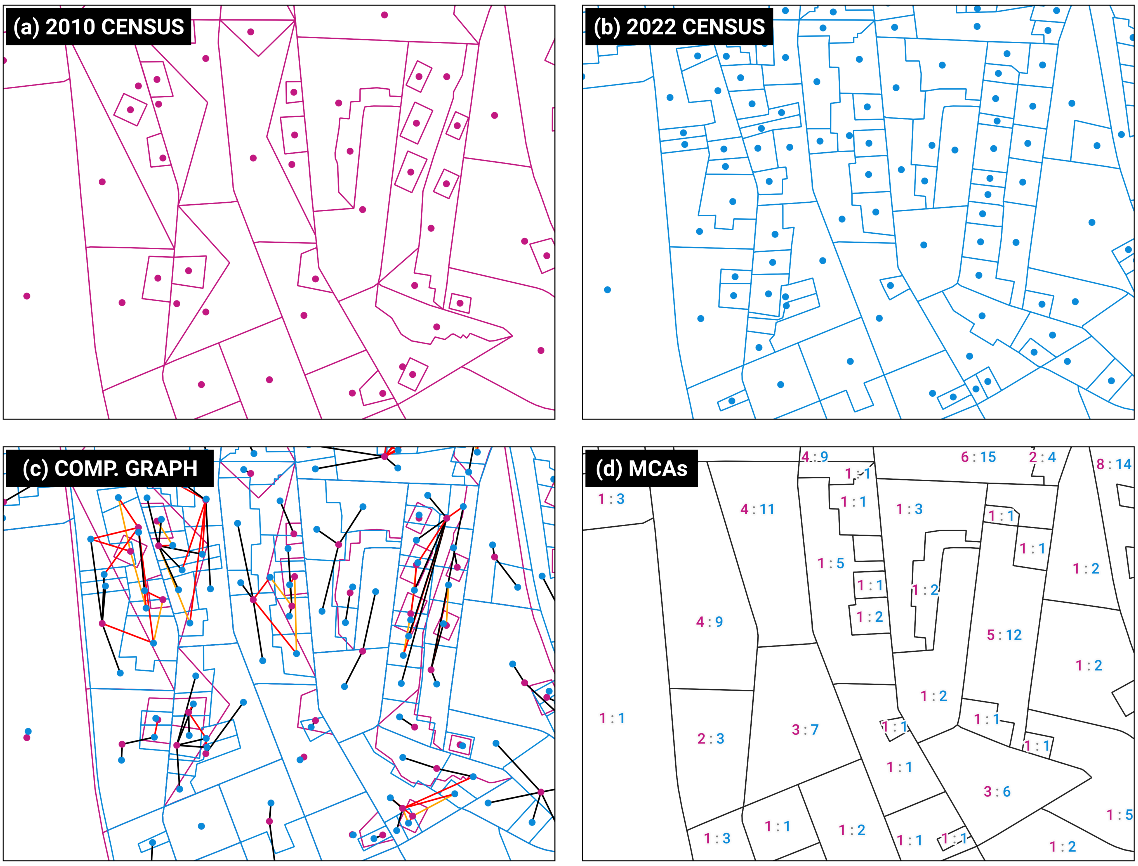

Figure 7. Example of comparability graph overlayed to the resulting MCA polygons for a portion of Bela Vista, central São Paulo. Red lines in frame c represent false positives, and yellow lines are false negatives. Ratios in frame d show the proportion of previous tracts (2010 census) to new tracts (2022 census).

ter variation, as expected due to the use of overlaying natural boundaries for polygon tracing. Nevertheless, no region has achieved a 100% score for both recall and precision, as expected, at least for NYC and Buenos Aires, suggesting that the gold standard data may be inconsistent. For the Brazilian regions, we identified inconsistent relationships between tracts in a visual assessment of the ground truth data. In other words, gold standard data does not correspond to ground truth, but it is the best publicly available reference for assessment. It nevertheless poses a mild risk to conclusion validity.

We also performed a visual inspection of the resulting graph and MCAs to understand the redesign patterns of false positives and negatives. There is a general pattern of misidentification related to illogical relationships between geometries and ids, as illustrated by Figure 12. Such cases would require new approaches for residue filtering and geometry assessment.

6.1 Threats to Validity

Regarding our study design, there are a few threats to validity to consider. All possible redesign operations are logically contained within our redesign typology ($1:1$, $1:n$, $n:1$, $m:n$), asserting construct validity. We chose large urban agglomerations from each targeted national census, assuming they pose the biggest computational challenges due to urban complexity. Many of the selected regions contain rural and semi-urban areas, but there is a risk that metrics and results might change when dealing with large extents of rural or forested regions – which were not the main focus of this work.

Also, regarding external validity, there is a risk of errors in the redesign identification being spatially correlated. As a consequence, incorrect data reaggregation would not be spatially random, violating a key quality indicator for spatial data. Nevertheless, this risk stems from the original tract redesign process, often defined by the needs and expertise of local census workers. That is, this risk originates in the dataset and is not exclusive to the aggregative heuristic.

As Figure 8 illustrates, parameters have different effects in different regions. Since correspondence errors might not be spatially random, there is a risk that the global parameters in our implementation could exacerbate the spatial concentration of errors in reaggregated data. A method for statistical assessment of this error could strengthen our conclusions, similar to those proposed for areal interpolation Cockings *et al.* [1997].

6.2 Conclusion

The adopted aggregative heuristic allows the resulting MCAs to be used in urban studies that rely on precise, unambiguous data reaggregation. This is the case in [Mendonça *et al.*, 2024], which used a preliminary version of this algorithm to classify demographic transformations in the SPMR between 2010 and 2022. Our method discards the base assumptions of interpolative heuristics, which are known sources of reaggregation errors and imprecision. In general, longitudinal studies

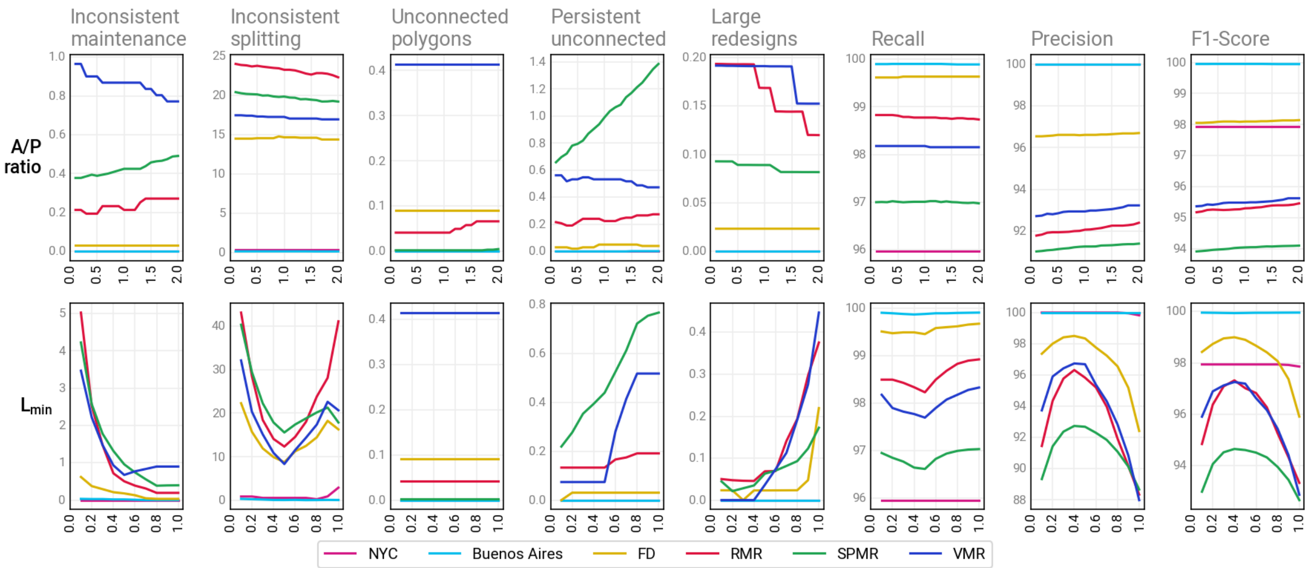

Figure 8. Variation of metrics for different values of the A/P ratio and L_{min}, with all other parameters held constant.

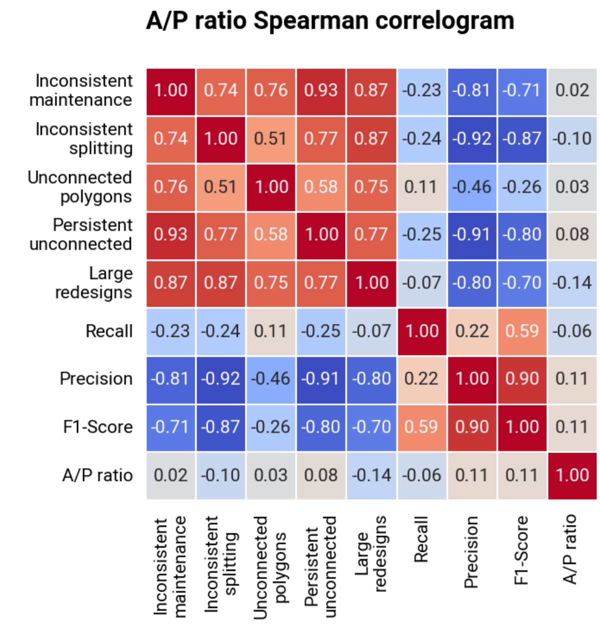

Figure 9. Spearman correlogram for A/P ratio sensibility tests on all regions.

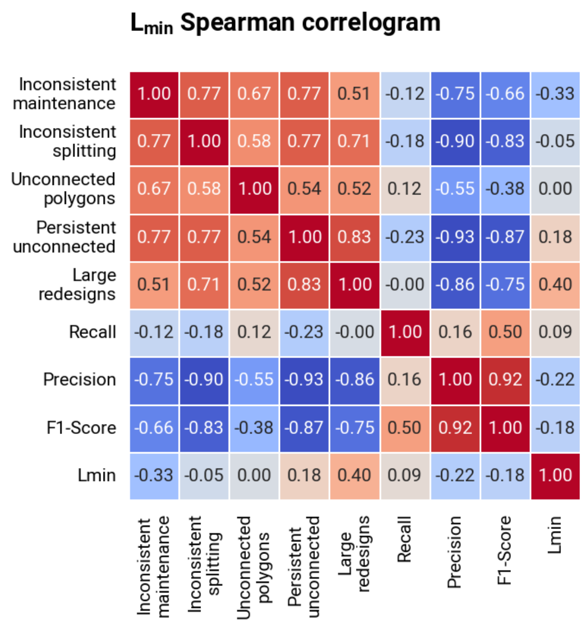

Figure 10. Spearman correlogram for L_{min} sensibility tests on all regions.

based on local-scale census series that depend on accurate spatial joins between georeferenced data – such as urban lots or points of interest – can benefit from the implementation of this methodology.

Compared to non-automated methods used in related work, the cost of performing this operation manually is significantly reduced, though human intervention can mitigate potential localized inaccuracies. Running the code on a commodity computer (Intel i7-9700F with 32GB RAM) took approximately 3–4 minutes to pre-process the census tract layers and run the comparability algorithm for the SPMR, the largest and most polygon-dense of the chosen regions. We expect this cost reduction to favor national demographic agencies and enable urban research that would otherwise be unfeasible.

We believe that our algorithm also facilitates the work of those who intend to perform a full manual correction for comparability, since most of the geometry tracing process can be replaced by simple additions and exclusions to the comparability graph. The implementation in Python also simplifies MCA creation for censuses such as the Argentine, where geometry imprecision is not a major issue, but direct comparability files are not available.

Regarding the Brazilian Census, the diverging, region-specific results shown in Figure 8 reveal the need for a region-aware approach. Due to the decentralization of redesign operations, the precision and consistency of boundary layers between censuses may vary considerably across different cities. Consequently, the choice of parameters may have significantly different effects when applied to different spatial contexts. This is also true within metropolitan regions. As shown in Figure 11, reconfiguration operations tend to concentrate in peripheral, semi-rural, and rural areas, potentially leading to information loss in analyses focused on these regions. In denser areas, such as highly vertical neighborhoods,

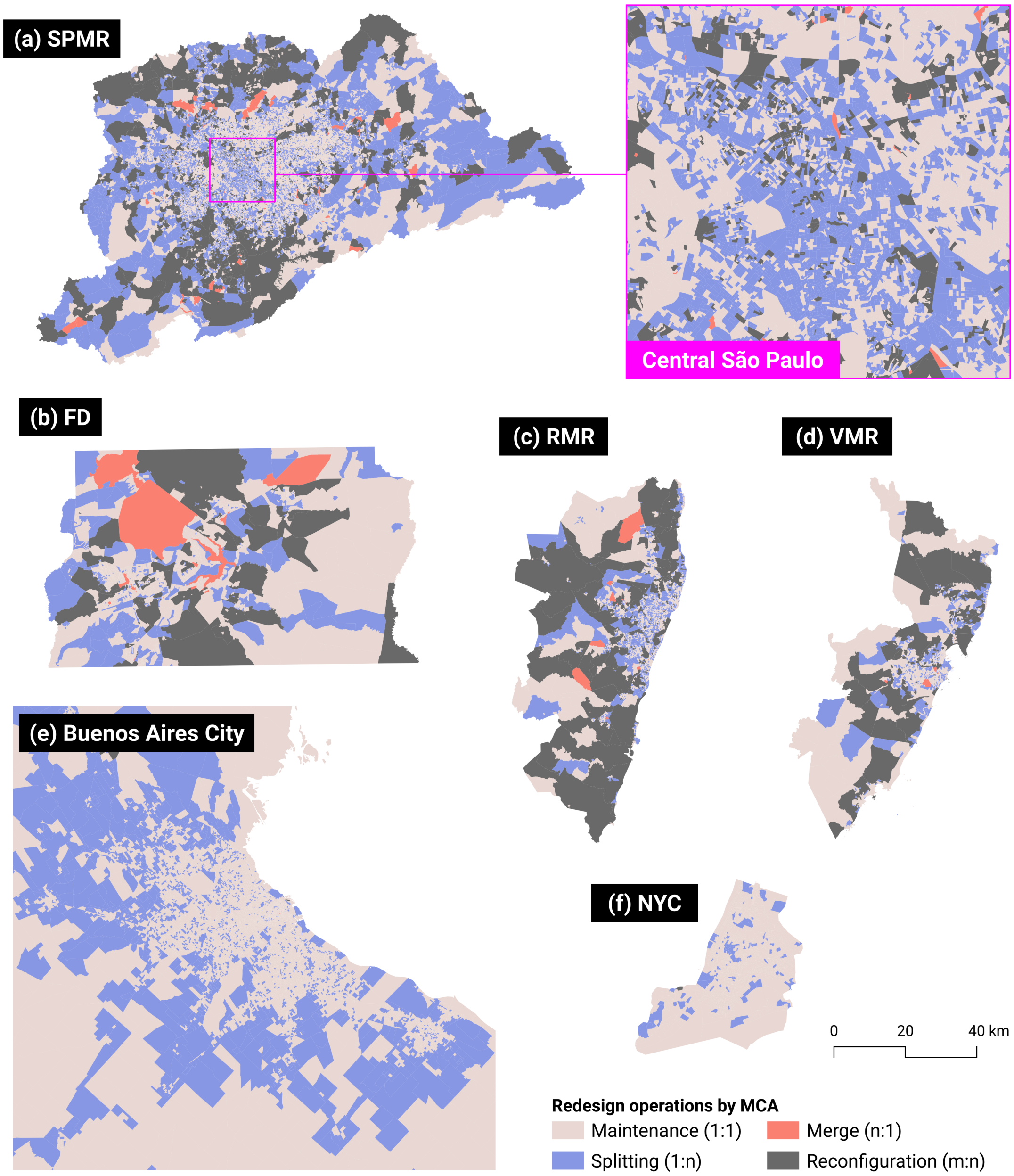

Figure 11. MCAs by identified redesign operation. Map (e) was clipped to the city of Buenos Aires due to size.

census tracts tend to be very small and more sensitive to minor boundary adjustments.

Based on the evaluation of the results, we identify two main directions for future improvements to the proposed methodology.

The first is the development of a more robust approach for selecting the parameters A/P, L_{min}, and B. The A/P ratio may be complemented or replaced by other geometric metrics such as elongation, compactness, and curvature. A systematic assessment of the combined effects of these parameters could improve the selection of their values. However, the choice of global parameters tends to balance distortions between areas with dense census tract boundaries (urban centers) and sparse ones (semi-rural and rural regions). Thus, we consider automated, region-dependent parameter selection a more promising alternative.

A second path for future enhancements follows the example of comparability approaches that employ auxiliary data. The use of satellite imagery, land-use data, identification of uninhabited areas, and road network layers could improve the comparability results by reducing the influence of uninhabited regions — that is, areas of low relevance for demographic studies.

Acknowledgements

We thank the research teams of LabCidade FAUUSP and Instituto Peregum for their early contributions to the methodological design and concept testing, and IPEA for the later exchanges in results assessment and testing design.

Funding

This research was funded by the São Paulo Research Foundation

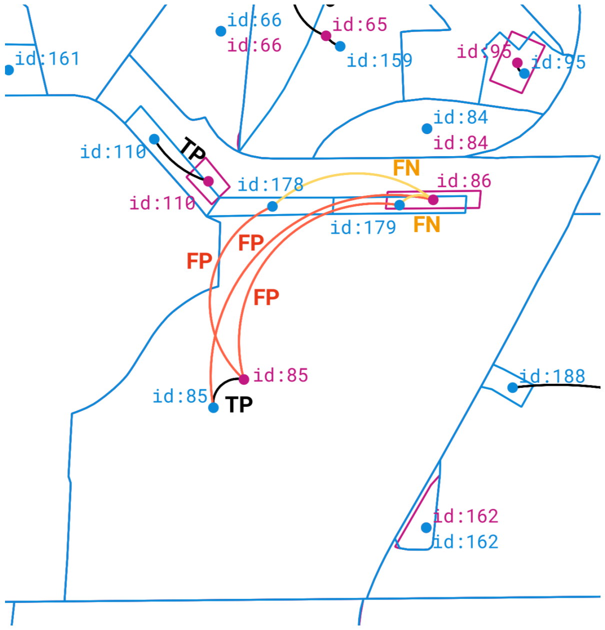

Figure 12. Example of a comparability graph component where geometry tracing has no logic connection to the true redesign operation. FN indicates false negative edges, FP is for false positives and TP for true positives. Magenta tracts are from the 2010 census and cyan tracts are from the 2022 census. Tracts 178 and 179 are formed by the splitting of tract 86, but there is no id or geometric relationship between tracts 86 and 178 after the operation, resulting in redesign misidentification. The imprecision is aggravated by the thin shape of tracts 86, 178, and 179, which tend to generate geometries mistaken for A/P ratio residues after overlay. Frame located at Mooca industrial park, in SPMR.

(FAPESP), under grant #2024/12863-8. This grant is associated with research project grants #2023/12851-7 (*Evictions Observatory: interdisciplinary dimensions of risk in times of climate change and housing crisis*) and 2023/00811-0 (*EcoSustain: computer and data science for the environment*).

Authors' Contributions

PM led the study's conception, experimental work, and manuscript preparation. FK supervised the research and provided critical oversight of the study design and writing.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets and source code produced and analyzed during the current study are freely available at https://github.com/peredrozende/compat_censos.

Abbreviations

MAUP: Modifiable Areal Unit Problem
IBGE: Brazilian Institute of Geography and Statistics
Conicet: National Scientific and Technical Research Council of Argentina
MCA: Minimum Comparable Area
IPEA: Brazilian Institute of Applied Economic Research
SPMR: São Paulo Metropolitan Region
RMR: Recife Metropolitan Region
VMR: Vitória Metropolitan Region
FD: Federal District
NYC: New York City

References

Allen, J. and Taylor, Z. (2018). A new tool for neighbourhood change research: The Canadian Longitudinal Census Tract Database, 1971–2016. *Canadian Geographies / Géographies canadiennes*, 62(4):575–588. DOI: 10.1111/cag.12467.

Barbosa, R. J. (2014). Comparabilidade das informações disponíveis nos Censos (1960-2010) e PNADs (1976, 1985, 1995 e 2005). *São Paulo: Centro de Estudos da Metrópole*. Available at:`https://centrodametropole.fflch.usp.br/en/file/16809/download?token=kdwWiMmW`.

Breen, C. F. and Feehan, D. M. (2025). New Data Sources for Demographic Research. *Population and Development Review*, 51(1):539–573. DOI: 10.1111/padr.12671.

Buzzelli, M. (2020). Modifiable Areal Unit Problem. In *International Encyclopedia of Human Geography*, pages 169–173. Elsevier. DOI: 10.1016/B978-0-08-102295-5.10406-8.

Cockings, S., Fisher, P. F., and Longford, M. (1997). Parameterization and Visualization of the Errors in Areal Interpolation. *Geographical Analysis*, 29(4):314–328. DOI: 10.1111/j.1538-4632.1997.tb00967.x.

Comber, A. and Zeng, W. (2019). Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. *Geography Compass*, 13(10):e12465. DOI: 10.1111/gec3.12465.

Dias, F. and Silver, D. (2021). Neighborhood Dynamics with Unharmonized Longitudinal Data. *Geographical Analysis*, 53(2):170–191. DOI: 10.1111/gean.12224.

Flowerdew, R. and Openshaw, S. (1987). *A review of the problems of transferring data from one set of areal units to another incompatible set*. Northern Regional Research Laboratory. Book.

Gehlke, C. E. and Biehl, K. (1934). Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association*, 29(185A):169–170. DOI: 10.1080/01621459.1934.10506247.

GeoLitycs (2007). CensusCD Neighborhood Change-Database (NCDB) 1970-2000US Census Tract Data - User Guide. Technical report, GeoLytics, East Brunswick. Available at:`https://geolytics.com/media/pdf/NCDB-final-LF-version.pdf`.

Goodchild, M. F., Anselin, L., and Deichmann, U. (1993). A Framework for the Areal Interpolation of Socioeconomic Data. *Environment and Planning A: Economy and Space*, 25(3):383–397. DOI: 10.1068/a250383.

Gregory, I. (2002). The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment* *and Urban Systems*, 26(4):293–314. DOI: 10.1016/S0198-9715(01)00013-8.

Hirye, M. C. d. M., Amaral, S., Monteiro, A. M. V., and Alves, D. S. (2016). Interpolação de dados censitários para análise da ocupação intraurbana em Altamira (PA) em 2000 e 2010. *Revista Brasileira de Cartografia*, 68(8):2016. DOI: 10.14393/rbcv68n8-44381.

IBGE (1980). *IX Recenseamento Geral - Instruções para delimitação dos setores censitários*. IBGE, 1 edition. Available at:https://biblioteca.ibge.gov.br/visualizacao/livros/liv17621.pdf. Last access on: 2025-03-02.

Junior, J. U. P., Louro, T. V., Assis, L. B. M. d., and Brito, P. L. (2025). Measuring land-use mix with address-level census data. DOI: 10.31224/5975.

Jurjevich, J. R., Meehan, K., Chun, N. M. J. W., and Schrock, G. (2025). Advancing methods for comparative urban research: A city-centric protocol and longitudinal dataset for US metropolitan statistical areas. *PLOS ONE*, 20(3):e0316750. DOI: 10.1371/journal.pone.0316750.

Lobo, M. A. A. (2009). Método para compatibilizar setores censitários urbanos de 1991 e 2000 aplicado ao estudo da dinâmica populacional da região metropolitana de Belém (PA). *Revista Brasileira de Gestão Urbana*, 1(1):71–84. Available at:https://periodicos.pucpr.br/Urbe/article/view/4264.

Logan, J. R., Xu, Z., and Stults, B. J. (2014). Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database. *The Professional Geographer*, 66(3):412–420. DOI: 10.1080/00330124.2014.905156.

Markley, S. N., Holloway, S. R., Hafley, T. J., and Hauer, M. E. (2022). Housing unit and urbanization estimates for the continental U.S. in consistent tract boundaries, 1940–2019. *Scientific Data*, 9(1):82. DOI: 10.1038/s41597-022-01184-x.

Mendonça, P. H. and Kon, F. (2025). Compatibilização de setores censitários baseada em grafos para análise histórica de processos urbanos em escala local. In *Workshop de Computação Urbana (CoUrb)*, pages 29–42. SBC. DOI: 10.5753/courb.2025.7950.

Mendonça, P. H. R., Lima, P. H. B. M., Costa, D. F., Canan, H. G., Benedusi, A. A., Giacomini, L. A., Azzolini, G. S., Andrade, L. G. E., Stroher, L. E., Santoro, P. F., and Rolnik, R. (2024). A expansão – com desadensamento – da Região Metropolitana de São Paulo entre 2010 e 2022. *E-metropolis*, 15. Available at:https://revistas.ufrj.br/index.php/emetropolis/article/view/63730.

Nguyen, T., Bernard, A., Lee, R., Wilson, T., and Argent, N. (2023). Do Co-Ethnic Neighbourhoods Affect the Labour Market Outcomes of Immigrants? Longitudinal Evidence from Australia. *Applied Spatial Analysis and Policy*, 16(2):831–850. DOI: 10.1007/s12061-023-09505-2.

Norman, P., Colbert, J., and Exeter, D. J. (2023). Linking Individuals to Areas: Protecting Confidentiality While Preserving Research Utility. *Spatial Demography*, 11(3):10. DOI: 10.1007/s40980-023-00121-9.

Openshaw, S. (1979). A million or so correlated coefficients: three experiment on the modifiable areal unit problem. *Statistical applications in the spatial sciences*. Available at:https://cir.nii.ac.jp/crid/1571135650655620224.

Pérez, V. and Pavía, J. M. (2024). Automating the transfer of data between census sections and postal codes areas over time. an application to spain. *Journal of Regional Research*. Available at:https://investigacionesregionales.org/wp-content/uploads/sites/3/2024/10/101745-Perez-Online.pdf.

Reis, E., Pimentel, M., Alvarenga, A. I., and Santos, M. d. C. H. (2008). Áreas mínimas comparáveis para os períodos intercensitários de 1872 a 2000. *Rio de Janeiro: Ipea/Dimac*, 40. Available at:https://www.ufmg.br/rededemuseus/crch/simposio/REIS_EUSTAQUIO_JOSE_R_ET_AL.pdf.

Reis, I. A. (2013). Compatibilização de populações entre malhas censitárias diferentes com o uso de imagens de sensores orbitais. *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR - INPE*. Available at:http://bibdigital.sid.inpe.br/dpi.inpe.br/marte2/2013/05.28.23.11.10.

Rodríguez, G. M. (2021). Comparabilidad retrospectiva en la cartografía censal digital del INDEC. estado actual, avances y desafíos en Argentina y la ciudad de Buenos Aires. *Población de Buenos Aires*, 18(30):22–33. Available at:https://ri.conicet.gov.ar/handle/11336/169850.

Rodriguez, G. M. and de Grande, P. E. (2024). Base cartográfica de radios del censo argentino 2022: Primera versión revisada y corregida para uso en Sistemas de Información Geográfica. Available at:https://ri.conicet.gov.ar/handle/11336/238198. Last access on: 2025-03-02.

Ruther, M., Leyk, S., and Buttenfield, B. P. (2015). Comparing the effects of an nlcd-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. *GIScience & Remote Sensing*, 52(2):158–178. DOI: 10.1080/15481603.2015.1018856.

Schroeder, J., Van Riper, D., Manson, S., Knowles, K., Kugler, T., Roberts, F., and Ruggles, S. (2025). IPUMS National Historical Geographic Information System. DOI: 10.18128/D050.V20.0.

Schroeder, J. P. (2007). Target-Density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data. *Geographical Analysis*, 39(3):311–335. DOI: 10.1111/j.1538-4632.2007.00706.x.

US Census Bureau (2024a). History of census tracts and blocks. Available at:https://www.census.gov/about/history/historical-censuses-and-surveys/census-programs-surveys/geography/tracts-and-block-numbering-areas.html. Last access on: 2025-03-02.

US Census Bureau (2024b). Relationship files. Available at:https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.2020.html#tract. Last access on: 2025-03-02.

Yamaguchi, F. Y. (2017). Avaliação de dados de grades regulares para fins estatísticos. Master's thesis, Escola

Politécnica da Universidade Federal da Bahia. Available at:`https://repositorio.ufba.br/handle/ri/24036`.

7 Appendix

In this section, we present additional details of the comparability algorithm and its implementation in our study.

7.1 Comparability Algorithm

Let us consider two census layers M_I from census I, and M_J from census J, both composed of tracts t. The aggregative comparability procedure consists of constructing an undirected graph $G = (V, E)$ as follows:

$$V_I = \{t \mid t \in M_I\}$$
$$V_J = \{t \mid t \in M_J\}$$
$$V = V_I \cup V_J$$
$$E = MakeComparable(V_I, V_J)$$

Where $MakeComparable$ is a function implementing Algorithm 1. The notation $geom(t)$ in the algorithm represents the polygon of census tract t, and $id(t)$ represents the identifier code of tract t. The notations $geom(T)$ and $id(t)$ represent the set of polygons and identifier codes for a set of tracts T. Each component of the resulting graph represents a redesign operation, with a set of tracts from each census involved in that operation.

7.2 Python Implementation

Figures 13 and 14 illustrate the comparability pipeline implemented in Python for this study. Additional Jupyter Notebooks with examples are available at `https://github.com/peredrozende/compat_censos`.

Algorithm 1: MakeComparable

[p]
Input: V_I, V_J
Output: Set E of edges of type $t_I t_J,\ t_I \in V_I,\ t_J \in V_J$
$E, C \leftarrow \emptyset$;
`// Step 1: Identify maintenance operations`
for *each* $t_I \in V_I$ **do**
  if $id(t_I) \in id(V_J)$ **then**
    $t_J \leftarrow$ tract in V_J with $id = id(t_I)$;
    if *t_I and t_J are in the same absolute and relative location and preserve their shapes* **then**
      $E \leftarrow E \cup \{t_I t_J\}$;
      $C \leftarrow C \cup \{t_I, t_J\}$;
    end
  end
end
for *each* $t_J \in V_J\ :\ t_J \notin C$ **do**
  `// Identify disassociated maintenance`
  if *$geom(t_J)$ intersects a $geom(t_I)$ of some $t_I \in V_I$ such that $id(t_J) = id(t_I)$* **then**
    $E \leftarrow E \cup \{t_I t_J\}$;
    $C \leftarrow C \cup \{t_I, t_J\}$;
  end
  `// Step 2: Identify splitting operations`
  if *$geom(t_J)$ has a portion of its area larger than L_{min} in $geom(t_I)$ of some $t_I \in V_I$ and the overlay intersection $geom(t_I) \cap geom(t_J)$ is not residual* **then**
    $E \leftarrow E \cup \{t_I t_J\}$;
    $C \leftarrow C \cup \{t_I, t_J\}$;
  end
end
`// Step 3: Forced overlay`
`// Redefine polygons of` V_I `and` V_J `with buffer`
$geom(V_I) \leftarrow \{buffer(geom(t), B) \mid t \in V_I\}$;
$geom(V_J) \leftarrow \{buffer(geom(t), B) \mid t \in V_J\}$;
$N \leftarrow \{t \mid t \in V_I \cup V_J,\ t \notin C\}$;
`// Step 3a: Force overlay from` I `to` J
for *each* $t_I \in (V_I \cap N)$ **do**
  if *$geom(t_I)$ intersects $geom(t_J)$ of some $t_J \in V_J$ and $geom(t_I) \cap geom(t_J)$ is not residual* **then**
    $E \leftarrow E \cup \{t_I t_J\}$
  end
end
`// Step 3b: Force overlay from` J `to` I
for *each* $t_J \in (V_J \cap N)$ **do**
  if *$geom(t_J)$ intersects $geom(t_I)$ of some $t_I \in V_I$ and $geom(t_J) \cap geom(t_I)$ is not residual* **then**
    $E \leftarrow E \cup \{t_I t_J\}$
  end
end
return E;

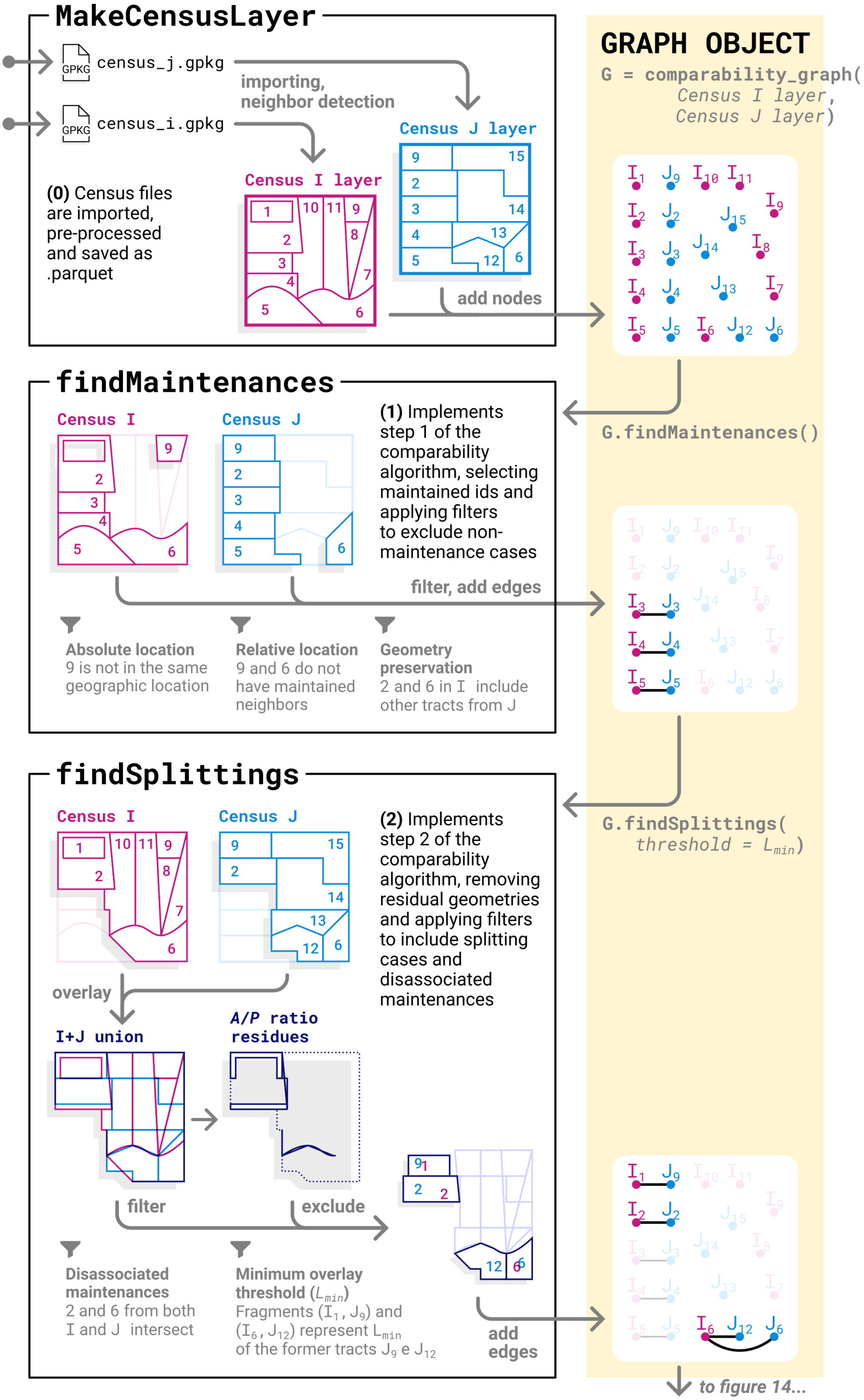

Figure 13. Diagram of the first two steps of the algorithm building the comparability graph, labeling classes and methods as implemented in Python.

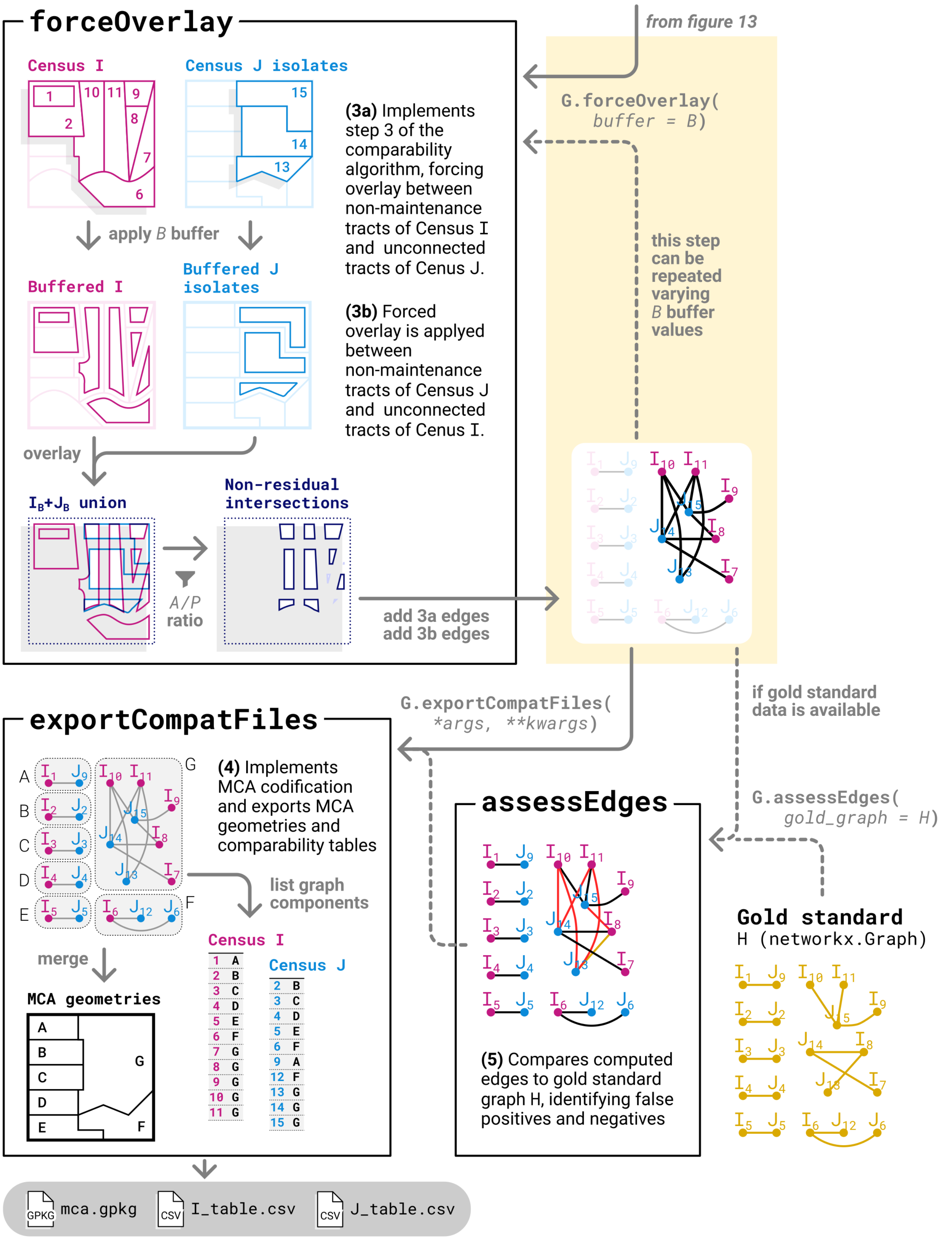

Figure 14. Diagram of the final steps of the algorithm building the comparability graph, labeling classes and methods as implemented in Python.