Measuring land-use mix with address-level census data

Jorge Ubirajara Pedreira Junior[1]* 0000-0002-8243-5395
Thiago Vinícius Louro[2,3] 0000-0002-1118-3119
Lucas Brandão Monteiro de Assis[2] 0000-0002-8338-8535
Patrícia Lustosa Brito[1] 0000-0002-3987-7331

[1] Department of Transportation Engineering and Geodesy, Polytechnic School of the Federal University of Bahia, Brazil.
[2] Department of Transportation Engineering, São Carlos School of Engineering, University of São Paulo, Brazil.
[3] Department of Civil Engineering, University of Twente, Netherlands.

* corresponding author: jorge.ubirajara@ufba.br

Abstract

This paper introduces a data-driven framework to evaluate mixed land use in Brazilian cities using the National Address File for Statistical Purposes (CNEFE), an address-level dataset from the 2022 census that records over 110 million geocoded establishments. We treat CNEFE records as point-based observations of functional use and aggregate them into H3 hexagonal grids to compute local residential and non-residential shares. Building on this representation, we calculate two standard indices – the Entropy Index (EI) and the Herfindahl–Hirschman Index (HHI) – and propose two directional extensions: an adapted HHI (aHHI), which maps functional dominance to a $[-1, 1]$ scale, and the Bidirectional Global-centered Balance Index (BGBI), which measures deviations from the citywide residential proportion. The method is implemented in an open R workflow that automates data retrieval, preprocessing, and index computation for any municipality. Applying it to six major metropolitan areas at two H3 resolutions, we show that EI and HHI behave as expected but are blind to the direction of homogeneity, whereas aHHI and BGBI clearly distinguish predominantly residential from predominantly non-residential cells and highlight areas that match the global functional balance. Cross-scale comparisons document non-trivial sensitivity to grid resolution, reinforcing the need to report and test scale in mixed land-use studies.

Keywords: land use; urban planning; spatial analysis

1. Background

Land-use mix (LUM) is a central principle in contemporary urban planning and sustainability research, defined as the spatial integration of residential, commercial, institutional, and recreational functions within the same area or in close proximity. It underpins smart growth, new urbanism, and transit-oriented development, promoting compact and efficient urban forms (Jiao & Fu, 2020). By countering the limitations of single-use zoning, such as isolation, lack of vitality, and increased travel distances, LUM enhances space utilization, accessibility, and local economic activity (Raman & Roy, 2019).

Evidence consistently links LUM to improved sustainability outcomes. Compact, mixed environments reduce horizontal expansion, infrastructure redundancy, and long-distance commuting, enabling more efficient land consumption and revitalization of underused areas (Bahadure & Kotharkar, 2015; Chen et al., 2022). Environmental benefits follow from lower greenhouse gas emissions and greater energy efficiency, as the proximity of housing and employment supports walking, cycling, and transit use (Bordoloi et al., 2013). Studies indicate that land-use diversity is associated with reduced CO_2 emissions, although excessive mixing may generate diminishing returns, producing a U-shaped relationship between entropy-based diversity and emissions (Q. Li et al., 2022).

LUM also contributes to urban vitality and social cohesion by supporting continuous street activity, strengthening local interaction, and enhancing economic and cultural resilience (J. Li et al., 2024; Nabil & Eldayem, 2015). Economically, mixed-use areas attract higher property values and stronger business performance, though they may also accelerate gentrification and displacement if unmanaged (Koster & Rouwendal, 2010). Safety benefits arise from increased pedestrian presence and "eyes on the street," although incompatible use combinations can produce opposite effects (Wo, 2019; Zahnow, 2018).

Despite its recognized importance, measuring LUM remains challenging due to data limitations and methodological constraints. Many cities lack accessible, high-resolution land-use datasets, and integrating heterogeneous sources, such as cadastral records, imagery, and POIs, requires intensive preprocessing (J. Li et al., 2024). Additionally, numerous indices have been proposed, including the Entropy Index, the Herfindahl–Hirschman Index, the Dissimilarity Index, the Gini Index, the Simpson's Diversity Index, the Shannon Index, the Mixed-Use Index, the Land Use Interaction Index, and the Multidimensional Mixed-Degree Index. For a comprehensive discussion and classification of these indices, along with an examination of their relative advantages, readers are referred to Song et al. (2013).

Considering this background, the present study introduces a methodological framework for computing mixed land-use indices using a national-level dataset of georeferenced addresses collected by the Brazilian demographic census — the National Address File for Statistical Purposes (in Portuguese, *Cadastro Nacional de Endereços para Fins Estatísticos*, CNEFE). First established in 2005, the CNEFE represents the most comprehensive and publicly

accessible repository of address-level information in Brazil. It systematically records all built and under-construction units observed within each census tract across the national territory. Updated after every demographic census, this database serves as a key instrument for the Brazilian Institute of Geography and Statistics (IBGE) to locate and classify buildings by their primary use, thereby supporting a detailed spatial understanding of where people live and how the built environment is utilized.

Building upon this foundation, the study makes three main methodological contributions to the analysis of urban land use. The first contribution concerns the adoption of a standardized, nationwide data source (the CNEFE) for the spatial interpretation of land use. Traditional land-use studies in Brazil often depend on administrative or cadastral datasets that vary substantially among municipalities, limiting comparability across regions. While remotely sensed data have been increasingly used to overcome such limitations, they are typically constrained by classification inaccuracies that reduce their reliability for fine-grained urban characterization. In contrast, the CNEFE provides point-based, rather than area-based information, allowing for a more detailed spatial depiction of urban structure and facilitating the detection of mixed-use patterns at multiple scales.

The second contribution lies in the development of two new land-use indices designed to overcome some conceptual limitations of conventional heterogeneity measures. Standard indices, such as entropy-based or concentration metrics, quantify the degree of land-use mixture but fail to indicate *which* type of use dominates. As a result, highly homogeneous areas cannot be interpreted as predominantly residential or non-residential without recourse to additional data. The proposed indices address this limitation by employing a directional scale ranging from -1 to $+1$, where negative values denote predominantly non-residential areas, values around zero represent balanced or mixed configurations, and positive values indicate residential predominance. This directional property enhances interpretability and enables more nuanced comparisons across heterogeneous urban contexts.

The third contribution concerns the creation of an open-source R repository that enables the automatic download, processing, and computation of all proposed indices for any Brazilian municipality. The repository includes functions for retrieving CNEFE address data and aggregating them within user-defined H3 hexagonal grids at multiple spatial resolutions. This flexible structure allows researchers and practitioners to systematically evaluate and visualize mixed land-use patterns, while explicitly addressing issues related to the Modifiable Areal Unit Problem (MAUP), a recurring challenge in spatial analysis that affects the interpretation of results across different geographic scales.

2. Materials and Method

2.1. Data sources

The empirical analysis draws upon the CNEFE, produced by the IBGE as part of the 2022 Population Census (IBGE, 2024). Beyond serving as an operational basis for census enumeration, the CNEFE functions as a nationwide geospatial register of all built or under-construction addresses.

During the 2022 census, each address was verified or newly added in the field by enumerators equipped with handheld mobile data collection devices integrating GNSS receivers. Coordinates were recorded in three operational contexts: confirmation or inclusion of an address, return visits when the resident was initially absent, and during the household interview. The data collection system required up to three GNSS readings before allowing the enumerator to classify the address by its land-use category. To optimize fieldwork, each device was preloaded with a preexisting list of addresses for the census tract, which enumerators systematically updated through on-site verification.

The resulting coordinates, referenced to the SIRGAS 2000 geodetic system (EPSG:4674), achieved high spatial precision. Pre-census testing by the IBGE measured an average root mean square error (RMSE) of 5.84 meters under ideal open-sky conditions and about 11.71 meters in typical collection environments. In dense urban areas, positional accuracy could deteriorate further due to satellite obstruction, though such instances were mitigated through subsequent validation procedures.

After field collection, IBGE implemented an extensive post-processing and validation workflow to ensure coordinate reliability. Invalid or missing coordinates were replaced through a multi-source geocoding estimation process, which used the geometry of block faces, previously confirmed address locations, nearby addresses within the same street segment, and the enumerator's recorded position at the time of interview. This hierarchical procedure produced a complete and topologically consistent national database of coordinates, with quality levels classified according to the geocoding resolution: address, block face, locality, or census tract.

Following these validation and estimation steps, the consolidated dataset comprised 111.1 million geographic records, of which 98.95% were validated as spatially coherent. More than 110.7 million (99.68%) of these were geocoded precisely at the address level, including 93.8% with original field coordinates, 5.5% with standardized corrections (e.g., apartment units sharing the same number), and 0.7% estimated via spatial interpolation. This extensive positional accuracy ensures that the CNEFE represents the most reliable geospatial basis currently available for nationwide urban analyses.

Each address record includes a unique identifier, its geographic coordinates, and the classification of the establishment's functional type, distinguishing between constructed and under-construction buildings. Within the universe of constructed addresses, the

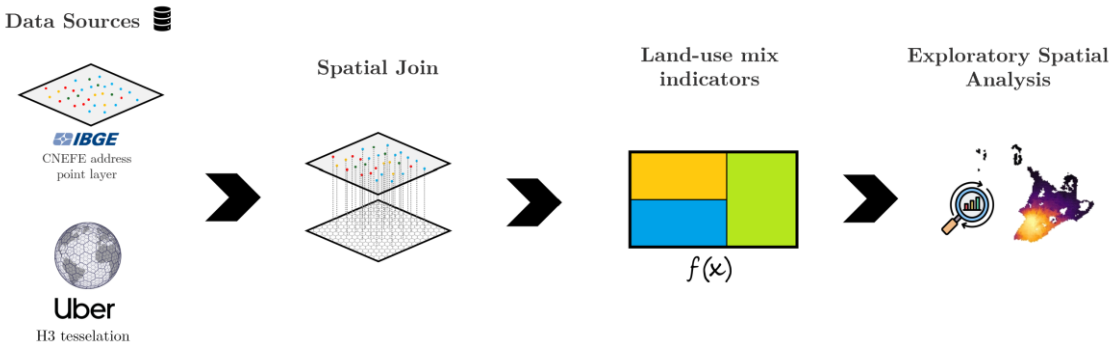
CNEFE identifies residential and non-residential categories. Residential uses comprise approximately 90.6 million private dwellings (81.5%) and 104.5 thousand collective households (0.1%). Among non-residential establishments, the most frequent are other-purpose establishments (≈ 11.7 million, 10.5%), followed by agricultural establishments (≈ 4.1 million, 3.7%), while educational, health, and religious establishments jointly account for about 1.1 million records (≈ 1%). In addition, there are approximately 3.5 million buildings under construction (≈ 3.2%), which, although registered in the census, were excluded from the analyses in this study due to their transitional nature.

To generate spatially explicit land-use indices, this study also employed H3 hexagonal grids at multiple spatial resolutions (https://h3geo.org/). The use of regular hexagonal tessellations offers several methodological advantages over traditional administrative or square-based grids. First, hexagons minimize edge effects and directional bias, as each cell has six equidistant neighbors, ensuring more uniform adjacency relationships and isotropic spatial connectivity, properties particularly relevant when analyzing neighborhood composition and spatial mixing. Second, the H3 hierarchical indexing system allows for seamless aggregation and disaggregation of data across spatial scales, providing a consistent framework for multi-resolution analysis and cross-city comparison. Finally, the grid can be generated directly from municipal boundary geometries using open-source tools, facilitating reproducibility, national-scale application, and integration with other geospatial datasets in standard coordinate reference systems.

2.2 Method

The methodological workflow developed in this study comprises three main stages: (i) acquisition and preprocessing of the address data from CNEFE; (ii) computation of mixed land-use indices through the spatial aggregation of address points within regular H3 hexagonal grids; and (iii) an exploratory spatial analysis for comparative evaluation across multiple urban contexts and spatial resolutions. This methodological sequence is depicted in the flowchart presented in Figure 1 below. All R functions required to implement these steps are openly available in the GitHub repository at github.com/pedreirajr/lumi_cnefe.

Figure 1. Methodological workflow



2.2.1. Data acquisition and preprocessing

Municipal-level CNEFE datasets were automatically retrieved from the IBGE's open data repository[1]. The retrieval process was automated using an R routine that first employed the geobr package (Pereira & Gonçalves, 2019) to obtain the complete list of Brazilian municipalities, including their official names, codes, and geographic boundaries. From this list, the procedure systematically scraped the hierarchical structure of the IBGE repository, thereby constructing a reference index table linking every municipality to its respective downloadable dataset. This index also contained metadata such as the state abbreviation, download URL, and municipal geometry, serving as a master lookup table for subsequent data access.

A dedicated R function, `read_cnefe()`, was then used to automate the download, extraction, and loading of each CNEFE dataset into R. The function locates the municipality's ZIP file using the reference index, downloads it directly from the IBGE server, identifies the enclosed CSV file, extracts it to a temporary directory, and reads it efficiently using the arrow package. At this stage, no preprocessing, filtering, or transformation is performed, ensuring that the dataset remains exactly as provided by IBGE for maximum reproducibility and transparency in subsequent analytical steps.

Finally, the cleaned data were converted into a georeferenced address point layer in the SIRGAS 2000 reference system (EPSG:4674). Each record in this layer includes the address identifier, its spatial coordinates, and the land-use classification, which distinguishes between residential and non-residential categories (e.g., commercial, agricultural, institutional, religious, or health-related). Records corresponding to addresses under construction were excluded, as mentioned earlier.

2.2.2. Computation of land-use mix indices

The computation of land-use mix indices was implemented through the `compute_lumi()` function, which integrates spatial processing, land-use classification, and index calculation in a single workflow. This function builds directly upon the address data loaded by `read_cnefe()`, performing all subsequent validation, spatial aggregation, and computation of land-use measures.

The procedure begins by reading the municipal-level CNEFE dataset returned by `read_cnefe()`, harmonizing variable names, and validating the key columns required for spatial analysis, namely, longitude, latitude, and the land-use classification variable (COD_ESPECIE column). Coordinate values expressed in microdegrees are automatically detected and converted by dividing by 10^6, ensuring that all locations are properly referenced in the SIRGAS 2000 coordinate system (EPSG 4674). Invalid or missing

[1] Available at https://ftp.ibge.gov.br/Cadastro_Nacional_de_Enderecos_para_Fins_Estatisticos/Censo_Demografico_2022/Arquivos_CNEFE/CSV/UF/

coordinates and records without a valid land-use code are removed at this stage, as are entries labeled as *under construction*, whose temporary character prevents an unambiguous land-use classification.

After turning each valid address into a georeferenced point, these are spatially joined to a regular hexagonal H3 grid covering the municipal boundary. The grid is generated directly from the municipality's polygon geometry with the R package 'h3jsr' (O'Brien, 2024) and can be produced at different resolutions.

Subsequently, the number of residential and non-residential addresses was counted, forming the basis for computing local proportions of land-use composition. Let p denote the proportion of residential addresses and $q = 1 - p$ the proportion of non-residential ones. This dichotomous representation allows the functional balance of each cell to be evaluated independently of the absolute number of establishments, focusing instead on relative spatial composition.

From these proportions, four indices were computed to describe land-use balance. Two of them, Entropy Index (EI) and Herfindahl–Hirschman Index (HHI) are conventional measures widely used in the literature, formulated as in Equations (1) and (2), respectively. In this study, the non-percentage form of the HHI (i.e., not multiplied by 100) was adopted to maintain numerical consistency with the other normalized indices.

$$EI = -\frac{p\ln(p) + q\ln(q)}{\ln(2)} \qquad \text{Eq. (1)}$$

$$HHI = p^2 + q^2 \qquad \text{Eq. (2)}$$

Both indices quantify the degree of diversity or concentration in land-use composition within each hexagonal cell, ranging from perfect balance to total dominance of a single use. The EI varies between 0 and 1, where values close to 1 indicate an even distribution between residential and non-residential addresses ($p \cong q \cong 0.5$), while values near 0 reflect homogeneity, that is, cells dominated by only one land-use type.

Conversely, the HHI (here expressed in its non-percentage form) ranges from 0.5 to 1 for a binary classification, but in an inverse manner: higher values denote stronger land-use concentration, while values closer to 0.5 indicate a balanced mix. Together, these complementary indices provide a first approximation of how heterogeneous or specialized the local urban fabric is within each spatial unit.

In addition to the conventional measures, two new indices are proposed in this study to address specific limitations commonly observed in traditional land-use mix metrics. One of these limitations is that when land use within a spatial unit is highly homogeneous,

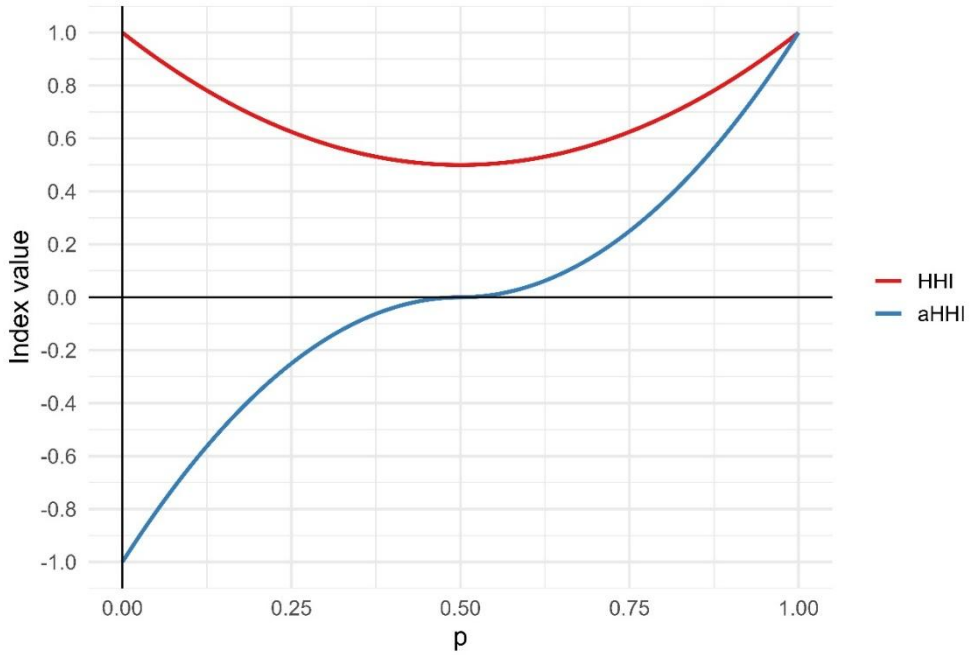
conventional indices (such as EI or HHI) do not indicate *in which direction* the homogeneity occurs, whether the area is predominantly residential or predominantly non-residential.

To overcome this issue, the first proposed measure, the Adapted Herfindahl–Hirschman Index (aHHI), introduces directional information into the conventional HHI by re-scaling its values to a symmetric range between −1 and +1. In this formulation, −1 represents the least residential possible configuration (fully non-residential), 0 denotes a balanced mix, and +1 represents the most residential configuration. The transformation is achieved in two steps: (i) rescaling the standard HHI, which varies between 0.5 and 1 for binary land-use categories, into a [0, 1] interval; and (ii) applying a sign function to preserve the directionality of homogeneity. The resulting formulation is expressed as in Equation (3):

$$aHHI = \text{sign}(p - q) \times \frac{(HHI - 0.5)}{1 - 0.5} \qquad \text{Eq. (3)}$$

This directional adaptation allows distinguishing between equally homogeneous but functionally opposite urban patterns without the need to consult additional maps or indices to understand the direction of homogeneity, thereby improving interpretability in the context of mixed land-use analysis. To illustrate the behavior of the conventional HHI and the aHHI, Figure 2 compares both measures across the full range of residential proportions p, highlighting how the adaptation incorporates directionality while maintaining boundedness.

Figure 2. Comparison between the Herfindahl–Hirschman Index (HHI) and its directional adaptation (aHHI).



A second limitation of conventional land-use mix measures lies in their implicit assumption that a perfectly balanced configuration corresponds to an even 50/50 distribution between residential and non-residential uses. In real urban systems, however, such symmetry seldom reflects the actual functional composition of the broader study area. For instance, in a city where 70% of all addresses are residential and 30% are non-residential, a hexagon with p = 0.7 should reasonably be regarded as "functionally balanced." To account for this, a global parameter P, representing the overall residential share in the study area, is introduced as the reference value for balance (in this example, $P = 0.7$).

To address this specific issue, we propose the Bidirectional Global-centered Balance Index (BGBI), to capture directional deviations in local land-use composition relative to the overall equilibrium of the study area. Its bidirectional scaling (–1 to +1) allows distinguishing whether homogeneous areas are predominantly residential or non-residential, while centering the measure on the global balance ensures interpretability and comparability across different urban contexts.

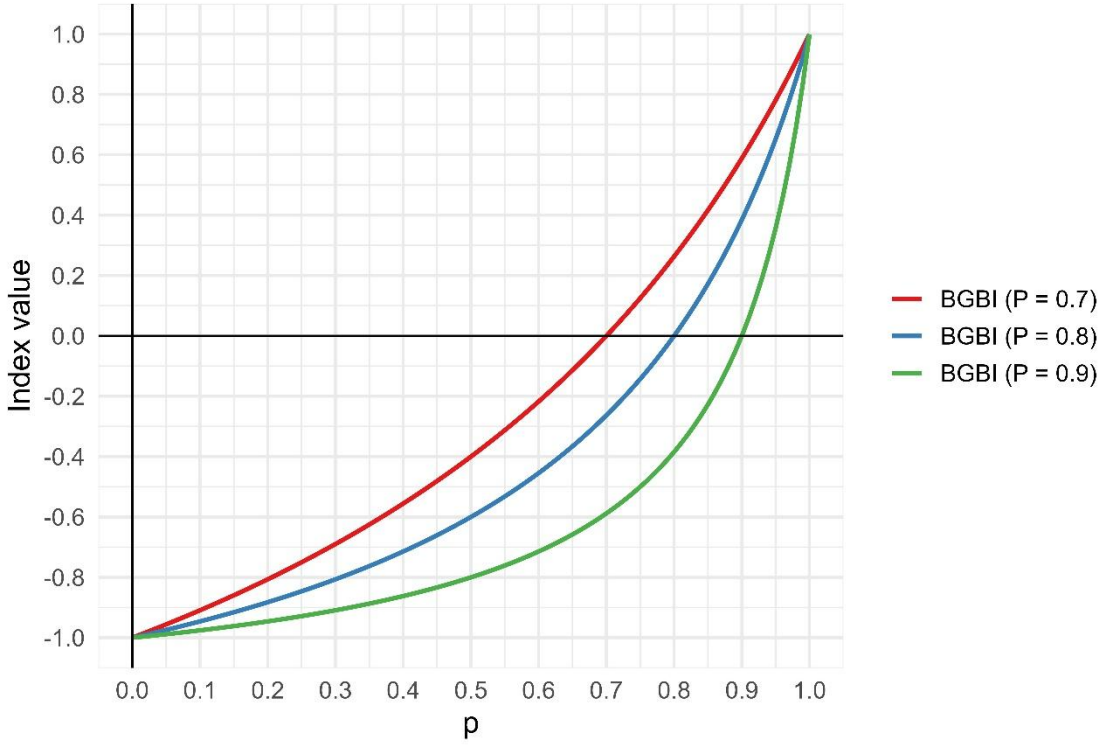
To rescale and center this deviation symmetrically within the interval [−1,1], while keeping the "balanced" situation ($p = P$) fixed at zero, a transformation inspired by a linear fractional mapping is applied, as shown in Equation (4):

$$BGBI = \frac{(2p - 1) - (2P - 1)}{1 - (2p - 1)(2P - 1)} \qquad \text{Eq. (4)}$$

The numerator $(2p - 1) - (2P - 1)$ represents the directional deviation between the local and global residential proportions, once p and P have been rescaled to $p' = 2p - 1$ and $P' = 2P - 1$, so that they range from −1 (fully non-residential) to +1 (fully residential). The denominator $(1 - P'p')$ acts as a nonlinear scaling term that keeps the resulting index within $[-1,1]$, regardless of the global residential share P. In addition, the presence of P' in the denominator controls the curvature of the transformation. When $P = 0.5$, we have $P' = 2P - 1 = 0$, so the mapping becomes linear and the index reduces to a simple centered difference, corresponding to a conventional linear interpolation. A detailed mathematical derivation is provided in Appendix A.

To illustrate, the plot in Figure 3 shows the behavior of the BGBI for three different values of the global residential share: $P = 0.7$, $P = 0.8$, and $P = 0.9$. In each case, the index remains centered at zero when the local proportion p equals P, while smoothly approaching –1 and +1 at the fully non-residential and fully residential extremes, respectively. This visualization highlights the bounded and symmetric nature of the transformation, as well as the shifting of the equilibrium point according to the global balance of land uses.

Figure 3. Behavior of the Bidirectional Global-centered Balance Index (BGBI) for different global residential proportions ($P = 0.7, 0.8, 0.9$).



2.2.3. Exploratory data analysis

The exploratory data analysis aimed to evaluate how the proposed and conventional land-use mix indices behave across different urban contexts and spatial resolutions. For this purpose, the analysis was conducted for six major Brazilian urban centers — São Paulo, Rio de Janeiro, Brasília, Fortaleza, Salvador, and Belo Horizonte — selected for their demographic significance, economic weight, and morphological diversity. Together, these cities provide a robust comparative foundation for examining spatial patterns of mixed land use under distinct metropolitan configurations.

To ensure a systematic spatial framework and to assess potential scale effects, all indices were computed using H3 hexagonal grids at resolutions 8 and 9, corresponding approximately to edge lengths of 0.46 km and 0.17 km, respectively. The use of multiple resolutions was essential to assess the well-known modifiable areal unit problem (MAUP), whereby integral measures are sensitive to the size of the analytical units (Dark & Bram, 2007). In this case, larger areas may appear more mixed than smaller ones simply because they encompass a broader scale.

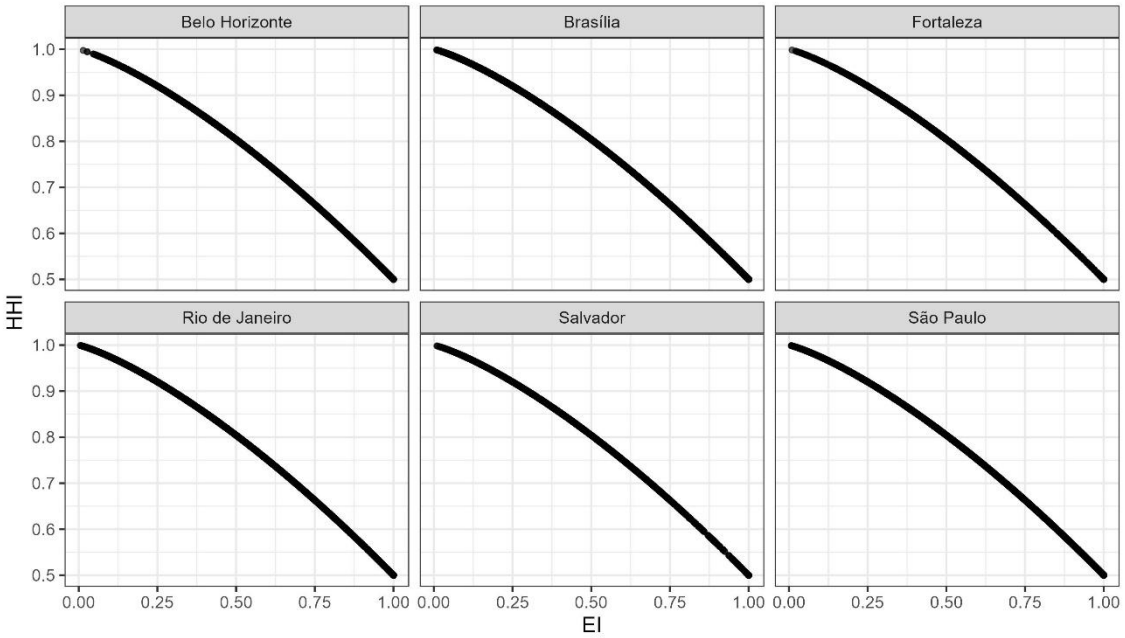
The application of a consistent hexagonal framework allows direct comparison between resolutions while maintaining topological regularity and uniform neighborhood relationships. This facilitates multiscale interpretation of spatial heterogeneity and supports an evaluation of the robustness of each indicer across varying spatial granularities.

3. Results and discussions

This section presents the main results obtained from the computation of land-use mix indices across the six selected Brazilian metropolitan areas — São Paulo, Rio de Janeiro, Brasília, Salvador, Fortaleza, and Belo Horizonte. To begin, we elucidate the relationships among the different indices (HHI vs. EI and BGBI vs. aHHI) to better understand their internal coherence and comparative behavior across distinct urban contexts.

As expected, an inverse relationship is observed between the EI and the HHI across all cities (Figure 4). As mentioned earlier, this pattern reflects their conceptual opposition: while higher HHI values indicate greater dominance of a single land-use category (homogeneity), higher EI values denote more balanced or mixed conditions ($p \cong 0.5$).

Figure 4. Scatterplot of HHI vs. EI, computed at H3 resolution 9.

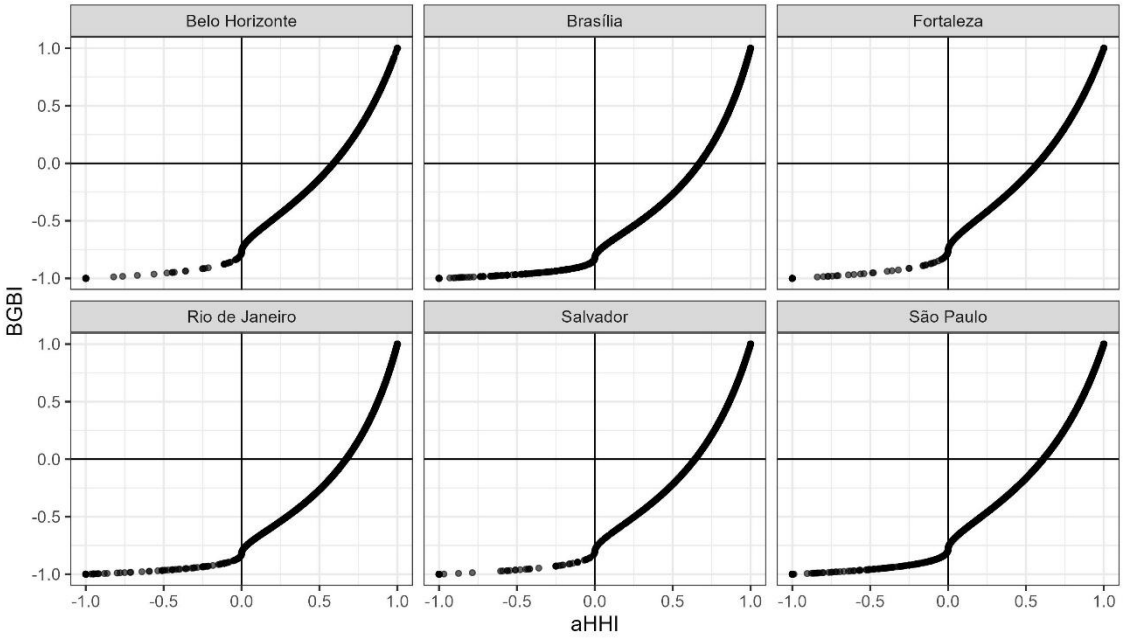


A strong and positive relationship is observed between the aHHI and the BGBI across all metropolitan areas (Figure 5) as both indices share a common directional logic, where -1 represents predominantly non-residential cells, $+1$ predominantly residential ones, and 0 denotes functional balance. However, despite their similar orientation, the BGBI exhibits a nonlinear transformation due to the rescaling applied around the global residential share (P).

This rescaling results in asymmetric sensitivity around the equilibrium: values above P (residential dominance) vary more steeply, whereas values below P (non-residential dominance) vary more gradually. Consequently, the BGBI should not be interpreted as a perfectly symmetric scale (for instance, BGBI $= -0.8$ and BGBI $= 0.8$ do not represent equivalent magnitudes of deviation from balance). Furthermore, because all cities analyzed

exhibit an overall predominance of residential addresses, the distribution of BGBI values tends to be more negative than that of aHHI, reflecting the stronger influence of the global residential proportion on the index's calibration.

Figure 5. Scatterplot of BGBI vs. aHHI, computed at H3 resolution 9.



Figures 6 to 11 display the spatial distribution of the four land-use indices (EI, HHI, aHHI, and BGBI) for the six metropolitan areas analyzed, computed using H3 hexagonal grids at resolution 9. Overall, the maps exhibit comparable spatial patterns across all indices, consistently identifying the main areas of homogeneous or mixed land use.

However, the proposed indices (aHHI and BGBI) offer a clearer and more interpretable depiction of the direction of land use when comparing to the EI and the HHI. The aHHI differentiates whether homogeneity is driven by residential ($+1$) or non-residential (-1) predominance, whereas the BGBI further enhances this interpretation by centralizing balance relative to the global proportion of residential addresses within each city.

Among all indices, the BGBI stands out visually, as balanced areas (those whose local composition reflects the overall residential–nonresidential distribution) are distinctly highlighted in white tones (values near zero). This contrast allows a more intuitive recognition of spatial equilibrium patterns, making the BGBI particularly effective for comparative urban analysis across different metropolitan contexts. Consequently, this visual property makes the BGBI especially useful for mapping equilibrium zones within complex and heterogeneous urban fabrics.

Figure 6. Spatial distribution of mixed land-use indices in Salvador, computed at H3 resolution 9.

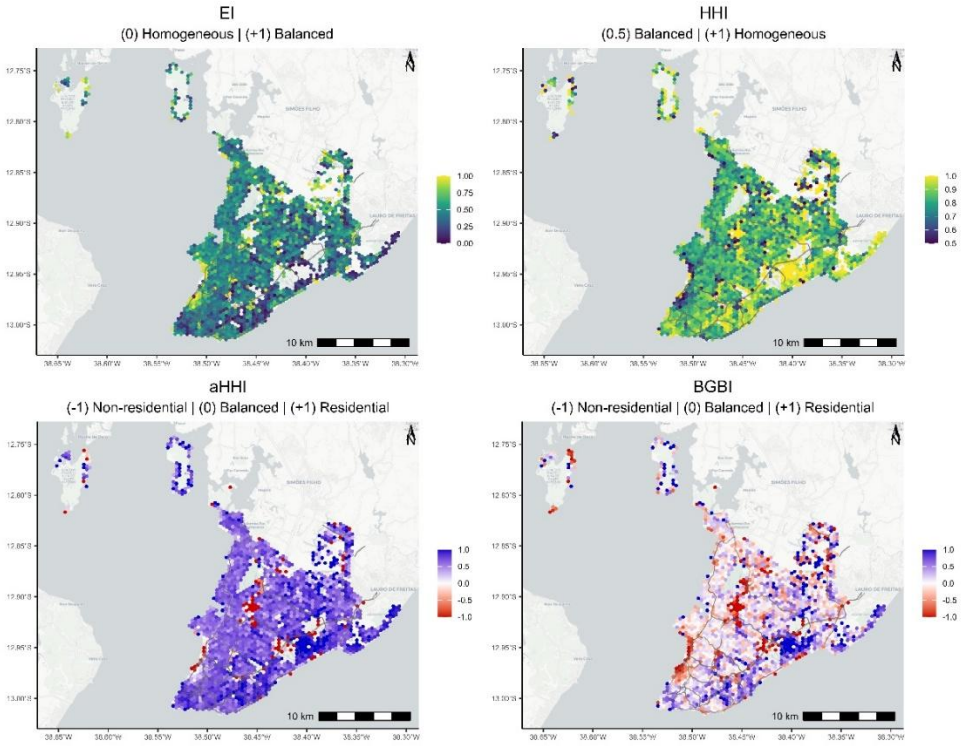


Figure 7. Spatial distribution of mixed land-use indices in São Paulo, computed at H3 resolution 9.

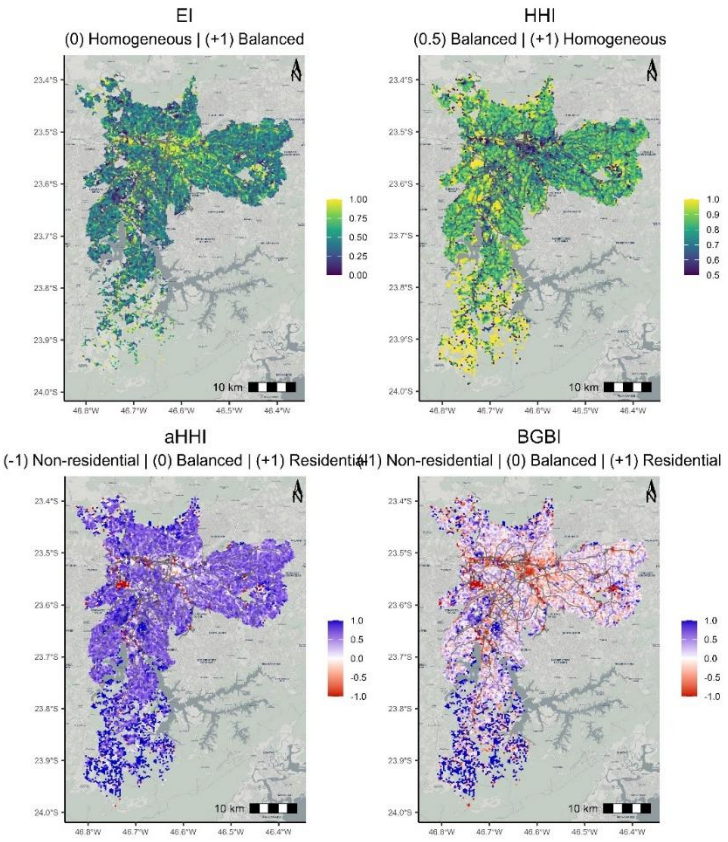


Figure 8. Spatial distribution of mixed land-use indices in Rio de Janeiro, computed at H3 resolution 9.

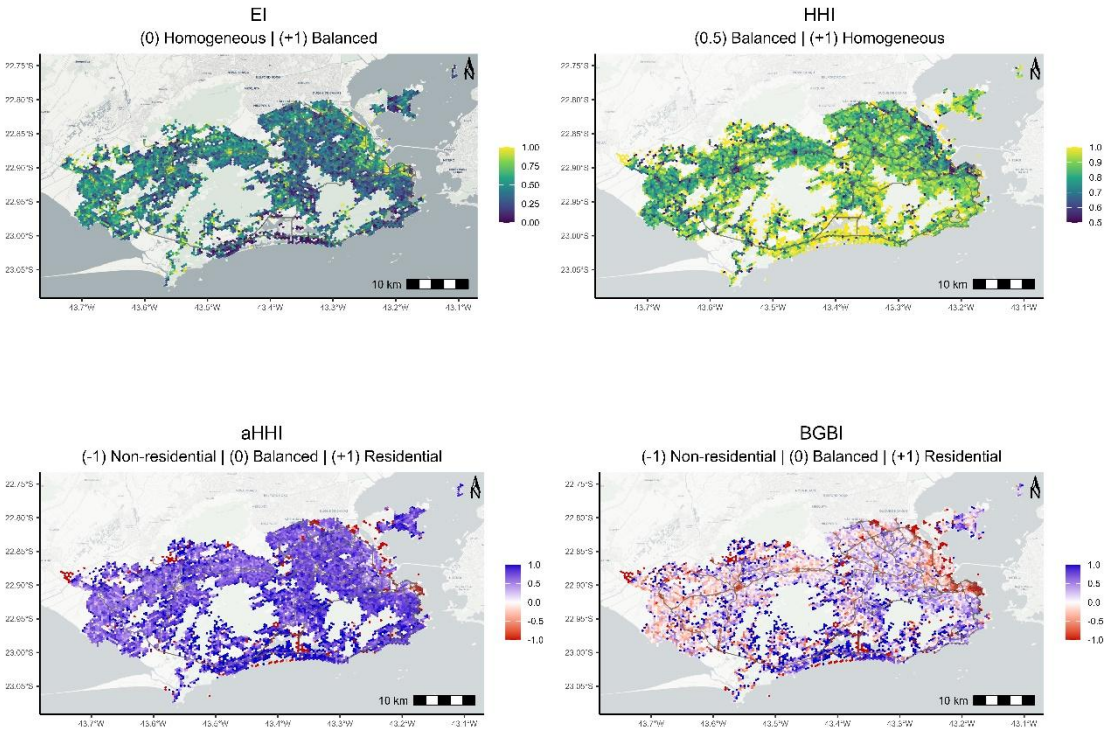


Figure 9. Spatial distribution of mixed land-use indices in Fortaleza, computed at H3 resolution 9.

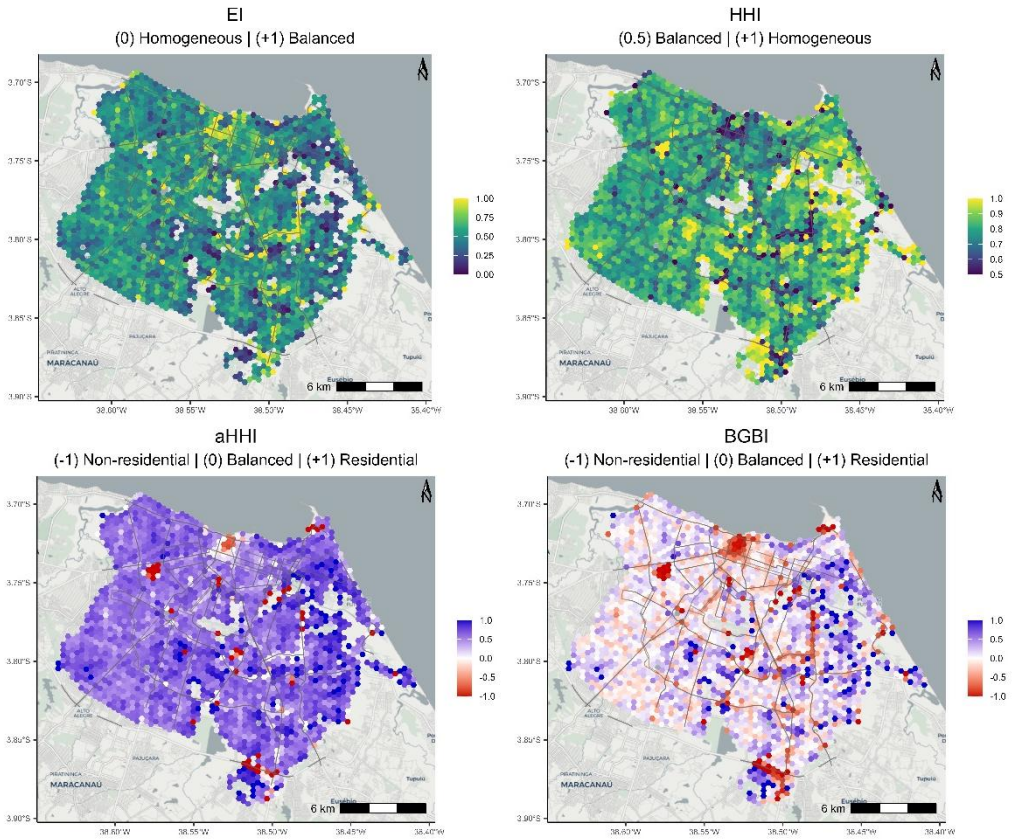


Figure 10. Spatial distribution of mixed land-use indices in Belo Horizonte, computed at H3 resolution 9.

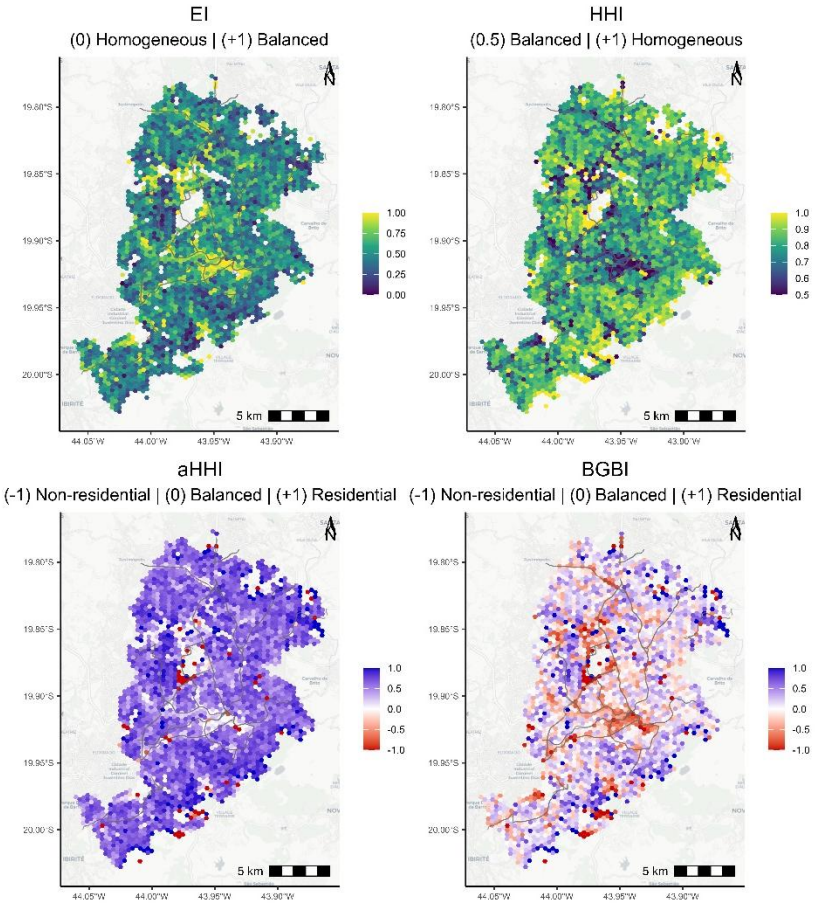
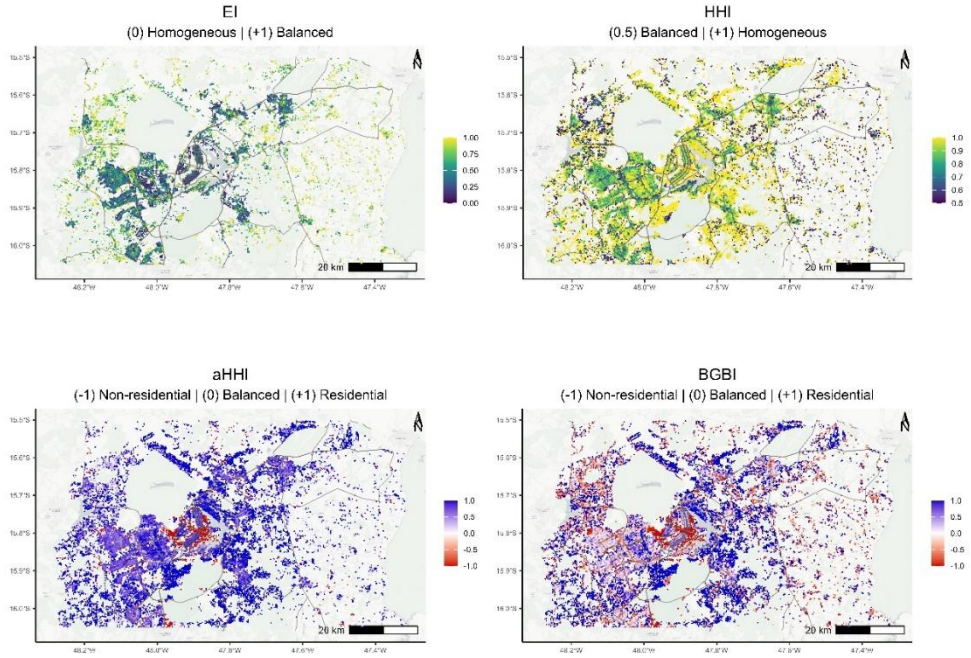


Figure 11. Spatial distribution of mixed land-use indices in Brasília, computed at H3 resolution 9.



Finally, Figures 12 and 13 illustrate the effect of the MAUP on the computation of the BGBI for Salvador and São Paulo, respectively. Although both maps display broadly consistent spatial patterns of land use, discrepancies can be observed in specific areas as a consequence of the aggregation process when moving from finer (H3 resolution 9) to coarser (H3 resolution 8) spatial scales. This variation highlights that the spatial resolution at which land-use indices are computed directly influences the resulting interpretation of urban patterns. Therefore, the selection of the analysis scale must be aligned with the spatial dimension of the phenomenon under study to ensure meaningful and accurate conclusions.

In this regard, the availability of a flexible computational workflow capable of producing the indices across multiple spatial resolutions represents an important contribution of this research, as it allows researchers and practitioners to explicitly assess and visualize the scale sensitivity inherent in mixed land-use analyses.

Figure 12. Spatial distribution of BGBI indices in Salvador, computed at H3 resolutions 8 and 9.

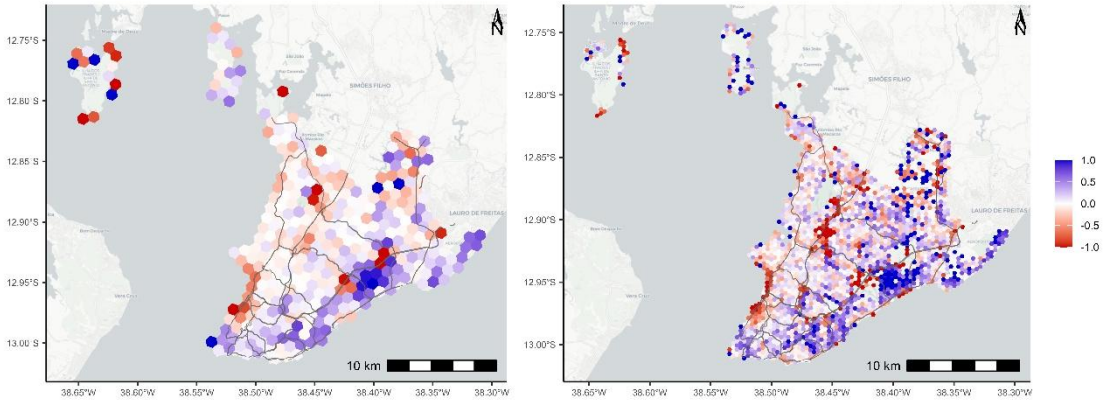
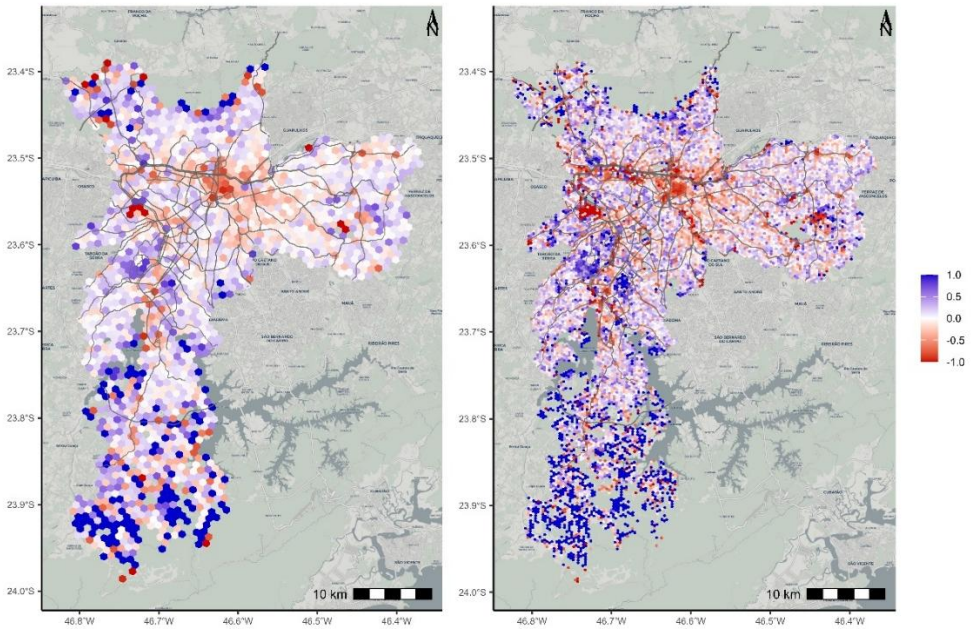


Figure 13. Spatial distribution of BGBI in São Paulo, computed at H3 resolutions 8 and 9.



4. Conclusions

This study presented a methodological framework for computing mixed land-use indices using a comprehensive, nationally standardized dataset, the National Address File for Statistical Purposes (CNEFE), developed and maintained by the Brazilian Institute of Geography and Statistics (IBGE). By leveraging this institutional database, which covers the entire national territory, the proposed approach enables systematic, replicable, and fine-grained analyses of land-use composition across all Brazilian municipalities.

The study makes two primary contributions to the field of urban spatial analysis. First, it demonstrates the potential of institutional, census-based address data as a consistent and reliable source for deriving indices of land-use mix. Unlike local cadastral systems or remote-sensing classifications, which often vary in quality and methodology across municipalities, the CNEFE provides harmonized, point-based information that facilitates direct comparison between cities and scales. Second, it introduces two new directional indices, the *adapted Herfindahl–Hirschman Index (aHHI)* and the *Bidirectional Global-centered Balance Index (BGBI)*, that overcome key interpretative limitations of conventional land-use heterogeneity measures. By capturing both the degree and the direction of functional balance, these indices provide a more nuanced and interpretable picture of urban structure, particularly in identifying areas where residential and non-residential uses coexist in equilibrium.

Despite these advantages, some limitations remain. The first relates to the modifiable areal unit problem (MAUP), which affects all spatial aggregation methods. Although the proposed workflow allows for the computation of indices at any desired level of spatial resolution, results may still vary depending on the chosen grid size or geometry. A second limitation concerns the temporal update cycle of the CNEFE, which is revised only once every decade, potentially restricting the method's applicability for monitoring rapid urban dynamics. Finally, the current formulation distinguishes only between residential and non-residential uses, and thus does not yet provide directional insights when more than two land-use categories are considered.

Future work may address these issues by extending the method in several directions. One priority is to incorporate flexible spatial geometries beyond the H3 hexagonal grid, enabling the computation of indices using alternative zoning systems or custom spatial boundaries. A second direction is to improve computational efficiency, particularly in the spatial joins between address points and grid cells, by leveraging scalable spatial database frameworks such as the spatial extension of DuckDB or distributed engines like Sedona/SedonaDB. A third direction involves the development of an R package that consolidates all functions presented in this study, including data retrieval, preprocessing, and index computation, to facilitate wider adoption, reproducibility, and integration into other spatial analytical workflows. By making these tools publicly available and adaptable, the framework proposed

here lays a foundation for large-scale, open, and comparable analyses of land-use patterns across Brazil and potentially other national contexts.

References

Bahadure, S., & Kotharkar, R. (2015). Assessing sustainability of mixed use neighbourhoods through residents' travel behaviour and perception: The case of Nagpur, India. *Sustainability (Switzerland)*, *7*(9), 12164–12189. https://doi.org/10.3390/su70912164

Bordoloi, R., Mote, A., Sarkar, P. P., & Mallikarjuna, C. (2013). Quantification of Land Use Diversity in The Context of Mixed Land Use. *Procedia - Social and Behavioral Sciences*, *104*, 563–572. https://doi.org/10.1016/j.sbspro.2013.11.150

Chen, H., Su, K., Peng, L., Bi, G., Zhou, L., & Yang, Q. (2022). Mixed Land Use Levels in Rural Settlements and Their Influencing Factors: A Case Study of Pingba Village in Chongqing, China. *International Journal of Environmental Research and Public Health*, *19*(10). https://doi.org/10.3390/ijerph19105845

Dark, S. J., & Bram, D. (2007). The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography*, *31*(5), 471–479. https://doi.org/10.1177/0309133307083294

IBGE. (2024). *Coordenadas geográficas dos endereços no Censo Demográfico 2022 : nota metodológica n. 01*. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2102063

Jiao, J., & Fu, B. (2020). Overview and Applicability of Land Use-mixed Indices in the Smart City. *4th International Conference on Smart Grid and Smart Cities, ICSGSC 2020*, 118–123. https://doi.org/10.1109/ICSGSC50906.2020.9248552

Koster, H. R. A. ;, & Rouwendal, J. (2010). *The impact of mixed land use on residential property values. (TI Discussion Papers Series; No. 10-105/3)*. http://www.tinbergen.nl/ti-publications/discussion-

Li, J., Chen, Y., Zhao, D., & Zhai, J. (2024). The Impact of Built Environment on Mixed Land Use: Evidence from Xi'an. *Land*, *13*(12). https://doi.org/10.3390/land13122214

Li, Q., Chen, X., Jiao, S., Song, W., Zong, W., & Niu, Y. (2022). Can Mixed Land Use Reduce CO2 Emissions? A Case Study of 268 Chinese Cities. *Sustainability (Switzerland)*, *14*(22). https://doi.org/10.3390/su142215117

Nabil, N. A., & Eldayem, G. E. A. (2015). Influence of mixed land-use on realizing the social capital. *HBRC Journal*, *11*(2), 285–298. https://doi.org/10.1016/j.hbrcj.2014.03.009

O'Brien, L. (2024). *h3jsr: Access Uber's H3 Library - R package version 1.3.1.* https://obrl-soil.github.io/h3jsr/

Pereira, R. H. M., & Gonçalves, C. N. (2019). *geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil.* Github Repository. https://github.com/ipeaGIT/geobr

Raman, R., & Roy, U. K. (2019). Taxonomy of urban mixed land use planning. *Land Use Policy*, *88*. https://doi.org/10.1016/j.landusepol.2019.104102

Song, Y., Merlin, L., & Rodriguez, D. (2013). Comparing measures of urban land use mix. *Computers, Environment and Urban Systems*, *42*, 1–13. https://doi.org/10.1016/j.compenvurbsys.2013.08.001

Wo, J. C. (2019). Mixed land use and neighborhood crime. *Social Science Research*, *78*, 170–186. https://doi.org/10.1016/j.ssresearch.2018.12.010

Zahnow, R. (2018). Mixed Land Use: Implications for Violence and Property Crime. *City and Community*, *17*(4), 1119–1142. https://doi.org/10.1111/cico.12337

APPENDIX A – Derivation of the Bidirectional Global-centered Balance Index (BGBI)

A.1. Background and objective

The Bidirectional Global-centered Balance Index (BGBI) is inspired from a linear fractional mapping, applied to normalized proportions of residential and non-residential addresses. The objective of using this transformation is to construct a bounded and directionally meaningful index of local functional balance relative to the global composition of land use within a study area.

The desired mathematical and interpretive properties of this index are:

1. **Centering** — zero represents the global balance, where the local proportion of residential use equals the global proportion ($p = P$).
2. **Boundedness** — values are constrained to the interval $[-1, 1]$.
3. **Directionality** — positive values indicate a local dominance of residential uses ($p > P$), and negative values indicate non-residential dominance ($p < P$).

A.2. The linear fractional mapping

A linear fractional mapping is a transformation that is represented by a fraction whose numerator and denominator are linear. In our study, since land-use proportions p and P are defined in the range $[0, 1]$, a preliminary transformation is applied to rescale and center them into $[-1, 1]$:

$$p' = 2p - 1$$

$$P' = 2P - 1$$

This rescaling ensures symmetry around zero:

- $p = 0$ (fully non-residential) $\rightarrow p' = -1$;
- $p = 0.5$ (balanced) $\rightarrow p' = 0$;
- $p = 1$ (fully residential) $\rightarrow p' = +1$.

The same transformation applies to the global reference P, producing $P' = 2P - 1$.

For the BGBI, a simplified and normalized version of the linear fractional mapping is then employed:

$$f(p') = \frac{p' - P'}{1 - P'p'}$$

The numerator $(p' - P')$ represents the signed deviation between local and global residential shares. The denominator $(1 - P'p')$ acts as a nonlinear scaling term that keeps the resulting index within $[-1,1]$, regardless of the global residential share P. In addition, the presence of P' in the denominator controls the curvature of the transformation. When $P = 0.5$, we have $P' = 2P - 1 = 0$, so the mapping becomes linear and the index reduces to a simple centered difference, corresponding to a conventional linear interpolation.

This form has some desirable properties:

- $f(P') = 0$, establishing the equilibrium at the global balance;
- $f(p') \in [-1,1]$, whenever $p' \in [-1,1]$; and
- the mapping is smooth, monotonic, and bijective from $[-1,1]$onto $[-1,1]$.

A.3. The Bidirectional Global-centered Balance Index (BGBI)

Therefore, substituting $p' = 2p - 1$ and $P' = 2P - 1$ into the linear fractional mapping yields:

$$BGBI = \frac{(2p - 1) - (2P - 1)}{1 - (2p - 1)(2P - 1)}$$

This formulation provides a bounded and directionally oriented index of local deviation from the global land-use balance.

A.4. Interpretation and properties

1. **Boundedness:**
   The denominator $1 - (2p - 1)(2P - 1)$ acts as a scaling term that ensures $BGBI \in [-1,1]$ for all $p, P \in [0,1]$.
2. **Centering:**
   When $p = P$, the numerator equals zero, yielding $BGBI = 0$.
3. **Directionality:**
   - $BGBI > 0$, when $p > P$(more residential than expected globally);
   - $BGBI < 0$, when $p < P$(more non-residential than expected).
4. **Curvature:**
   The term $s = 2P - 1$defines the curvature of the mapping:
   - If $P = 0.5$, the function reduces to a linear mapping $BGBI = 2(p - P)$;

- If $P \neq 0.5$, the transformation becomes nonlinear, stretching the side corresponding to the dominant global land-use type.

A.5. Example

If the global residential proportion is $P = 0.7$, then $P' = 0.4$, and:

$$BGBI = \frac{(2p - 1) - 0.4}{1 - 0.4(2p - 1)}$$

- $BGBI = 0$, when $p = 0.7$;
- $BGBI = +1$, when $p = 1$;
- $BGBI = -1$, when $p = 0$.