Geographical Analysis

ORIGINAL ARTICLE OPEN ACCESS

A Spatial Clustering Approach of Economic Activities: An Application to the Agribusiness Sector in Brazil's Midwest Region

Gilvan Santos do Rosario Junior[1,2] | Ticiana Grecco Zanon Moura[2] | Mauro Ferrante[3]

[1]Gran Sasso Science Institute, L'Aquila, Italy | [2]Universidade Estadual de Santa Cruz, Ilhéus, Brazil | [3]University of Palermo, Palermo, Italy

Correspondence: Ticiana Grecco Zanon Moura (tgzmoura@uesc.br)

Received: 3 March 2026 | **Revised:** 30 June 2026 | **Accepted:** 9 July 2026

ABSTRACT

The spatial distribution of economic activities is central to regional economics. However, empirical tools for comparing industries based on the similarity of their spatial distributions remain limited. Traditional cluster analysis typically ignores geographical distance and connectivity, while spatial methods focus on clustering places rather than economic activities. This paper addresses this methodological gap by proposing a framework for clustering economic activities based on the similarity of their spatial distributions, explicitly incorporating geographical space. The Wasserstein distance is employed with a geographically informed cost matrix, that is, the Shimbel matrix, to measure the effort required to transform one activity's spatial distribution into another. This approach is compared with traditional non-spatial distance metrics through an illustrative example and then applied to 272 agribusiness activities across 466 municipalities in Brazil's Midwest region from 2013 to 2021. The results reveal four main clusters: one spread throughout the region and others concentrated in northern, southern, or eastern areas, consistent with regional endowments, historical trajectories, and agro-industrial linkages. The case study suggests that the approach offers a useful complementary tool for identifying co-distributed activities and spatially similar industrial patterns, with implications for regional development analysis, place-based policy design and future studies of economic linkages and shock exposure.

1 | Introduction

It is widely acknowledged that the spatial distribution of economic activities is not arbitrary but an intrinsic characteristic of the economic landscape, which in turn is shaped by underlying geographical, historical, and economic forces. The uneven arrangement of firms, industries, and population across space profoundly influences regional growth, innovation, and resilience. Seminal theories of agglomeration (Marshall 1920) and New Economic Geography (Krugman 1991; Fujita et al. 2001) highlight how spatial clustering arises from increasing returns, specialized labor markets, knowledge spillovers, and reduced transaction costs, leading to the formation of distinct regional economic structures. More recently, concepts of *relatedness* (Hidalgo et al. 2007; Boschma and Iammarino 2009; Neffke et al. 2011) have underscored how diversification into cognitively similar activities can drive regional development. A broader literature of economic geography has operationalized industrial clusters and relatedness through: co-agglomeration, input–output linkages, co-occurrence across regions, product-space measures, and network or regionalization approaches (Ellison et al. 2010; Hidalgo et al. 2007; Neffke et al. 2011; Feser and

Bergman 2000; Feser et al. 2005; Duque and Rey 2008; Delgado et al. 2016). Yet the spatial dimension of this relatedness remains an underexplored avenue.

While traditional cluster analysis (Ward Jr. 1963; MacQueen 1967) represents a powerful statistical tool for grouping similar objects, its application to regional science often faces critical limitations: it typically compares objects based solely on their attribute values, often overlooking their intrinsic spatial arrangement. Existing spatial analysis techniques, such as Exploratory Spatial Data Analysis (Bivand 2010) or density-based clustering like DBSCAN (Ester et al. 1996), are used to identify geographical hotspots or classify locations based on their characteristics. Localization tests, such as Duranton and Overman (2005), assess whether individual industries are spatially concentrated or dispersed. However, these approaches answer questions which are related to, but distinct from, that addressed in this Paper: a framework for comparing several economic activities according to the similarity of their full spatial distributions once geographical cost or connectivity is explicitly incorporated.

This Paper will address this challenge by proposing a new methodological framework for clustering economic activities, which is based on the similarity of their spatial distributions, thereby explicitly incorporating the underlying geographical space. The aim of this approach is to provide a granular, data-driven understanding of how industries are spatially organized across a regional system. By embedding geographical cost into the clustering procedure, the method identifies activities with similar spatial footprints and offers a distribution-based perspective on one spatial dimension of relatedness. Specifically, this Paper seeks to answer the following research questions:

- Which economic activities exhibit similar spatial distribution patterns within a given region when geographical location and connectivity are explicitly considered?
- How do the resulting clusters of spatially similar activities differ from those generated by traditional, non-spatial distance metrics?
- What insights do these spatially informed clusters of co-distributed activities offer regarding regional economic specializations, and the spatial targeting of policies or future analyses of economic linkages and shock exposure?

To answer these questions, the Wasserstein distance will be proposed as a proximity metric within traditional clustering algorithms. The key innovation therein lies in incorporating a geographically-informed cost matrix, specifically the Shimbel matrix, into the Wasserstein distance calculation. This permits us to quantify the 'effort' required to transform the spatial distribution of one economic activity into another, taking into account the actual geographical separation and connectivity between locations. In the empirical application, we also assess the sensitivity of the results to an alternative ground-distance matrix, which is based on intermunicipal travel times, thereby permitting us to compare the baseline topological specification with a road-based accessibility specification.

The approach can be demonstrated with an illustrative example and thereafter applied to a real-world case study of agribusiness economic activities in Brazil's Midwest region. The authors of this Paper contend that the results will be of use to policymakers, regional planners, and researchers, all of whom seek a more nuanced understanding of regional economic structures. By identifying activities with similar spatial footprints, this method can support a more geographically informed regional analysis and provide a basis for future studies of economic linkages, shock exposure, and place-based development strategies.

The remainder of this Paper is organized as follows. Section 2 will review the background relating to cluster analysis and its applications to regional science, thereby highlighting the need for spatially-aware proximity metrics. Section 3 will detail the authors' proposed methodology, also by means of an illustrative example. Section 4 will present a case study using real data from Brazil's agribusiness sector. And finally, Section 5 will discuss the findings of this study, their theoretical and policy implications, and avenues for future research.

2 | Background

The identification and analysis of spatial clusters have long been a cornerstone of regional science, driven by the understanding that the geographical arrangement of economic activities, populations, and resources is intrinsically linked to regional development, competitiveness, and resilience (Marshall 1920; Krugman 1991; Porter 1998). Researchers have extensively employed cluster analysis to unravel complex spatial datasets, aiming at distilling vast quantities of information into interpretable groupings, thereby informing theoretical understanding and practical policy across different fields.

A significant piece of this research has focused on the typology and delineation of geographical units. Early and ongoing studies consistently group regions, municipalities, or census tracts as being based on shared socio-economic, demographic, or environmental indicators. This practice has assisted in delineating functional regions, such as labor market areas (Smart 1974; Coombes et al. 1986; Farmer and Fotheringham 2011) or in creating typologies of urban and rural areas, which are based on their development patterns, industrial profiles, or socio-economic challenges (Cörvers et al. 2009; Van Oort et al. 2012). Such regional classifications are invaluable for designing targeted, place-based policies and for conducting comparative analyses across similar contexts, thereby providing a foundation for understanding differential regional trajectories.

Complementary to this, cluster analysis has been crucial for understanding industrial structures, firms' dynamics and regional specialization. Studies frequently categorize industries based on attributes such as technological intensity (Frenken et al. 2007), market orientation, or input–output linkages (McCann and Fingleton 1996). Furthermore, firms within a region are often clustered by characteristics such as size, innovation capacity, or internationalization strategies, in order to gain insight into the composition and evolution of local industrial ecosystems (Delgado et al. 2014). The mapping of such industrial ecosystems is fundamentally linked to a broader literature in economic geography regarding industrial clusters and *relatedness*. In the Porterian tradition, clusters involve more than spatial concentration: they are understood as systems of interconnected industries connected through supplier relationships, labor-market pooling, knowledge spillovers, institutions, and shared competitive conditions (Porter 1998, 2003; Delgado et al. 2016). In parallel, the *relatedness* literature emphasizes that regions tend to diversify into activities which are technologically, cognitively, or economically close to their existing capabilities (Boschma and Iammarino 2009; Neffke et al. 2011). Together, these approaches can contribute to understanding regional specialization, sources of competitive advantage, and the mechanisms driving economic growth and diversification.

Empirically, this literature has operationalized industrial clustering and relatedness through several strategies. Co-agglomeration measures identify industries which tend to locate near one another, investigating the mechanisms behind such patterns; the latter include input–output linkages, labor pooling, and knowledge spillovers (Ellison et al. 2010). Product-space and co-occurrence approaches infer relatedness from the tendency of products or industries to appear together across countries or regions (Hidalgo et al. 2007; Neffke et al. 2011). Other studies use input–output linkages, national cluster templates, or network and regionalization methods to delineate industrial systems and groups of related industries (Feser and Bergman 2000; Feser et al. 2005; Duque and Rey 2008; Delgado et al. 2016). These approaches provide important evidence on economic, technological, and relational forms of industrial relatedness.

Despite the widespread and successful application of traditional cluster analysis to regional science, conventional distance metrics still pose challenges when the objective is to compare spatial distributions. Metrics, such as Euclidean, Manhattan, or Pearson correlation distances (Kassambara 2017), operate in an abstract feature space, evaluating similarity solely based on numerical attribute values. While valid for non-spatial data, this approach often introduces a significant analytical bias, and it yields incomplete insights when applied to phenomena which are fundamentally shaped by geography. This oversight is problematic for several reasons which are central to regional science: firstly, these methods often do not explicitly incorporate geographical proximity and connectivity. Two economic activities might exhibit numerically similar distributions (e.g., both are highly concentrated, or both are broadly dispersed) yet they may be situated in entirely different, non-interacting parts of a region. Traditional metrics might erroneously deem them to be similar, overlooking the fact that they are too geographically distant or disconnected to share common localized factors, labor markets, or knowledge spillovers (Glaeser et al. 1992); secondly, this traditional approach can be said to struggle to identify spatially embedded patterns of co-distribution. The formation of agglomeration economies, innovative clusters, and regional production networks is deeply contingent upon the co-location and local interaction of economic actors (Combes and Overman 2004). By not incorporating geographical distance and connectivity into the similarity assessment, conventional clustering techniques can obscure groups of activities with shared spatial footprints. This can hinder the identification of activities which may share localized conditions, infrastructure needs, or exposure to similar spatial constraints; thirdly, this limitation can compromise the efficacy of place-based policy design. Policies aimed at fostering regional development, promoting industrial clusters, or mitigating the impact of economic shocks rely heavily on an accurate understanding of which activities are spatially co-distributed and may share common locational conditions. If clusters are defined without robustly accounting for actual geographical relationships, policies may be poorly targeted, thereby failing to leverage possibly shared locational conditions or address specific localized vulnerabilities, both of which can ultimately diminish their intended impact.

In recognizing these challenges, various spatial clustering techniques have been developed within the field of regional science and geographical analysis to explicitly integrate geographical information. These advancements primarily fall into two categories:

- The clustering of geographical units: methods such as the Local Indicators of Spatial Association (LISA) (Anselin 1995, 2019) within Exploratory Spatial Data Analysis (ESDA) permit the identification of local clusters of similar values (e.g., hotspots or coldspots) among geographical units. Similarly, density-based algorithms like DBSCAN (Ester et al. 1996) and FSDP (Rodriguez and Laio 2014) effectively group contiguous or dense geographical locations which share similar attributes and which are considered invaluable for mapping concentrations of social phenomena (Grubesic and Murray 2006; Arribas-Bel and Schmidt 2013).
- Identifying the localization of industries: more specific to economic geography, several statistical tools have been proposed by Ellison and Glaeser (1997), and Duranton-Overman (DO) statistic (Duranton and Overman 2005) to assess whether industries exhibit significant localization or dispersion on various geographical scales. This has been instrumental in characterizing the micro-geography of industries and understanding their spatial concentration patterns (Chatterji et al. 2014; Delgado et al. 2016).

While these literatures represent significant advancements, they primarily focus on either clustering places, analyzing the localization patterns of industries, or identifying economic relatedness through co-agglomeration, input–output, product-space, or network-based approaches. A persistent and critical methodological gap, however, remains: a framework for comparing and grouping different economic activities based on the similarity of their entire spatial distributions while explicitly integrating the geographical cost of interaction across space. Specifically, while traditional methods identify if two industries occupy the same unit, they lack a metric to quantify the similarity of industries that are proximate but not co-incident. This “binary” treatment of space—where things are either in the same unit or they are not—is the primary limitation addressed by the Wasserstein distance-based framework. That is, existing approaches answer complementary but distinct questions: they identify clusters of places, test whether individual industries are localized, or infer economic and technological relatedness among industries. The framework proposed here addresses another question: which economic activities share similar spatial patterns when their full distributions are compared under a geographical cost structure.

In doing so, it adds a spatially explicit, distribution-based layer to the analysis of industrial relatedness, a topic central to contemporary regional science.

3 | Spatial Clustering of Economic Activities: A Wasserstein Distance-Based Approach

3.1 | Proximity Metrics for Clustering Spatial Distribution

Cluster analysis is a fundamental pattern-recognition method which is widely used across disciplines (including regional science) for grouping objects such that those within the same cluster share similar characteristics and differ from other objects in other clusters (Delage and Nakata 2023; Regal et al. 2023; Zelasky et al. 2023). Milligan and Cooper (1987) have broadly categorized clustering methods into hierarchical, partitioning, overlapping, and ordination techniques, each offering different ways that reveal structure within a dataset. For all methods, the choice of distance (or proximity) metric is paramount as it dictates how similarity is quantified between objects and, consequently, it profoundly influences the resulting cluster configurations (Kumar et al. 2014). Traditional distance metrics, such as Euclidean distance (1), Manhattan distance (2), and Pearson correlation distance (3) (Kassambara 2017), are effective for vector-valued data where objects are compared based on their attribute values in a feature space.

3.1.1 | Euclidean Distance

$$d_{\text{euc}}(x, y) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2} \tag{1}$$

3.1.2 | Manhattan Distance

$$d_{\text{man}}(x, y) = \sum_{i=1}^{n} \mid \left(x_i - y_i\right) \mid \tag{2}$$

3.1.3 | Pearson Correlation Distance

$$d_{\text{cor}}(x, y) = 1 - \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \overline{x}\right)^2}} \tag{3}$$

However, these traditional distances might not always be optimal for high-dimensional data or when the distributions being compared have different shapes or supports (Villani 2009). Thus, new distance metrics have been developed, securing a place in the literature. Among them, one distance widely used in cluster analysis is the Wasserstein Distance (El Malki et al. 2020; Irpino et al. 2014; Li and Ma 2020; Zhuang et al. 2022).

The Wasserstein Distance (WD), also known as Earth Mover's Distance, is a metric, which has been developed in accordance with optimal mass transport theory in order to ascertain the minimum cost required to transform one distribution into another (Kantorovitch 1958). Despite having its roots in civil engineering and economics, the ability of Wasserstein distance to compare diverse distributions across spatial or temporal domains has extended its use to other fields. For example, it has been successfully applied to image analysis, such as medical imaging (Aoun et al. 2024; Oh et al. 2020) and facial recognition (He et al. 2019), and to data analysis in biological (Liu et al. 2023), big data and machine learning (Cantisani et al. 2024; Ponti et al. 2022), and decision-making (Zhang et al. 2022).

Given two distributions, μ and ν, defined on a metric space $(\mathcal{M}, d)$, the Wasserstein distance between μ and ν is defined by the formula (Villani 2009):

$$W(\mu, \nu) = \underset{\pi \in \Pi(\mu,\nu)}{inf} \int_{\mathcal{M}} d(x, y)\, \pi(x, y) \tag{4}$$

where $\Pi(\mu, \nu)$ is the set of all probability measures with marginals μ and ν, $d(x, y)$ is the cost for transporting one unit of mass from x to y, and $\pi(x, y)$ is the amount of mass moved from x in μ to y in ν. Among the set of probabilities, the Wassertein distance considers the minimum cost (*inf*). In other words, the Wasserstein distance is the product of the distance between two points in the distributions and the amount of mass transported to transform one distribution into another with the minimum effort. The calculation of $W(\mu, \nu)$ is performed by the R-package *transport* (Schuhmacher et al. 2024). Exhaustive mathematical treatment of the Wasserstein distance is beyond the scope of this study; see Peyré and Cuturi (2019) for an in-depth overview of the topic.

In the baseline specification, $d(x, y)$ is defined by the Shimbel matrix, which assigns a topological transport cost to each pair of administrative regions in the study area. The Shimbel matrix (Shimbel 1951) represents the shortest paths between the nodes of a network, thus determining the minimum number of steps to connect one node to all the others. If two regions share a border, they are considered as being first-order neighbors, while non-contiguous regions are assigned a cost according to the minimum number of intervening regions required to connect them. By incorporating this geographically informed cost, the Wasserstein distance calculation quantifies the 'effort' of transforming spatial distributions while respecting actual geographical connectivity. Specifically in this framework, the distance represents the total transportation cost—measured in topological steps (shortest paths)—necessary to 'move' the economic mass (e.g., firm density) of one industry to match the distribution of another across the regional network. Consequently, a lower 'effort' or distance indicates that two industries are not only numerically similar but also spatially proximate under the adopted connectivity cost structure, thereby suggesting that they may share similar locational patterns.

The Shimbel matrix is only one possible specification of the geographical cost structure. Since the choice of spatial weights or cost matrices can affect spatial results, it should be interpreted as a transparent topological baseline rather than as a direct measure of economic distance, transport costs, travel times, or market-access frictions (Anselin and Bera 1998; Anselin 2002; Griffith 1996; Getis 2009). This distinction is straightforward in an illustrative example, where the spatial units form a regular grid. In the empirical application, however, municipalities are heterogeneous areal units, and adjacency-based costs may be affected by differences in: polygon size, shape, centroid distance, internal heterogeneity, road accessibility, and infrastructure conditions. As a result, two first-order neighbors are assigned the same Shimbel cost, even when the actual geographical or economic friction between them differs substantially.

The Wasserstein framework remains flexible and can incorporate alternative ground-distance matrices based on: physical distance, travel time, transport costs, commuting flows, input–output linkages, or other context-specific measures. In the empirical application, the Shimbel matrix is used as the baseline cost structure, while an intermunicipal travel-time matrix is used as a sensitivity analysis. The purpose of this comparison is not to identify a universally superior matrix, but to assess how stable the resulting clusters are when geographical friction is represented by road-based accessibility rather than by topological adjacency.

The following section will illustrate how the Wasserstein distance, particularly with a geographically informed cost matrix, operates in comparison with other traditional distances, and how it effectively captures the geographical space in clustering. In addition, it will be demonstrated why the ESDA is limited for the purposes of comparing different spatial distributions.

3.2 | An Illustrative Example

In order to highlight the differences between traditional, non-spatial cluster analysis and the spatially informed approach proposed in this Paper, let us consider a hypothetical region composed of several geographical units, and the spatial distribution of six different economic activities (labeled a–f) in this region, as depicted in Figure 1 below. Distributions a and c represent two economic activities, which are highly concentrated in a single geographical unit but in geographically distinct areas (e.g., northwest vs. southeast). Distributions b and d are equally concentrated around a geographical unit while being connected to each other, whereas distributions e and f are spread more diffusely throughout the region.

If traditional distance metrics (Euclidean, Manhattan, Pearson correlation) were to be employed, those not accounting for explicit geographical location, these methods would primarily group economic activities based on the level of concentration or overall shape of their numerical distributions, regardless of where they are situated in space. For example, (a–c) might be grouped together due to their high concentration, (b–d) due to their moderate concentration, and (e, f) due to their dispersed nature. However, by acknowledging the geographical space and utilizing the Shimbel matrix as the cost function within the Wasserstein distance, a different, more spatially relevant set of patterns can be identified: economic activities concentrated in the region's northwestern part (a, b), economic activities concentrated in the southeastern part of the region (c, d), and economic activities truly dispersed throughout the region (e, f).

In order to empirically demonstrate the different results from each distance metric, first the distance matrix for each of the distance metrics[1] is calculated. As shown in the matrices in Figure 2 below, the distance between the spatial distributions varies, depending on the chosen distance metric. The closest distributions for the Euclidean distance are (b–e) and (d, e). Since the economic activities are present in the same geographical unit, the distance between them is reduced. The same result is obtained for the Manhattan distance but, as it is not sensitive to outliers (such as the Euclidean distance), all activities which share an area share the same proximity. The Pearson correlation, being sensitive to outliers, has obtained the same proximity

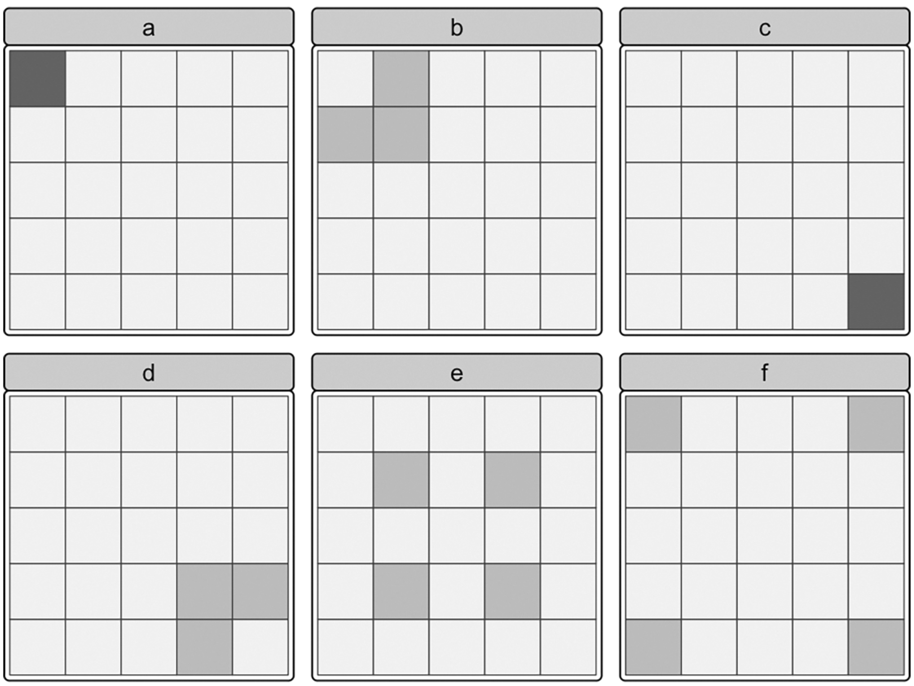

FIGURE 1 | Spatial distribution of hypothetical data.

Euclidean distance matrix

	a	b	c	d	e	f
a	0					
b	115	0				
c	141	115	0			
d	115	82	115	0		
e	112	65	112	65	0	
f	87	76	87	76	71	0

Manhattan distance matrix

	a	b	c	d	e	f
a	0					
b	200	0				
c	200	200	0			
d	200	200	200	0		
e	200	150	200	150	0	
f	150	200	150	200	200	0

Pearson correlation distance matrix

	a	b	c	d	e	f
a	0					
b	1.07	0				
c	1.04	1.07	0			
d	1.07	1.14	1.07	0		
e	1.09	0.82	1.09	0.82	0	
f	0.53	1.16	0.53	1.16	1.19	0

Wasserstein distance matrix

	a	b	c	d	e	f
a	0					
b	100	0				
c	400	367	0			
d	367	267	100	0		
e	250	167	250	167	0	
f	300	250	300	250	100	0

FIGURE 2 | Euclidean, Manhattan, Pearson correlation, and Wasserstein distance matrices of hypothetical data.

between (a–f) and (c–f) since their deviation from the mean is similar. The spatial proximity pertaining to the Wasserstein distance is critical. As depicted in Figure 2 below, the most similar spatial distributions are those in which the economic activities are located closer to each other in geographical space and which can be 'transformed' into one another with a lower geographical cost. Thus, groups like (a, b), (c, d), and (e, f) emerge as the most similar pairs, thereby directly reflecting their spatial coherence.

The fundamental differences among the proximity metrics are clearly reflected in the clustering results of hypothetical data presented in this Paper. Figures 3 and 4 illustrate the hierarchical and k-means clustering analyses respectively, using each of the four distance metrics. As anticipated, the Wasserstein distance is the only metric which successfully groups the economic activities based on their precise spatial distribution patterns, accounting for *where* they are located and their geographical connectivity. In contrast, traditional metrics fail to group these distribution patterns based on these spatially coherent patterns, instead yielding clusters driven by non-spatial aspects of the distributions. Thus, this illustrative example provides strong evidence for the capability of the methodology proposed in this Paper and meaningful spatial relationships between economic activities—otherwise obscured—are revealed.

Regarding the spatial clustering techniques developed in regional science and geographical analysis, the same hypothetical data were used to calculate the Global and Local Indicators of Spatial Association (LISA) within the Exploratory Spatial Data Analysis (ESDA). This method captures the spatial correlation in the distributions but it fails to compare different spatial distributions. The Global Moran's I (Table 1 below) produces the same results for distributions b, d and a–c, even though they are located on opposite sides of the region. The same occurs in the Local Moran's I (Figure 5), where distributions b–d exhibit the same pattern.

It is important to note that the Wasserstein distance matrix used in this Paper compares economic activities, whereas ESDA requires a spatial weights matrix defined among geographical units. Therefore, Wasserstein distances cannot be directly substituted for W in Moran-type statistics without changing the object of analysis. The comparison below is therefore intended to clarify the different analytical purposes of the two approaches: ESDA identifies spatial autocorrelation within each distribution, while the Wasserstein-based framework compares the similarity between different spatial distributions.

4 | Case Study: Agribusiness Economic Activities in Brazil

In order to demonstrate the utility of the Wasserstein distance-based approach, as presented in this Paper, in a real-world scenario, hierarchical clustering was applied to classify 272 agribusiness economic activities[2] in Brazil's Midwest. This analysis utilized the spatial distribution of firms across the region's 466 municipalities from 2013 to 2021. The underlying data were obtained from Brazil's Annual List of Social Information (RAIS, 2013–2021; Brazilian Ministry of Labor 2025) dataset, which is an annual administrative record of every formal establishment in Brazil managed by the Ministry of Labor. For the sensitivity analysis, intermunicipal travel times were

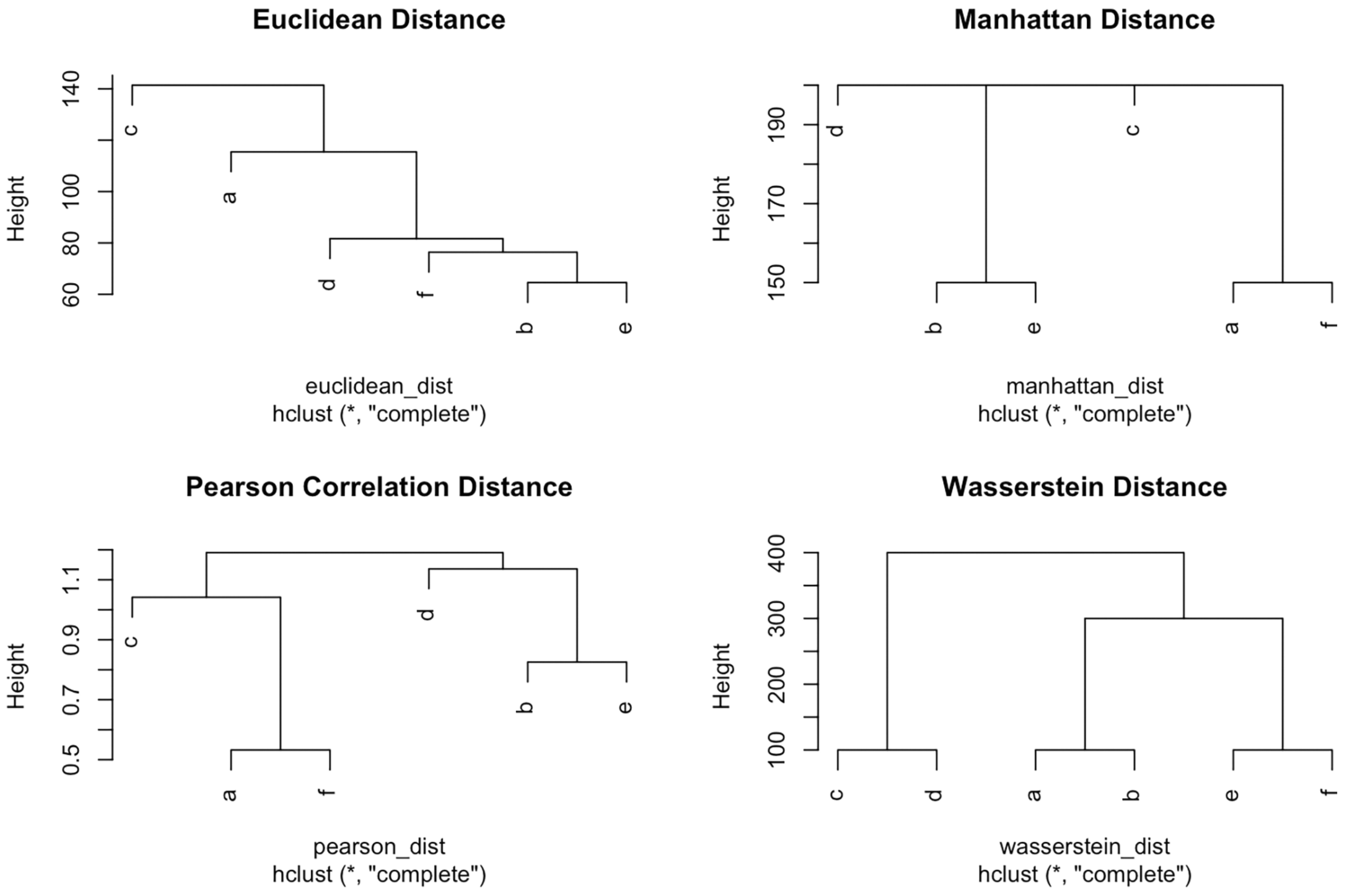

FIGURE 3 | Hierarchical cluster analysis of hypothetical data.

calculated from historical OpenStreetMap road-network data (OpenStreetMap contributors 2026), reconstructed for the relevant reference year and processed through a locally implemented OSRM routing engine (Luxen and Vetter 2011). The resulting origin–destination travel-time matrix was used as an alternative ground-distance matrix in the Wasserstein calculation.

Brazil is a continental-sized country with many natural advantages for agricultural activities. Its favorable climate, relief, abundant water resources, and significant investment in science and technology have transformed the country into a global agribusiness powerhouse. The expansion of Brazilian agricultural frontiers, coupled with increasing demand from emerging countries (primarily China) for agricultural inputs since 2002, has converted the Midwest into the most productive region of Brazil in the agribusiness sector (Vieira Filho and Gasques 2020).

The development of the agribusiness sector in the Midwest region significantly took off in the 1970s: soybean production, the development of adapted crop varieties, and public incentives shifted Brazilian agriculture from subsistence to a business-oriented model, drawing other economic activities into the region (Dall'Agnol 2016). The strong linkages to the soybean system also facilitated the development and establishment of other activities in the region: for instance, meat and egg production depend on animal feed which is produced with soybean meal, thus making livestock another main activity in the Midwest region (Gazzoni and Dall'Agnol 2018).

Soybean farming has also supported other key crops in the Midwest, particularly rice, cotton, and corn farming. Rice farming was initially utilized in the Midwest to open up new agricultural frontier areas. Typically, its cultivation lasts for 2 to 3 years, after which it is replaced by pastures or crops such as soybeans, cotton, or corn. Later, rice farming played a transitional role: reconverting old lands or degraded pastures for soybean farming (Mendez del Villar and Ferreira 2005). Cotton and corn farming exhibit similar dynamics, as both are cultivated in rotation with soybean crops. This crop rotation strategy enhances soil attributes, controls diseases and pests, replenishes organic matter, optimizes fertilizer use, and increases profitability through crop diversification (Silva et al. 2022).

Despite being characterized by a relatively homogeneous set of economic activities, the Midwest region states—Goiás (GO), Mato Grosso (MT), and Mato Grosso do Sul (MS)—also display various unique economic activities due to their natural advantages, strategic location, or historical formation. They have undergone different processes of settlement and agro-industrialization, and they have, therefore, been targeted with distinct public policies. Thus, the spatial-distribution clustering approach, outlined in this Paper, was employed to compare and group agribusiness economic activities exhibiting similar spatial distributions and patterns across the region at the municipality level. This method permits the identification of activities with similar spatial footprints, which may point to shared locational drivers or spatially embedded economic environments.

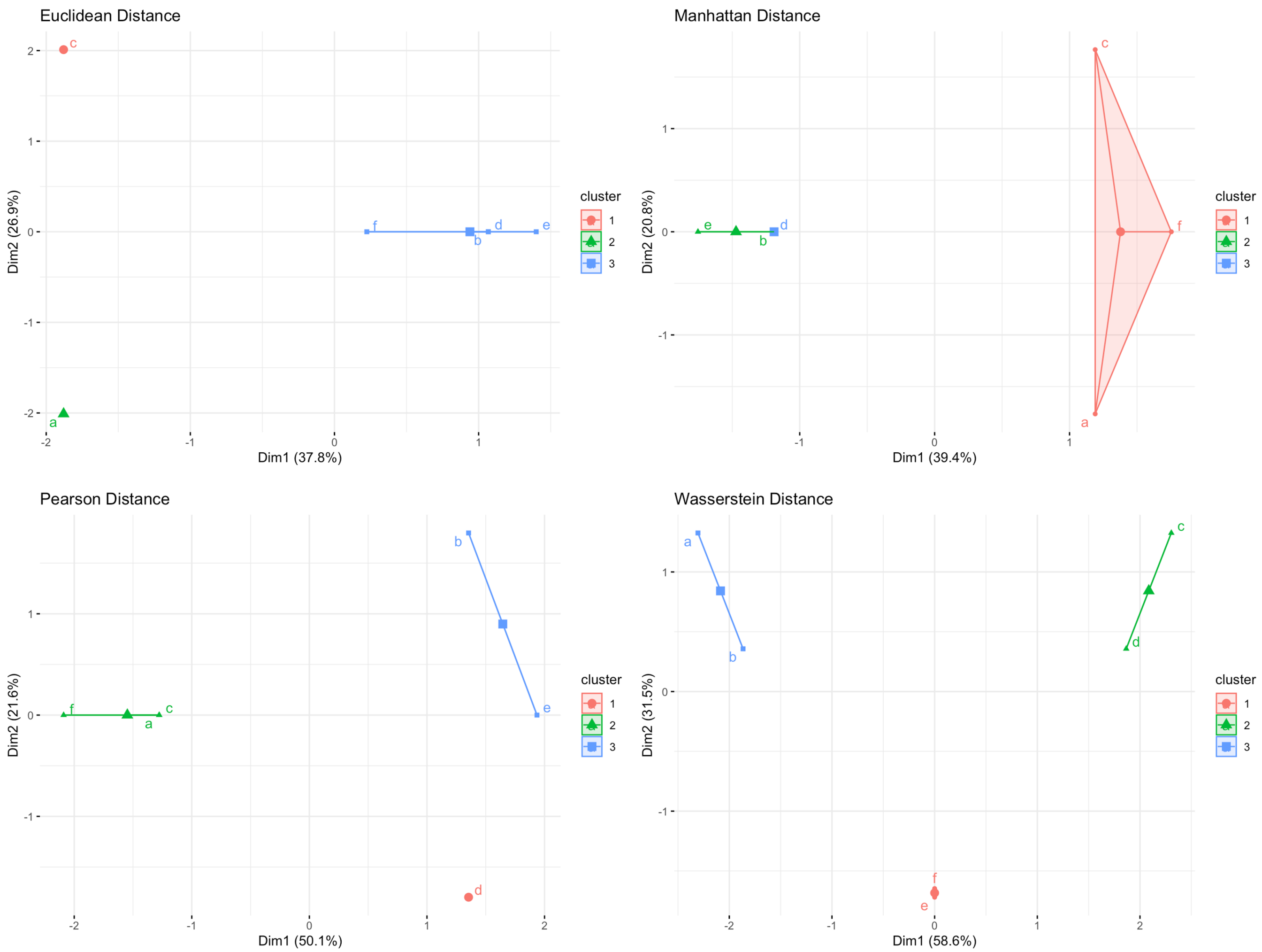

FIGURE 4 | K-means clustering of hypothetical data.

TABLE 1 | Global Moran's I coefficients relating to the hypothetical data.

Spatial distribution	Global Moran's I	*p*
b	0.24684	0.00074
d	0.24684	0.00074
a	−0.00174	0.06441
c	−0.00174	0.06441
f	−0.00794	0.36235
e	−0.33862	0.99903

In order to account for temporal changes regarding the location of firms, the spatial distribution of each economic activity for each year from 2013 to 2021 has been considered. Thus, in addition to comparing different economic activities, the same economic activity across different years has also been compared. The characteristics of each cluster focused on economic activities with consistent spatial distributions, that is, those that were consistently grouped in the same cluster for the same time period being analyzed. The list of economic activities and respective clusters is shown in the Appendix A, with the economic activities considered in the analysis (Table A1) and those not considered due to the temporal changes (Table A2).

The results for the Hierarchical Cluster Analysis using the Wasserstein distance and complete-linkage, with six clusters as the optimal solution, are presented in the dendrogram (Figure 6 below). The differently colored branches in the tree indicate the six different clusters, which have been generated by the analysis, and the height of the branches indicates the similarity of the observations. The groups were defined to maximize the similarity among the observations within the same cluster.

The visual analysis in Figure 6 above revealed four well-defined clusters with similar significant economic activities (Clusters 1–4). Two additional clusters (Clusters 5 and 6) were formed by a very small number of activities and are, therefore, treated as residual clusters. Since they do not provide a clear substantive pattern for regional interpretation, they are reported in the Appendix A; Table A3 but are not analyzed in detail. The economic activities of Cluster 1, here nominated the Spread cluster, are spread throughout Brazil's Midwest region. Economic activities in Cluster 2 (the South cluster) are concentrated in the state of Mato Grosso do Sul.

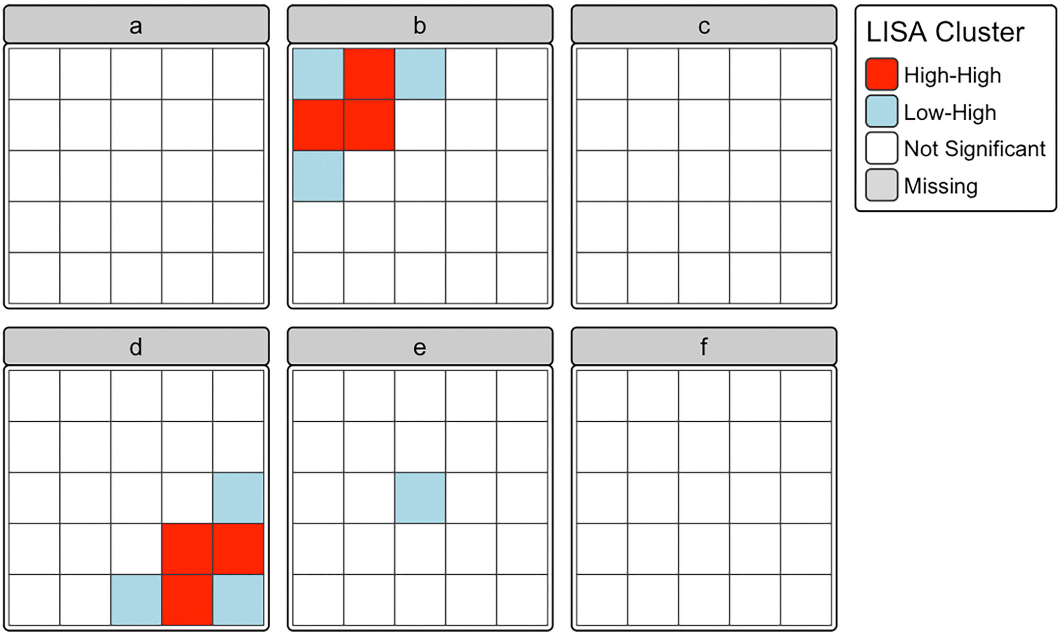

FIGURE 5 | Local Moran's I for hypothetical data.

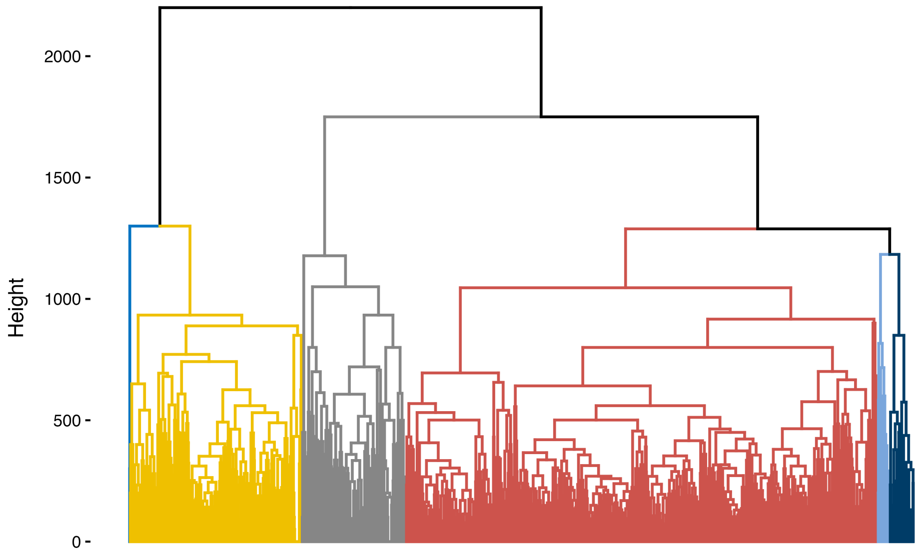

FIGURE 6 | Dendrogram with the economic activity clusters.

Cluster 3 (the East cluster) is concentrated in the state of Goiás and Cluster 4 (the North cluster) is concentrated in the state of Mato Grosso.

The Spread cluster contains the majority of economic activities. Its spatial distribution, shown in Figure 7 below, reveals its dispersed pattern. However, the state capitals are characterized by a higher concentration of firms than the other municipalities. Despite possessing economic activities that are geographically dispersed, the region's urban centers still play a key role in the agribusiness sector, concentrating most of the firms in these centers and functioning as agribusiness cores. Among its economic activities, soybean-related activities stand out. Concurrent with soybean farming, economic activities in its input (pesticides and fertilizers, agricultural machinery, and seeds sectors), and output sides (derived from soybean oil and meal, such as food, biofuel, and animal feed sectors) are located in this cluster. Moreover, the industries linked with soybean production (including meat production and rice, corn, and cotton farming) and their input and output activities (for instance, slaughterhouses and manufacturers and wholesalers of their by-products) are located in this cluster. In summary, the Spread cluster comprises economic activities with spatial distribution patterns that are dispersed across the Midwest states, despite some degree of agglomeration in urban centers. Of note, this includes the region's most prominent and established agro-industrial systems, where soybean serves as a beacon for livestock activities, concurrent with rice, corn, and cotton cultivation.

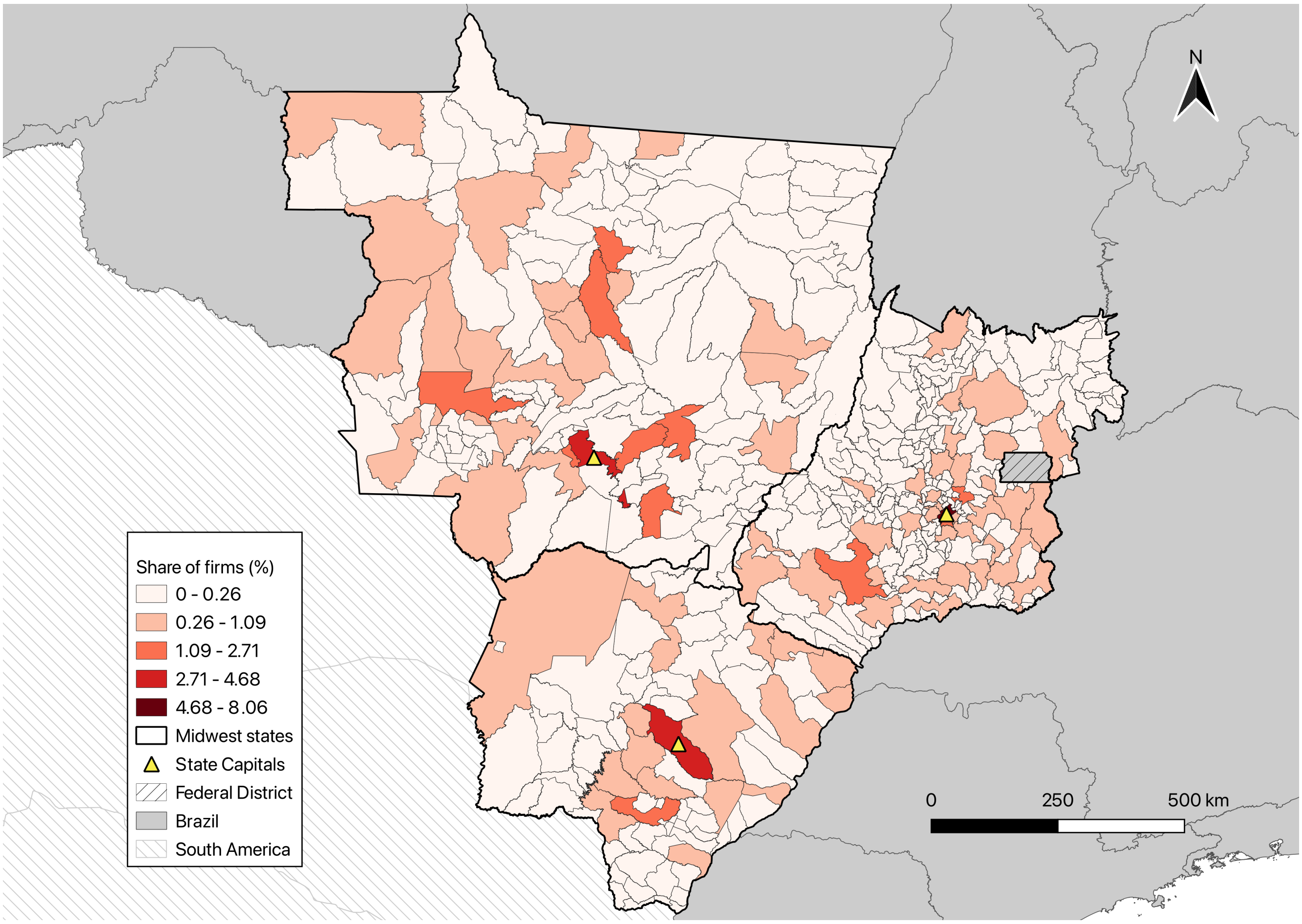

FIGURE 7 | Spatial distribution of firms in Cluster 1 economic activities.

In contrast, the spatial distribution of the South cluster (Figure 8 below) reveals a concentration of firms in the southern and central areas of the Midwest region, particularly in the state of Mato Grosso do Sul and the southern part of Mato Grosso. Economic activities grouped in this cluster are consistent with historical accounts of Mato Grosso do Sul absorbing spillovers from other Brazilian regions. This state was the gateway for soybean farming, a spillover from Brazil's southern region, and it subsequently attracted other economic activities in the same manner. However and unlike soybean farming, which was adapted and extended to the other Midwest states, other economic activities (e.g., sugarcane and yerba mate farming) have been restricted to Mato Grosso do Sul, which was introduced from the neighboring states of São Paulo and Paraná. In addition, oat farming and eucalyptus cultivation are also located in this cluster. In summary, the South cluster consists of economic activities concentrated in the state of Mato Grosso do Sul, with some spillover into the southern regions of the Midwest's neighboring states. These activities include localized oat and eucalyptus farming, which can be attributed to natural advantages, in addition to sugarcane and yerba mate farming due to their proximity to the largest producers.

As presented in Figure 9 below, the spatial distribution of the East cluster displays a concentration in the east of the Midwest region, which is located in the state of Goiás. The latter is the oldest and most diverse economy among the Midwest states. Thus, when soybeans were introduced to the region, the state of Goiás already had an established and varied economy due to the establishment of Goiânia, an important urban center, concurrent with the relocation of Brazil's capital to the Federal District (enclaved within Goiás). While soybean farming and cattle breeding expanded and replaced other land uses, some traditional activities persisted in the state, and agro-industrial activities increased possibly due to the state's superior infrastructure.

Raising cattle for milk is the most significant economic activity within the East cluster. For instance, Goiás accounts for nearly 20% of Brazil's dairy cattle herd, making it the second largest producer (IBGE 2024). In addition, olericulture plays an important role in the state's economy; growing tomato, coffee, beans, potatoes, garlic, spice plants, and onions are grouped in this cluster, in addition to the manufacture of spices, sauces, seasonings and condiments, and canned vegetables. Moreover, Goiás is also responsible for adding value to the agricultural products from the Midwest region. For instance, while the economic activity of tanning and other leather goods is present in all Midwest states, leather products are primarily manufactured in the state of Goiás; this is indicated by the economic activities of the manufacturing of leather shoes and the previously unspecified manufacturing of leather goods, as classified in the East cluster. In summary, the latter is a cluster characterized by moderate-sized

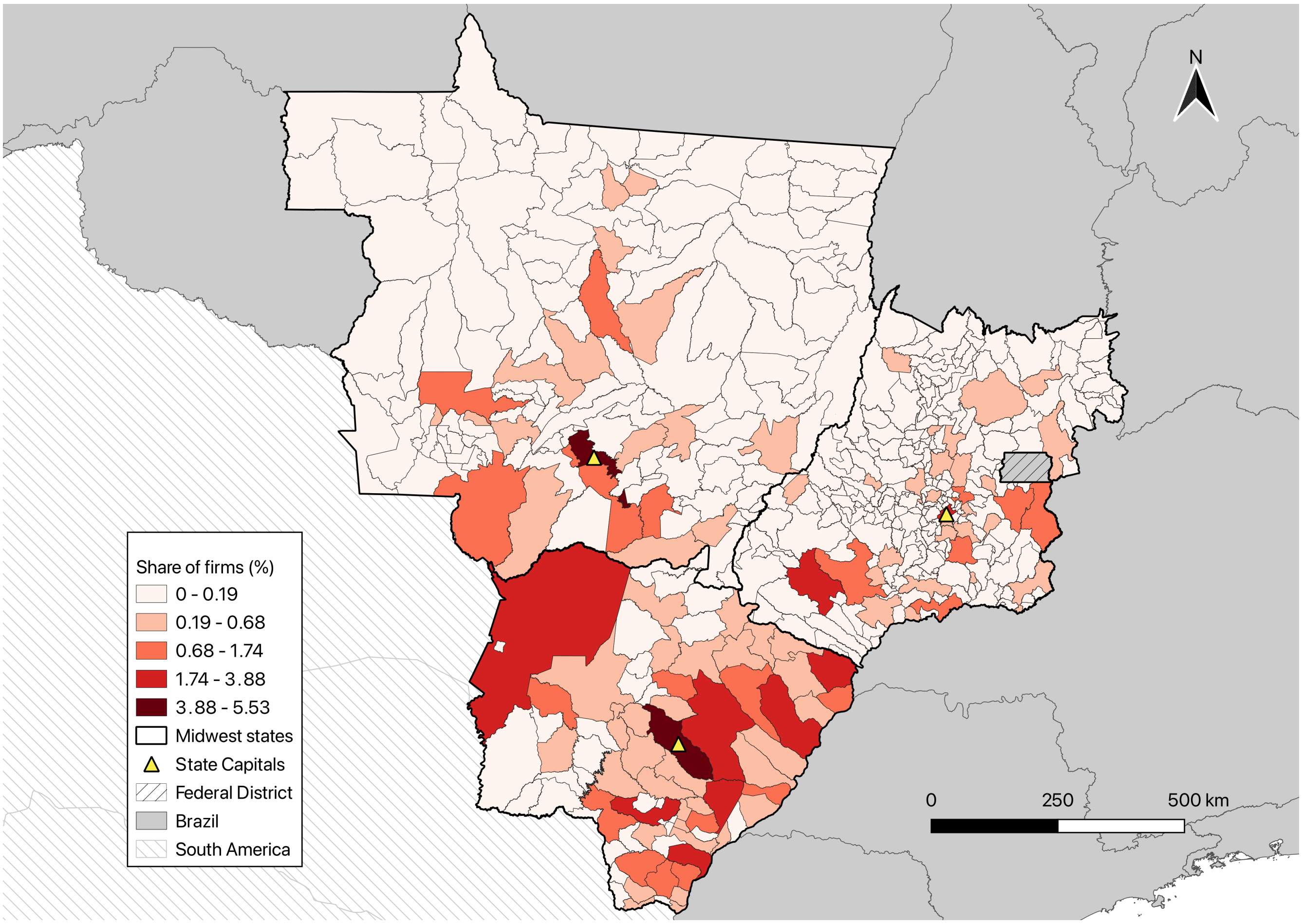

FIGURE 8 | Spatial distribution of firms in Cluster 2 economic activities.

economic activities, with some large-scale exceptions, and a variety of sectors, including dairy cattle, olericulture, and the leather industry.

Finally, the spatial distribution pattern for the North cluster reveals economic activities that are concentrated in the north of the region in the state of Mato Grosso, as shown in Figure 10 below. The state of Mato Grosso is the only state in Brazil's Midwest that includes part of the Amazon Rainforest. Consequently, all economic activities in this cluster are related to the biome, except that of teak farming. Logging activities in native rainforests, sawmills for wood and raw wood splitting are part of the native forest logging supply, while the growing of guarana appears as an extractive activity. In summary, the few economic activities characterizing this cluster reveal a spatial concentration due to a natural advantage.

The generation of distinct clusters composed of economic activities with similar spatial distributions indicates that the Wasserstein distance can compare a large set of economic activities based on their explicit geographical locations and connectivity. The resulting groups identify economic activities with similar spatial distributions under an explicit geographical cost structure. Such co-distribution provides a method for examining common locational conditions, regional specialization, and possible economic connections among activities.

The comparison with the travel-time matrix indicates that the broad regional structure identified with the Shimbel matrix is largely preserved, although the detailed assignment of some activities is sensitive to the adopted cost structure. Among the activities assigned to one of the four substantive clusters under both specifications, 118 remained in the same cluster, whereas 37 changed cluster.

5 | Discussion and Conclusion

The authors of this Paper contend that it contributes to a methodological gap in regional science: the comparison and clustering of economic activities based on the similarity of their full spatial distributions while explicitly incorporating a geographical cost structure. The proposed methodology combines the Wasserstein distance with a geographically informed ground-distance matrix. In the baseline specification, this matrix is defined by the Shimbel matrix, which captures topological connectivity among municipalities. In addition, an intermunicipal travel-time matrix is used as a sensitivity analysis to assess whether the broad regional structure, which is identified under the topological specification, is preserved when geographical friction is represented by road-based accessibility. The research addressed in this Paper has addressed three fundamental questions: firstly, which economic activities exhibit similar spatial distribution patterns within a

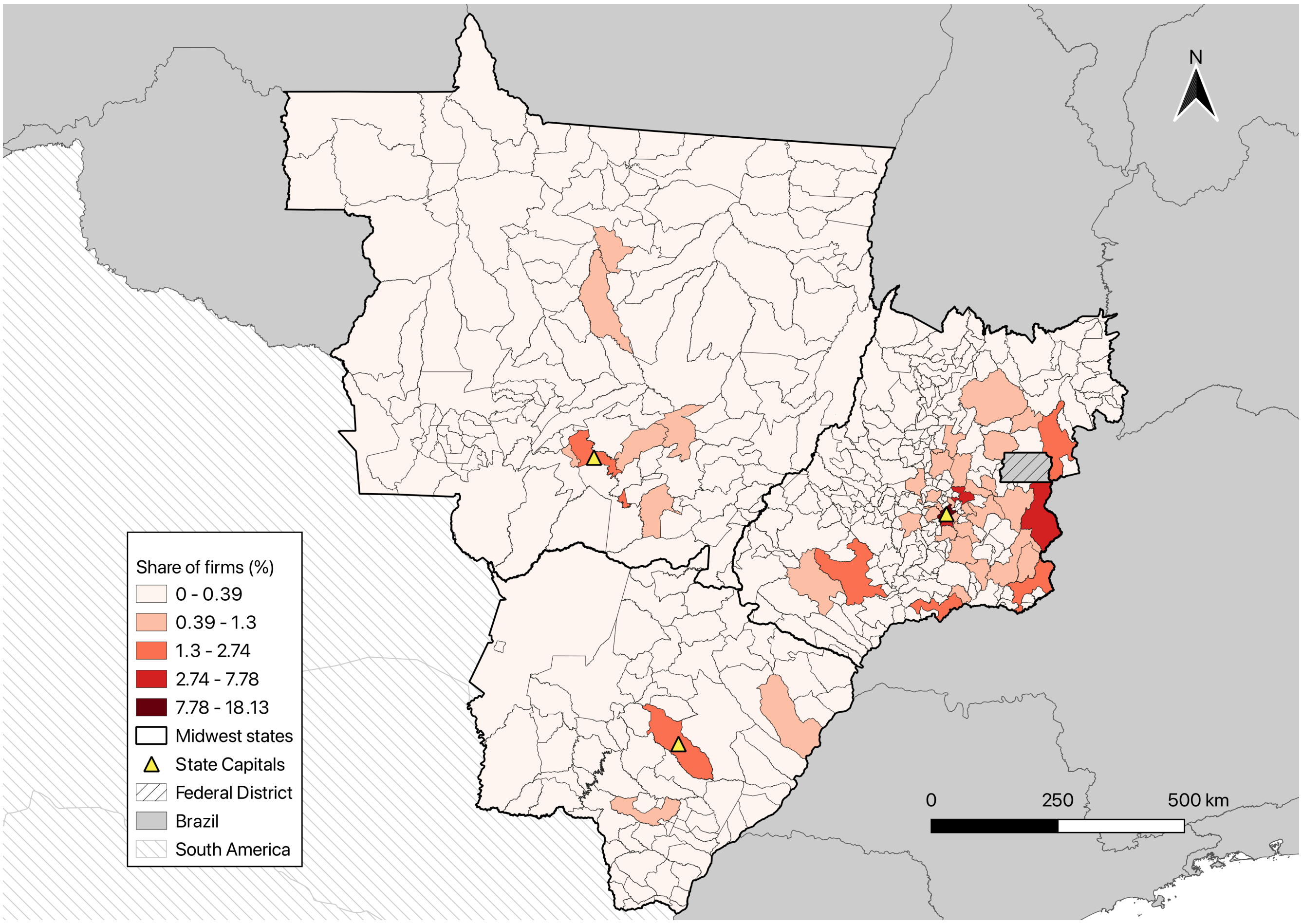

FIGURE 9 | Spatial distribution of firms in Cluster 3 economic activities.

given region when geographical distance and connectivity are explicitly considered? The authors of this Paper hold that the illustrative example with hypothetical data has clearly demonstrated that, by incorporating the Shimbel matrix into the Wasserstein distance, groups of activities (e.g., sectors a and b, or c and d), whose distributions are spatially proximate under the adopted cost structure, can be identified; this implies that their distributions are physically proximate and that they may be shaped by similar locational logic.

The case study in Brazil's Midwest illustrates the empirical usefulness of the approach, revealing distinct clusters of agribusiness activities that are either concentrated in specific sub-regions (North, South, and East clusters) or widely dispersed (the Spread cluster). The sensitivity analysis based on intermunicipal travel times indicates that this broad regional structure is largely preserved when the ground-distance matrix is changed, although the assignment of some activities is sensitive to the adopted cost structure. This result reinforces the interpretation of the Shimbel matrix as a transparent topological baseline, while also demonstrating that the Wasserstein framework can accommodate alternative representations of geographical friction; secondly, how do the resulting clusters of spatially similar activities differ from those generated by traditional, non-spatial distance metrics? The illustrative example with hypothetical data outlined in this Paper demonstrates that traditional distance metrics (the Euclidean, Manhattan, Pearson correlation) may fail to capture these spatially coherent groupings. They tend to group activities that are based on abstract, numerical similarities in their distributions, often ignoring their actual geographical location and connectivity. This highlights an important limitation of conventional approaches when applied to spatially explicit phenomena, thereby underscoring the unique advantage of the Wasserstein-based method in revealing meaningful geographical patterns that would otherwise be obscured.

Finally, what insights do these spatially informed co-distributed activities offer regarding regional economic specializations, potential interdependencies (e.g., spillovers), and the localized impacts of policies or economic shocks? The empirical results from Brazil's Midwest provide useful insights, indicating distinct regional specializations driven by a combination of geographical advantages, historical trajectories, and market dynamics: the North cluster (linked to specific natural endowments like the Amazon rainforest (e.g., indigenous forest logging)); the South cluster (influenced by historical spillovers from neighboring states (e.g., sugarcane, oats, eucalyptus)); the diverse East cluster (shaped by major urban centers (e.g., dairy, olericulture, leather)); and the Spread cluster (representing the region's core with its widely adapted agribusiness systems (e.g., soybean, meat, corn, cotton)). This detailed spatial clustering highlights groups of activities that are co-distributed across space and may

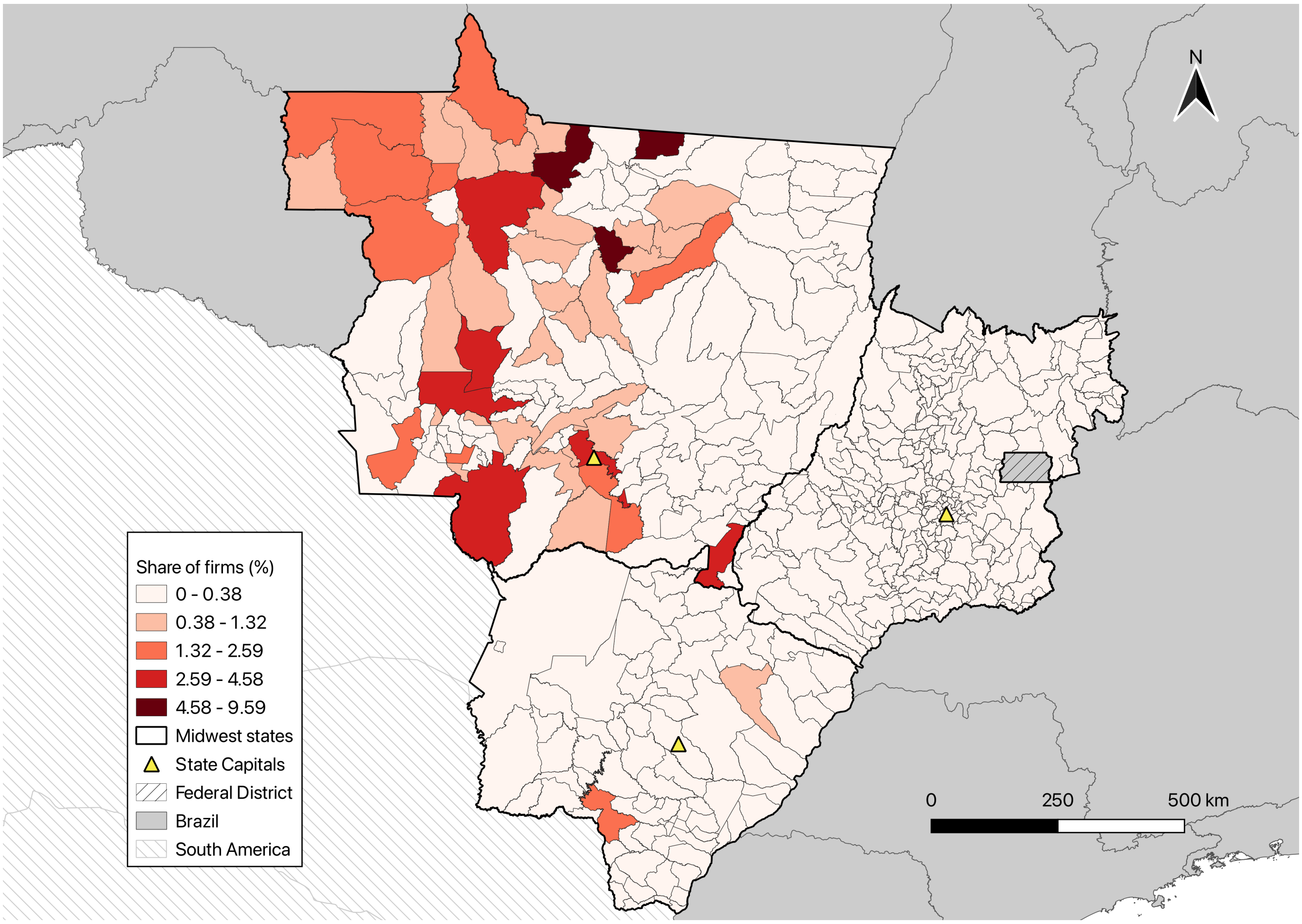

FIGURE 10 | Spatial distribution of firms in Cluster 4 economic activities.

therefore, share locational conditions, infrastructure needs, or exposure to localized shocks. For example, activities grouped within the Spread cluster suggest a widespread spatial system where shocks to one key sector (e.g., soybean farming) could have broad and similar impacts across related activities. Conversely, the more concentrated North cluster suggests that policies or shocks may be more geographically focused, permitting precise, localized interventions. By offering a method to empirically ground relatedness in spatial proximity, the authors of this Paper assert that these findings make a direct contribution to the literature regarding regional economic complexity and diversification (Boschma 2017).

The relevance of considering location in space while clustering variables cannot be overstated. From a theoretical perspective, the methodology outlined in this Paper provides an empirical representation of one spatial dimension of relatedness: similarity in the full geographical distribution of economic activities. This can contribute to a deeper understanding of how industries coalesce and interact across regions, moving beyond abstract theoretical constructs to observable spatial patterns.

From a policy perspective, the methodology can be considered useful as an exploratory tool for informing place-based analysis. Policymakers can leverage the similarity in the spatial distribution of a set of variables in identifying groups that may share locational constraints, infrastructure needs, or exposure to localized shocks. Instead of generic policies, interventions can be tailored to support specific spatially defined value chains or industrial ecosystems, thereby maximizing their localized impact and efficiency. For example, supporting the South cluster in Mato Grosso do Sul might involve specific incentives for sugarcane processing and bioenergy, thereby leveraging the existing infrastructure and expertise, rather than providing a broad agricultural subsidy across the entire Midwest. Moreover, by identifying which economic activities are spatially co-distributed, policymakers can develop more spatially informed hypotheses regarding the potential exposure of co-located activities to economic or environmental shocks. A negative shock in a core sector within a spatially defined cluster (e.g., a drought impacting soybean yields in the Spread cluster) may indicate a shared exposure among co-located activities and it can provide a basis for the further investigation of possible spillover effects (e.g., meat production due to feed reliance), thus permitting proactive mitigation strategies. Conversely, a shock to a highly concentrated but spatially isolated cluster (like the North cluster logging activities) might require more contained and localized policy responses.

Furthermore, the identified spatial clusters can inform infrastructure planning (e.g., transportation networks for specific agricultural products, energy supply for industrial parks) and the allocation of public resources (e.g., extension services, research funding). Activities grouped together spatially often have common demands, thus making it possible to invest efficiently in supporting an entire cluster and potentially improving the targeting of public investment. For instance, infrastructure improvements in the East cluster could prioritize connectivity to urban markets for dairy and olericulture products.

Finally, understanding a region's distinct spatially defined specializations permits a more robust assessment of its economic resilience and potential diversification pathways. Regions with diversified spatial patterns or strong, complementary clusters might exhibit greater stability in the face of economic turbulence. The methodology outlined in this Paper can help identify related industries (Hidalgo et al. 2018) in a spatial context, suggesting new activities which are geographically and functionally in the proximity of existing clusters; targeted diversification policies (e.g., fostering related value-added processing within existing spatial clusters) can thus be formulated.

While the approach described in this Paper offers significant advancements in spatial data cluster analysis, several limitations and avenues for future research could be expanded upon. Firstly, exploring the temporal dynamics of these spatial clusters, and examining how they emerge, evolve, and dissolve over time would provide crucial insights into regional transformation and adaptability. This could involve applying time-series clustering methods to the Wasserstein distance matrices in order to track cluster shifts. Secondly, integrating other dimensions of relatedness (e.g., technological, cognitive, social proximity) with spatial distribution similarity could offer an even richer, multi-faceted understanding of industrial ecosystems and their evolutionary paths (Boschma 2017). A multi-dimensional clustering approach combining the spatial Wasserstein distance with other proximity metrics could be particularly insightful. Thirdly, applying this methodology to different geographical scales (e.g., intra-urban, national) and diverse economic contexts (e.g., manufacturing, services) could test its generalizability and reveal unique, global regional patterns, thereby contributing to a more universally applicable framework for spatial economic analysis. The delimitation of the study area may generate boundary or edge effects: activities located near the limits of the study area may be connected to activities in neighboring regions, such as São Paulo, Paraná, or northern frontier areas, but these external linkages are not observed in the current spatial system. Nevertheless, for the purpose of this study, the Midwest region serves as an internally consistent unit for analyzing the region's core agribusiness logic, which operates under shared state-level policies and environmental conditions. Finally, further methodological refinements could explore alternative ground distances for the Wasserstein calculation, beyond the Shimbel matrix, possibly incorporating economic flows or other context-specific measures in order to enhance the capturing of nuanced geographical interactions.

In conclusion, by proposing a spatially aware clustering methodology, this Paper aspires to contribute to a powerful new lens through which to analyze regional economic structures. It moves beyond abstract measures of similarity by identifying economic activities with similar spatial distributions under an explicit geographical cost matrix. This distribution-based perspective does not directly measure economic linkages but it can support regional analysis by indicating where shared locational drivers, possible inter-industry connections, or policy-relevant spatial patterns may warrant further investigation.

Acknowledgments

G.S.R.J. gratefully acknowledges the financial support of CNPq (National Council for Scientific and Technological Development, Brazil), Grant #131029/2023-1. The Article Processing Charge for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614).

Funding

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, 131029/2023-1.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available in Brazilian Ministry of Labor through the Annual Social Inform at https://www.gov.br/trabalho-e-emprego/pt-br/acesso-a-informacao/acoes-e-programas/programas-projetos-acoes-obras-e-atividades/estatisticas-trabalho/rais/rais-2025/rais-2025. These data were derived from the following resources available in the public domain: RAIS, https://www.gov.br/trabalho-e-emprego/pt-br/acesso-a-informacao/acoes-e-programas/programas-projetos-acoes-obras-e-atividades/estatisticas-trabalho/rais/rais-2025/rais-2025.

Endnotes

[1]The data has been transformed as a percentage of the total to calculate the distances.

[2]The selection of economic activities analyzed in this Paper is based on Feix and Leusin Júnior (2016) list of economic activities related to the agribusiness sector.

References

Anselin, L. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27, no. 2: 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L. 2002. "Under the Hood: Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics* 27, no. 3: 247–267. https://doi.org/10.1016/S0169-5150(02)00077-4.

Anselin, L. 2019. "A Local Indicator of Multivariate Spatial Association: Extending Geary's C." *Geographical Analysis* 51: 133–150. https://doi.org/10.1111/gean.12164.

Anselin, L., and A. K. Bera. 1998. "Spatial Dependence in Linear Regression Models With an Introduction to Spatial Econometrics." In *Handbook of Applied Economic Statistics*, edited by A. Ullah and D. E. A. Giles. Marcel Dekker.

Aoun, A., O. Shetler, R. Raghuraman, G. A. Rodriguez, and S. A. Hussaini. 2024. "Beyond Correlation: Optimal Transport Metrics for Characterizing Representational Stability and Remapping in Neurons Encoding Spatial Memory." *Frontiers in Cellular Neuroscience* 17: 1273283. https://doi.org/10.3389/fncel.2023.1273283.

Arribas-Bel, D., and C. R. Schmidt. 2013. "Self-Organizing Maps and the US Urban Spatial Structure." *Environment and Planning B: Planning and Design* 40, no. 2: 362–371. https://doi.org/10.1068/b37014.

Bivand, R. S. 2010. "Exploratory Spatial Data Analysis." In *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, edited by M. Fischer and A. Getis. Springer. https://doi.org/10.1007/978-3-642-03647-7_13.

Boschma, R. 2017. "Relatedness as Driver of Regional Diversification: A Research Agenda." *Regional Studies* 51, no. 3: 351–364. https://doi.org/10.1080/00343404.2016.1254767.

Boschma, R., and S. Iammarino. 2009. "Related Variety, Trade Linkages, and Regional Growth in Italy." *Economic Geography* 85, no. 3: 289–311. https://doi.org/10.1111/j.1944-8287.2009.01034.x.

Brazilian Institute of Geography and Statistics (IBGE). 2024. *Pesquisa da Pecuária Municipal (PPM)*. IBGE. https://sidra.ibge.gov.br/acervo.

Brazilian Ministry of Labor. 2025. *Annual Social Information Report (RAIS): Microdata 2013–2021*. Ministry of Labor. https://www.gov.br/trabalho-e-emprego/.

Cantisani, G., G. Del Serrone, R. Mauro, P. Peluso, and A. Pompigna. 2024. "From Radar Sensor to Floating Car Data: Evaluating Speed Distribution Heterogeneity on Rural Road Segments Using Non-Parametric Similarity Measures." *Science* 6, no. 3: 52. https://doi.org/10.3390/sci6030052.

Chatterji, A., E. Glaeser, and W. Kerr. 2014. "Clusters of Entrepreneurship and Innovation." *Innovation Policy and the Economy* 14, no. 1: 129–166. https://doi.org/10.1086/674023.

Combes, P. P., and H. G. Overman. 2004. "The Spatial Distribution of Economic Activities in the European Union." In *Handbook of Regional and Urban Economics*, vol. 4, 2845–2909. Elsevier.

Coombes, M. G., A. E. Green, and S. Openshaw. 1986. "An Efficient Algorithm to Generate Official Statistical Reporting Areas: The Case of the 1984 Travel-To-Work Areas Revision in Britain." *Journal of the Operational Research Society* 37, no. 10: 943–953.

Cörvers, F., M. Hensen, and D. Bongaerts. 2009. "Delimitation and Coherence of Functional and Administrative Regions." *Regional Studies* 43, no. 1: 19–31. https://doi.org/10.1080/00343400701654103.

Dall'Agnol, A. 2016. *A Embrapa Soja no Contexto do Desenvolvimento da Soja no Brasil: Histórico e Contribuições*. Embrapa.

Delage, R., and T. Nakata. 2023. "Cluster Analysis of Energy Consumption Mix in the Japanese Residential Sector." *Smart Energy* 12: 100122. https://doi.org/10.1016/j.segy.2023.100122.

Delgado, M., M. E. Porter, and S. Stern. 2014. "Clusters, Convergence, and Economic Performance." *Research Policy* 43, no. 10: 1785–1799. https://doi.org/10.1016/j.respol.2014.05.007.

Delgado, M., M. E. Porter, and S. Stern. 2016. "Defining Clusters of Related Industries." *Journal of Economic Geography* 16, no. 1: 1–38. https://doi.org/10.1093/jeg/lbv017.

Duque, J. C., and S. J. Rey. 2008. "A Network Based Approach Towards Industry Clustering." In *The Economics of Regional Clusters*, edited by U. Blien and G. Maier, 41–68. Edward Elgar.

Duranton, G., and H. G. Overman. 2005. "Testing for Localization Using Micro-Geographic Data." *Review of Economic Studies* 72, no. 4: 1077–1106. https://doi.org/10.1111/0034-6527.00362.

El Malki, N., R. Cugny, O. Teste, and F. Ravat. 2020. "DECWA: Density-Based Clustering Using Wasserstein Distance." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2005–2008. Association for Computing Machinery.

Ellison, G., and E. Glaeser. 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy* 105: 889–927. https://doi.org/10.1086/262098.

Ellison, G., E. L. Glaeser, and W. R. Kerr. 2010. "What Causes Industry Agglomeration? Evidence From Coagglomeration Patterns." *American Economic Review* 100, no. 3: 1195–1213.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise." *KDD* 96, no. 34: 226–231.

Farmer, C. J., and A. S. Fotheringham. 2011. "Network-Based Functional Regions." *Environment and Planning A* 43, no. 11: 2723–2741.

Feix, R. D., and S. Leusin Júnior. 2016. *Estatísticas e Indicadores do Emprego Formal do Agronegócio: Nota Técnica*. FEE.

Feser, E. J., and E. M. Bergman. 2000. "National Industry Cluster Templates: A Framework for Applied Regional Cluster Analysis." *Regional Studies* 34, no. 1: 1–19.

Feser, E. J., S. H. Sweeney, and H. C. Renski. 2005. "A Descriptive Analysis of Discrete U.S. Industrial Complexes." *Journal of Regional Science* 45, no. 2: 395–419.

Frenken, K., F. Van Oort, and T. Verburg. 2007. "Related Variety, Unrelated Variety and Regional Economic Growth." *Regional Studies* 41, no. 5: 685–697. https://doi.org/10.1080/00343400601120296.

Fujita, M., P. R. Krugman, and A. Venables. 2001. *The Spatial Economy: Cities, Regions, and International Trade*. MIT press.

Gazzoni, D. L., and A. Dall'Agnol 2018. "A Saga da Soja: De 1050 a.C. a 2050 d.C." Brasília: Embrapa.

Getis, A. 2009. "Spatial Weights Matrices." *Geographical Analysis* 41, no. 4: 404–410. https://doi.org/10.1111/j.1538-4632.2009.00768.x.

Glaeser, E. L., H. D. Kallal, J. A. Scheinkman, and A. Shleifer. 1992. "Growth in Cities." *Journal of Political Economy* 100, no. 6: 1126–1152.

Griffith, D. A. 1996. "Some Guidelines for Specifying the Geographic Weights Matrix Contained in Spatial Statistical Models." In *Practical Handbook of Spatial Statistics*, edited by S. L. Arlinghaus, 65–82. CRC Press.

Grubesic, T. H., and A. T. Murray. 2006. "Vital Nodes, Interconnected Infrastructures, and the Geographies of Network Survivability." *Annals of the Association of American Geographers* 96, no. 1: 64–83.

He, R., X. Wu, Z. Sun, and T. Tan. 2019. "Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, no. 7: 1761–1773.

Hidalgo, C. A., P. A. Balland, R. Boschma, et al. 2018. "The Principle of Relatedness." In *Unifying Themes in Complex Systems IX. ICCS 2018. Springer Proceedings in Complexity*, edited by A. Morales, C. Gershenson, D. Braha, A. Minai, and Y. Bar-Yam. Springer.

Hidalgo, C. A., B. Klinger, A. L. Barabási, and R. Hausmann. 2007. "The Product Space Conditions the Development of Nations." *Science* 317, no. 5837: 482–487. https://doi.org/10.1126/science.1144581.

Irpino, A., R. Verde, and F. d. A. T. De Carvalho. 2014. "Dynamic Clustering of Histogram Data Based on Adaptive Squared Wasserstein Distances." *Expert Systems With Applications* 41, no. 7: 3351–3366. https://doi.org/10.1016/j.eswa.2013.12.001.

Kantorovitch, L. 1958. "On the Translocation of Masses." *Management Science* 5, no. 1: 1–4.

Kassambara, A. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Sthda.

Krugman, P. 1991. "Increasing Returns and Economic Geography." *Journal of Political Economy* 99, no. 3: 483–499.

Kumar, V., J. K. Chhabra, and D. Kumar. 2014. "Performance Evaluation of Distance Metrics in the Clustering Algorithms." *INFOCOMP Journal of Computer Science* 13, no. 1: 38–52.

Li, T., and J. Ma. 2020. "Functional Data Clustering Analysis via the Learning of Gaussian Processes With Wasserstein Distance." In *Lecture Notes in Computer Science*, vol. 12533, 393–403. Springer International Publishing.

Liu, T., Y. Tian, Y. Cao, et al. 2023. "Isoelectric Point Barcode and Similarity Analysis With the Earth Mover's Distance for Identification of Species Origin of Raw Meat." *Food Research International* 166: 112600. https://doi.org/10.1016/j.foodres.2023.112600.

Luxen, D., and C. Vetter. 2011. “Real-Time Routing With OpenStreetMap Data.” In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 513–516. Association for Computing Machinery.

MacQueen, J. 1967. “Some Methods of Classification and Analysis of Multivariate Observations.” In *The Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 281–297. Berkeley, CA: University of California Press.

Marshall, A. 1920. *Principles of Economics*. 8th ed. Macmillan.

McCann, P., and B. Fingleton. 1996. “The Regional Agglomeration Impact of Just-In-Time Input Linkages: Evidence From the Scottish Electronic Industry.” *Scottish Journal of Political Economy* 43, no. 5: 493–518. https://doi.org/10.1111/j.1467-9485.1996.tb00946.x.

Mendez del Villar, P., and C. M. Ferreira. 2005. “Dinâmicas Territoriais do Arroz de Terras Altas na Região Centro-Oeste do Brasil.” *Cadernos de Ciência & Tecnologia* 22, no. 1: 97–107.

Milligan, G. W., and M. C. Cooper. 1987. “Methodology Review: Clustering Methods.” *Applied Psychological Measurement* 11, no. 4: 329–354. https://doi.org/10.1177/014662168701100401.

Neffke, F., M. Henning, and R. Boschma. 2011. “How Do Regions Diversify Over Time? Industry Relatedness and the Development of New Growth Paths in Regions.” *Economic Geography* 87, no. 3: 237–265. https://doi.org/10.1111/j.1944-8287.2011.01121.x.

Oh, J. H., M. Pouryahya, A. Iyer, A. P. Apte, J. O. Deasy, and A. Tannenbaum. 2020. “A Novel Kernel Wasserstein Distance on Gaussian Measures: An Application of Identifying Dental Artifacts in Head and Neck Computed Tomography.” *Computers in Biology and Medicine* 120: 103731. https://doi.org/10.1016/j.compbiomed.2020.103731.

OpenStreetMap contributors. 2026. “OpenStreetMap [Data Set].” OpenStreetMap Foundation. Available as Open Data Under the Open Data Commons Open Database License (ODbL).

Peyré, G., and M. Cuturi. 2019. “Computational Optimal Transport.” *Foundations and Trends in Machine Learning* 11, no. 5–6: 355–607. https://doi.org/10.1561/2200000073.

Ponti, A., I. Giordani, M. Mistri, A. Candelieri, and F. Archetti. 2022. “The “Unreasonable” Effectiveness of the Wasserstein Distance in Analyzing Key Performance Indicators of a Network of Stores.” *Big Data and Cognitive Computing* 6, no. 4: 138.

Porter, M. E. 1998. “Clusters and Competition: New Agendas for Companies, Governments and Institutions.” In *On Competition*, 197–287. Boston, MA: Harvard Business School Press.

Porter, M. E. 2003. “The Economic Performance of Regions.” *Regional Studies* 37: 549–578.

Regal, A., J. Gonzalez-Feliu, and M. Rodriguez. 2023. “A Spatio-Functional Logistics Profile Clustering Analysis Method for Metropolitan Areas.” *Transportation Research Part E: Logistics and Transportation Review* 179: 103312. https://doi.org/10.1016/j.tre.2023.103312.

Rodriguez, A., and A. Laio. 2014. “Clustering by Fast Search and Find of Density Peaks.” *Science* 344, no. 6191: 1492–1496. https://doi.org/10.1126/science.1242072.

Schuhmacher, D., B. Bähre, N. Bonneel, et al. 2024. “Transport: Computation of Optimal Transport Plans and Wasserstein Distances.” R Package Version 0.15–0.

Shimbel, A. 1951. “Applications of Matrix Algebra to Communication Nets.” *Bulletin of Mathematical Biophysics* 13, no. 3: 165–178.

Silva, M. A., A. Stephan Nascente, A. C. Lanna, et al. 2022. “Sistema de Plantio Direto e Rotação de Culturas no Cerrado.” *Research, Society and Development* 11, no. 13: e376111335568.

Smart, M. W. 1974. “Labour Market Areas: Uses and Definition.” *Progress in Planning* 2: 239–353.

Van Oort, F. G., M. J. Burger, J. Knoben, and O. Raspe. 2012. “Multilevel Approaches and the Firm-Agglomeration Ambiguity in Economic Growth Studies.” *Journal of Economic Surveys* 26, no. 3: 468–491. https://doi.org/10.1111/j.1467-6419.2012.00723.x.

Vieira Filho, J. E. R., and J. G. Gasques. 2020. *Uma Jornada Pelos Contrastes do Brasil: Cem Anos do Censo Agropecuário*. IPEA, IBGE.

Villani, C. 2009. *Optimal Transport: Old and New*. Springer.

Ward, J. H., Jr. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58, no. 301: 236–244.

Zelasky, S., C. L. Martin, C. Weaver, L. K. Baxter, and K. M. Rappazzo. 2023. “Identifying Groups of Children’s Social Mobility Opportunity for Public Health Applications Using k-Means Clustering.” *Heliyon* 9, no. 9: e20250. https://doi.org/10.1016/j.heliyon.2023.e20250.

Zhang, S., Z. Wu, Z. Ma, X. Liu, and J. Wu. 2022. “Wasserstein Distance-Based Probabilistic Linguistic TODIM Method With Application to the Evaluation of Sustainable Rural Tourism Potential.” *Economic Research-Ekonomska Istraživanja* 35, no. 1: 409–437. https://doi.org/10.1080/1331677X.2021.1894198.

Zhuang, Y., X. Chen, and Y. Yang. 2022. “Wasserstein K-Means for Clustering Probability Distributions.” *Advances in Neural Information Processing Systems* 35: 11382–11395.

Appendix A

CNAE stands for Brazil’s National Classification of Economic Activities. CNAE codes have been harmonized across the period to account for classification changes. When equivalent activities appear under old and new CNAE codes, they are now treated as the same economic activity in order to avoid double-counting. Residual clusters refer to small clusters which are generated by the hierarchical clustering procedure, in turn reported for transparency; they have not been substantively interpreted in the main text.

TABLE A1 | List of economic activities grouped in the same cluster throughout the entire period (2013–2021).

CNAE code	Economic activity	No. of firms (mean)	Shimbel matrix	Travel time
0151–2/02	Raising of dairy cattle	8022	East	East
1099–6/99	Manufacture of other food products n.e.s.	307	East	Spread
1531–9/01	Manufacture of leather footwear	246	East	East
1095–3/00	Manufacture of spices, sauces, seasonings and condiments	154	East	Spread
1529–7/00	Manufacture of leather artifacts n.e.s.	144	East	Spread
1092–9/00	Manufacture of cookies and crackers	133	East	Spread
0131–8/00	Growing of orange	73	East	East
0142–3/00	Production of certified seedlings and other forms of plant propagation	56	East	Spread
1731–1/00	Manufacture of paper packaging	56	East	Spread
1032–5/99	Manufacture of canned vegetables, except palm hearts	55	East	Spread
0119–9/09	Growing of tomato	42	East	East
0134–2/00	Growing of coffee	31	East	Spread
1742–7/01	Manufacture of disposable diapers	28	East	—
2122–0/00	Manufacture of veterinary medicines	25	East	Spread
0133–4/04	Growing of citrus fruits, except orange	24	East	Spread
0119–9/05	Growing of bean	20	East	East
0119–9/03	Growing of potato	14	East	—
0119–9/02	Growing of garlic	12	East	East
0139–3/04	Growing of spice plants, except black pepper	12	East	—
0119–9/04	Growing of onion	8	East	East
1732–0/00	Manufacture of paperboard and cardboard packaging	8	East	Spread
1322–7/00	Weaving of natural textile fibers, except cotton	7	East	East
2012–6/00	Manufacture of fertilizer intermediates	6	East	—
0133–4/10	Growing of mango	3	East	East
0321–3/99	Saltwater and brackish water aquaculture cultivation n.e.s.	3	East	East
1043–1/00	Manufacture of margarine, other vegetable fats, and non-edible animal oils	3	East	—
1742–7/02	Manufacture of sanitary napkins/pads	3	East	East
0311–6/03	Collection of other marine products	2	East	East
0321–3/03	Saltwater and brackish water oyster and mussel farming	2	East	East
1099–6/06	Manufacture of natural and artificial sweeteners	2	East	East
1610–2/01	Sawmilling with wood splitting	1364	North	North
1610–2/03	Sawmilling with raw wood splitting	1215	North	North
0220–9/01	Logging in native forests	283	North	North
0210–1/04	Growing of teak	78	North	North
0133–4/06	Growing of guarana	1	North	—
0111–3/99	Growing of other cereals n.e.s.	1635	South	South
0113–0/00	Growing of sugarcane	781	South	Spread
0154–7/00	Raising of pigs	457	South	South
0210–1/01	Growing of eucalyptus	359	South	South
0220–9/02	Production of charcoal—native forests	161	South	South
0141–5/02	Production of certified forage seeds for pasture training	125	South	Spread
0210–1/08	Production of charcoal—planted forests	121	South	South
0119–9/06	Growing of cassava	60	South	South
1065–1/01	Manufacture of vegetable starches	54	South	South
1071–6/00	Manufacture of raw sugar	51	South	Spread
1710–9/00	Manufacture of cellulose and other pulp for papermaking	23	South	South
0152–1/03	Raising of donkeys and mules	22	South	South
1099–6/05	Manufacture of infusion products (tea, mate, etc.)	15	South	South
4634–6/99	Wholesale of meat and other animal products	10	South	East
0220–9/99	Collecting non-timber products n.e.s. in native forests	7	South	—
0139–3/02	Growing of yerba mate	4	South	South

(Continues)

TABLE A1 | (Continued)

CNAE code	Economic activity	No. of firms (mean)	Shimbel matrix	Travel time
0322–1/06	Raising of alligators	2	South	
1210–7/00	Industrial processing of tobacco	2	South	—
0119–9/07	Growing of melon	1	South	South
0133–4/07	Growing of apple	1	South	—
1065–1/02	Manufacture of raw corn oil	1	South	—
1099–6/03	Production of yeast and ferment	1	South	South
1722–2/00	Manufacture of cardboard and paperboard	1	South	—
4623–1/07	Wholesale of sisal	1	South	South
0151–2/01	Raising of cattle for meat	38,529	Spread	South
0115–6/00	Growing of soybean	13,378	Spread	Spread
0161–0/99	Support activities for agriculture n.e.s.	2072	Spread	Spread
3101–2/00	Manufacture of furniture predominantly made of wood	2007	Spread	Spread
0151–2/03	Raising of cattle, except for meat and milk	1613	Spread	Spread
0162–8/99	Support activities for livestock n.e.s.	1603	Spread	Spread
4683–4/00	Wholesale trade of agricultural pesticides, fertilizers, fertilizers and soil improvers	1398	Spread	North
4617–6/00	Commercial representatives and agents for food, beverages, and tobacco products	1393	Spread	Spread
4611–7/00	Sales representatives and agents trading in agricultural raw materials and live animals	1381	Spread	Spread
0161–0/03	Services of land preparation, farming and harvesting	1204	Spread	Spread
1091–1/02	Manufacture of bakery and confectionery products with predominant in-house production	1110	Spread	Spread
4639–7/01	Wholesale trade of food products in general	916	Spread	Spread
4661–3/00	Wholesale trade of machinery, devices and equipment for agricultural use; parts and pieces	902	Spread	North
0155–5/01	Raising of chickens for slaughter	699	Spread	Spread
1066–0/00	Manufacture of animal feed	624	Spread	Spread
4632–0/01	Wholesale of processed cereals and pulses	578	Spread	Spread
1091–1/01	Manufacture of industrial bakery products	532	Spread	Spread
1052–0/00	Manufacture of dairy products	528	Spread	Spread
0111–3/02	Growing of corn	516	Spread	—
4622–2/00	Wholesale of soybean	467	Spread	North
7731–4/00	Rental of agricultural machinery and equipment without operator	466	Spread	Spread
0121–1/01	Horticulture, except strawberries	444	Spread	Spread
1053–8/00	Manufacture of ice cream and other edible frozen desserts	429	Spread	Spread
4623–1/06	Wholesale of seeds, flowers, plants and grasses	417	Spread	Spread
0119–9/99	Growing of other temporary crops n.e.s.	412	Spread	Spread
4623–1/99	Wholesale trade of agricultural raw materials n.e.s.	406	Spread	Spread
0161–0/01	Spraying and pest control services	405	Spread	Spread
4671–1/00	Wholesale trade of wood and wood products	394	Spread	North
4692–3/00	Wholesale trade of goods in general, with a predominance of agricultural inputs	356	Spread	Spread
0230–6/00	Support activities for forest production	351	Spread	Spread
4623–1/09	Wholesale of animal feed	348	Spread	Spread
0111–3/01	Growing of rice	344	Spread	Spread
1011–2/01	Slaughterhouse—slaughter of cattle	326	Spread	South
1629–3/01	Manufacture of various wood artifacts, except furniture	291	Spread	Spread
4691–5/00	Wholesale trade of general merchandise, with a predominance of food products	284	Spread	Spread
1061–9/01	Processing of rice	278	Spread	Spread
4634–6/01	Wholesale trade of beef and pork meat and by-products	266	Spread	Spread
4637–1/99	Specialized wholesale trade of other food products n.e.s.	264	Spread	Spread

(Continues)

TABLE A1 | (Continued)

CNAE code	Economic activity	No. of firms (mean)	Shimbel matrix	Travel time
0322–1/01	Freshwater fish farming	246	Spread	Spread
0112–1/01	Growing of herbaceous cotton	240	Spread	North
0210–1/07	Extraction of wood from planted forests	235	Spread	—
1013–9/01	Manufacture of meat products	234	Spread	Spread
0163–6/00	Post-harvest activities	212	Spread	Spread
0159–8/99	Raising of other animals n.e.s.	209	Spread	Spread
2013–4/00	Manufacture of fertilizers	199	Spread	Spread
0155–5/05	Production of eggs	173	Spread	South
1081–3/02	Roasting and grinding of coffee	163	Spread	Spread
0141–5/01	Production of certified seeds, except for forage for pasture	162	Spread	Spread
1051–1/00	Processing of milk	162	Spread	Spread
1610–2/02	Sawmilling without wood resawing	159	Spread	North
1621–8/00	Manufacture of laminated wood and plywood, pressed and particle boards	156	Spread	North
1069–4/00	Milling and manufacture of vegetable products n.e.s.	155	Spread	North
2833–0/00	Manufacture of machinery and equipment for agriculture and livestock, parts and accessories, except for irrigation	155	Spread	Spread
2013–4/02	Manufacture of fertilizers and fertilizers, except organomineral	151	Spread	Spread
1622–6/02	Manufacture of wooden door and window frames and wood parts for industrial and commercial installations	149	Spread	North
0162–8/03	Service of animal handling	148	Spread	Spread
4639–7/02	Wholesale trade of food products in general, with associated portioning and packaging activities	145	Spread	Spread
1099–6/04	Manufacture of common ice	141	Spread	Spread
4632–0/03	Wholesale of processed cereals and pulses, flours, starches and starches, with associated fractioning and packaging activity	141	Spread	—
0139–3/06	Growing of rubber tree	140	Spread	Spread
1311–1/00	Cotton fiber preparation and spinning	137	Spread	North
4623–1/01	Wholesale trade of live animals	135	Spread	—
1622–6/99	Manufacture of other joinery articles for construction	132	Spread	Spread
4631–1/00	Wholesale trade of milk and dairy products	130	Spread	Spread
1041–4/00	Manufacture of crude vegetable oils, except corn oil	122	Spread	Spread
0152–1/02	Raising of horses	117	Spread	Spread
0155–5/04	Raising of poultry, except gallinaceous animals	115	Spread	Spread
1031–7/00	Manufacture of canned fruits	115	Spread	Spread
4623–1/08	Wholesale of agricultural raw materials with associated fractioning and packaging activity	113	Spread	Spread
1610–2/04	Sawmilling without raw wood resawing—Resawing	108	Spread	North
1012–1/01	Poultry slaughter	105	Spread	Spread
1510–6/00	Tanning and other preparations of leather	105	Spread	South
0116–4/99	Growing of other temporary oilseed crops n.e.s.	104	Spread	South
0210–1/06	Growing of seedlings in forest nurseries	92	Spread	—
4637–1/04	Wholesale trade of bread, cakes, cookies, and similar items	82	Spread	Spread
1063–5/00	Manufacture of cassava flour and derivatives	80	Spread	South
4623–1/03	Wholesale trade of cotton	78	Spread	North
1013–9/02	Preparation of slaughter by-products	75	Spread	—
4634–6/03	Wholesale trade of fish and seafood	72	Spread	Spread
2013–4/01	Manufacture of organomineral fertilizers and fertilizers	67	Spread	Spread
1113–5/02	Manufacture of beers and draught beers	66	Spread	Spread
0112–1/99	Growing of other temporary fiber crops n.e.s.	65	Spread	—
0133–4/99	Growing of permanent fruit crops n.e.s.	64	Spread	Spread
1111–9/01	Manufacture of sugarcane spirit	64	Spread	Spread
4623–1/02	Wholesale trade of hides, wool, skins and other inedible by-products of animal origin	64	Spread	Spread

(Continues)

CNAE code	Economic activity	No. of firms (mean)	Shimbel matrix	Travel time
0133–4/02	Growing of banana	63	Spread	Spread
4637–1/01	Wholesale trade of roasted, ground, and instant coffee	62	Spread	Spread
4633–8/02	Wholesale trade of live poultry and eggs	60	Spread	Spread
1011–2/05	Slaughterhouse—slaughter of cattle under contract, except slaughter of pigs	59	Spread	Spread
4632–0/02	Wholesale trade of flours, starches, and meals	59	Spread	Spread
4637–1/07	Wholesale trade of chocolates, confectionery, candies, bonbons, and similar items	59	Spread	Spread
1064–3/00	Manufacture of corn flour and byproduct, except corn oil	57	Spread	Spread
4637–1/06	Wholesale trade of ice cream	56	Spread	Spread
0162–8/01	Service of artificial insemination for animals	54	Spread	—
1623–4/00	Manufacture of cooperage artifacts and wood packaging	54	Spread	Spread
1012–1/03	Slaughterhouse—slaughter of pigs	46	Spread	Spread
0152–1/01	Raising of buffaloes	42	Spread	—
0312–4/01	Freshwater fishing	42	Spread	—
0155–5/03	Raising of other gallinaceous animals, except for slaughter	40	Spread	Spread
0159–8/01	Beekeeping	40	Spread	South
1932–2/00	Manufacture of biofuels, except alcohol	40	Spread	North
1122–4/01	Manufacture of soft drinks	39	Spread	—
0153–9/02	Raising of sheep, including for wool production	37	Spread	Spread
0161–0/02	Tree pruning services for crops	36	Spread	North
4636–2/02	Wholesale trade of cigarettes, cigarillos, and cigars	33	Spread	Spread
4634–6/02	Wholesale trade of slaughtered poultry and by-products	32	Spread	Spread
0155–5/02	Production of one-day-old chicks	30	Spread	Spread
0159–8/02	Raising of pets	25	Spread	Spread
1111–9/02	Manufacture of other spirits and distilled beverages	25	Spread	Spread
4621–4/00	Wholesale trade of coffee beans	22	Spread	—
1020–1/02	Manufacture of preserved fish, crustaceans, and mollusks	20	Spread	Spread
1042–2/00	Manufacture of refined vegetable oils, except corn oil	19	Spread	Spread
2051–7/00	Manufacture of agricultural pesticides	19	Spread	—
0133–4/05	Growing of coconut	18	Spread	Spread
1742–7/99	Manufacture of paper products for household and sanitary use n.e.s.	18	Spread	—
4636–2/01	Wholesale trade of processed tobacco	18	Spread	Spread
1020–1/01	Preservation of fish, crustaceans, and mollusks	17	Spread	—
0119–9/01	Growing of pineapple	16	Spread	Spread
1061–9/02	Manufacture of rice products	16	Spread	Spread
0153–9/01	Raising of goats	15	Spread	—
0220–9/06	Conservation of native forests	15	Spread	Spread
1032–5/01	Manufacture of canned heart of palm	15	Spread	North
0119–9/08	Growing of watermelon	14	Spread	Spread
0132–6/00	Growing of grape	14	Spread	
0220–9/04	Latex collection in native forests	14	Spread	Spread
0311–6/04	Support activities for saltwater fishing	6	Spread	—
0159–8/03	Raising of snails	1	Spread	—

Note: CNAE stands for Brazil's National Classification of Economic Activities. Economic activities without clusters in the travel time analysis were not grouped in the same cluster throughout the entire period of analysis. During the period, three historical CNAE codes were replaced or reclassified in the CNAE update: 1610–2/01 was replaced by 1610–2/03; 1610–2/02 was replaced by 1610–2/04; and 2013–4/00 was replaced by the revised fertilizer categories 2013–4/01 and 2013–4/02.

TABLE A2 | List of economic activities grouped in different clusters throughout the period (2013–2021).

CNAE code	Economic activity	No. of firms (mean)	Travel time
4637–1/05	Wholesale trade of pasta	23	Spread
4637–1/03	Wholesale trade of oils and fats	25	Spread
4637–1/02	Wholesale trade of sugar	21	Spread
4633–8/03	Wholesale trade of rabbits and other small live animals for food	1	—
4633–8/01	Wholesale trade of fresh fruits, greens, roots, and vegetables	690	Spread
4623–1/05	Wholesale trade of cocoa	1	—
4623–1/04	Wholesale trade of unprocessed leaf tobacco	2	—
2832–1/00	Manufacture of agricultural irrigation equipment, parts, and accessories	6	—
2831–3/00	Manufacture of agricultural tractors, parts, and accessories	4	—
1931–4/00	Manufacture of alcohol	144	Spread
1749–4/00	Manufacture of pulp and paper products n.e.s.	32	Spread
1741–9/02	Manufacture of paper products for commercial and office use	53	Spread
1741–9/01	Manufacture of continuous forms	11	—
1733–8/00	Manufacture of corrugated cardboard sheets and packaging	23	Spread
1721–4/00	Manufacture of paper	20	South
1629–3/02	Manufacture of miscellaneous cork, bamboo, straw, and wicker articles	21	Spread
1622–6/01	Manufacture of prefabricated wooden houses	13	—
1610–2/05	Services of wood treatment performed under contract	6	—
1531–9/02	Finishing of leather footwear under contract	5	—
1312–0/00	Preparation and spinning of natural textile fibers, except cotton	6	—
1220–4/99	Manufacture of other tobacco products (except cigarettes/cigars)	6	—
1220–4/03	Manufacture of cigarette filters	1	—
1220–4/02	Manufacture of cigarillos and cigars	3	—
1220–4/01	Manufacture of cigarettes	6	—
1122–4/99	Manufacture of other non-alcoholic beverages n.e.s.	5	—
1122–4/04	Manufacture of isotonic beverages	2	—
1122–4/03	Manufacture of soft drinks, syrups, and powders for soft drinks	10	—
1122–4/02	Manufacture of yerba mate and other ready-to-drink teas	3	—
1113–5/01	Manufacture of malt, including whiskey malt	2	East
1112–7/00	Manufacture of wine	7	—
1099–6/07	Manufacture of dietary foods and food supplements	4	—
1099–6/02	Manufacture of food powders	7	—
1099–6/01	Manufacture of vinegars	9	Spread
1096–1/00	Manufacture of ready-to-eat meals and dishes	200	Spread
1094–5/00	Manufacture of pasta	167	Spread
1093–7/02	Manufacture of crystallized fruits, candies, and similar products	19	Spread
1093–7/01	Manufacture of cocoa products and chocolates	41	Spread
1082–1/00	Manufacture of coffee-based products	4	—
1081–3/01	Processing of coffee	13	North
1072–4/01	Manufacture of refined cane sugar	5	—
1062–7/00	Milling of wheat and manufacture of derivatives	28	Spread
1033–3/02	Manufacture of fruit and vegetable juices, except concentrates	34	Spread
1033–3/01	Manufacture of concentrated fruit and vegetable juices	33	Spread
1012–1/04	Slaughterhouse – slaughtering of pig under contract	5	Spread
1012–1/02	Slaughtering of small animals	33	—
1011–2/03	Slaughterhouse—slaughtering of sheep and goat	5	—
1011–2/02	Slaughterhouse—slaughtering of equine	2	—
0322–1/99	Freshwater aquaculture crops and semi-crops n.e.s.	8	—
0322–1/07	Support activities for freshwater aquaculture	16	Spread
0322–1/05	Frog farming	7	—
0322–1/04	Freshwater ornamental fish farming	4	—
0322–1/02	Freshwater shrimp farming	2	—

(Continues)

TABLE A2 | (Continued)

CNAE code	Economic activity	No. of firms (mean)	Travel time
0321–3/05	Support activities for aquaculture in salt and brackish water	5	—
0321–3/04	Ornamental fish farming in salt and brackish water	1	—
0321–3/02	Shrimp farming in salt and brackish water	2	—
0321–3/01	Fish farming in salt and brackish water	6	—
0312–4/04	Support activities for freshwater fishing	21	—
0312–4/03	Harvesting of other freshwater aquatic products	3	—
0312–4/02	Freshwater crustacean and mollusk fishing	2	—
0311–6/02	Saltwater crustacean and mollusk fishing	2	—
0311–6/01	Saltwater fishing	9	—
0220–9/05	Harvesting of heart of palm in natural forests	4	—
0220–9/03	Harvesting of Brazil nut in natural forests	6	Residual cluster
0210–1/99	Production of non-wood products n.e.s. in planted forests	10	—
0210–1/09	Production of black wattle bark – planted forests	1	—
0210–1/05	Growing of timber species (except eucalyptus, black wattle, pine, and teak)	19	South
0210–1/03	Growing of pine	10	East
0210–1/02	Growing of black wattle	2	East
0170–9/00	Services of hunting	4	—
0162–8/02	Services of sheep shearing	2	—
0159–8/04	Raising of silk worms	1	—
0139–3/99	Growing of other permanent crops n.e.s.	124	Spread
0139–3/05	Growing of oil palm	1	East
0139–3/03	Growing of black pepper	1	—
0139–3/01	Growing of tea plant	1	East
0135–1/00	Growing of cocoa	2	—
0133–4/09	Growing of passion fruit	7	Spread
0133–4/08	Growing of papaya	8	—
0133–4/03	Growing of cashew	2	—
0133–4/01	Growing of açaí	2	—
0122–9/00	Growing of flowers and ornamental plants	95	Spread
0121–1/02	Growing of strawberry	3	—
0116–4/03	Growing of castor bean	2	—
0116–4/02	Growing of sunflower	2	—
0116–4/01	Growing of peanut	3	—
0114–8/00	Growing of tobacco	5	—
0112–1/02	Growing of jute	10	—
0111–3/03	Growing of wheat	19	—

Note: CNAE: Brazil's National Classification of Economic Activities.

TABLE A3 | Descriptive statistics of economic activities firms' size by cluster.

	Number of exclusive economic activities			
Cluster	Shimbel matrix	Travel time	Economic activities in the same cluster	Economic activities in different clusters
Spread	127	122	86	22
South	24	23	14	4
East	30	20	14	11
North	5	19	4	0
Total	187	184	118	37

Source: Authors' own elaboration, based on the research data.