2021, Volume 44, Issue 87, 56-73 / ISSN 2304-4306

E C O N O M Í A

revistas.pucp.edu.pe/economia

FONDO EDITORIAL

www.fondoeditorial.pucp.edu.pe

Empirical Identification of Intra-Urban Subcentralities: A New Methodological Approach with an Application for a Developing Country

Rodger B. A. Campos[a,*], Carlos Azzoni[b]

[a]Regional and Urban Economimcs Lab (NEREUS), University of São Paulo

✉ rodgercampos@hotmail.com * Corresponding author

[b]Faculty of Economics, University of São Paulo

✉ cazzoni@usp.br

Abstract

We present a new empirical approach for identifying sub-centers within urban areas and apply it to the São Paulo metropolitan area (SPMA). We use geographically weighted regressions (GWR) to overcome the limitations presented by previous methods, which rely on previous knowledge of the employment distribution and use arbitrary threshold values and band sizes. We find three SBD in 2002 and only two in 2014, suggesting that SPMA is polycentric but presents only one business core that concentrates more than 90% of all employees working in an SBD. We apply the widely recognized method of McMillen and Smith (2003) to our database and compare the results. Our method is more conservative in identifying areas as sub-centers (SBD) and presents lower standard errors.

Article History: Received: May 21 2020 / Revised: July 01 2020 / Accepted: January 20 2021

Keywords: Urban economics; Spatial econometrics; Urban centralities

JEL Classification: C21, R32

https://doi.org/10.18800/economia.202101.004

1. Introduction

An extensive literature discusses the role of the Central Business Districts (CBD) and Subcenter Business Districts (SBD) areas in the urban economy and how these areas affect wages, land prices, and the transportation system. These are areas with high employment concentration, based on the premium generated by agglomeration economies (Anas et al., 1998). In the traditional models, a monocentric city arises from the interactions between firms and families. Firms consume land for production, and families consume land for housing, facing commuting costs. Empirical results derived from monocentric models identify the CBD as the area with the highest land prices and wages, as compared to any other intra-urban region, and these prices decrease with distance to the CBD (Alonso, 1964; Beckmann, 1974; Muth, 1969, 1975; Mills, 1967, 1972; Solow, 1973; Wheaton, 1974).

As the city grows, suburbanization of employment gives rise to multiple centers or SDB. The SDB can be defined as a place outside the historical center (typically identified as the CDB) where production, purchases, and employment are concentrated. Usually, it is served by a sound transportation system that provides accessibility to inputs and products (Ogawa and Fujita, 1980; Fujita and Ogawa, 1982; Helsley and Sullivan, 1991; McMillen, 2001b). Rents and wages are higher in the SDB than anywhere else but the CBD, due to the supply of infrastructure and proximity to the workplaces (Papageorgiou, 1971; White, 1976, 1988, 1999; Hartwick and Hartwick, 1974; Romanos, 1979; Sullivan, 1986; Wieand, 1987; Sivitanidou and Wheaton, 1992; Hotchkiss and White, 1993; Yinger, 1992; Ross and Yinger, 1995; Wrede, 2015). The SBD is a region with a concentration of firms with strong productive and communication linkages (Castells et al., 1994; Anas and Kim, 1996; Sasaki and Mun, 1996; Graham and Marvin, 1996). High-order activities, public and private administrations, and the central transportation system tend to cluster in these areas. These are activities that require face-to-face contact and create back-and-forth connections.

The suburbanization process implies a reduction in the fraction of population and employment located in the CBD, as economic activities and population increase in other areas. New centralities, or subcenters, appear as subsidiaries of the historic city center (Anas et al., 1998). The suburbanization of activities can alter spatial prices and reverse roles between the CBD and some SBD. White (1999) divides the theoretical literature on polycentric cities into two main categories.[1] The first deals with the dispute between the CBD and the SBD as an endogenous problem of agglomeration economies and transportation costs, in a general equilibrium approach (Fujita and Ogawa, 1982; Fujita, 1988; Helsley and Sullivan, 1991; Henderson and Slade, 1993; Anas and Kim, 1996; Lucas and Rossi-Hansberg, 2002; Ahlfeldt et al., 2016). The costs involved in commuting and the transportation of goods make the supply of labor endogenous due to the simultaneous locational choice of families and firms. However, the fundamental factors that cause the congregation of firms in the CBD or SBD are external agglomeration economies and transportation costs of goods, which are larger than the cost of commuting (White, 1999).

The second category of models assumes the existence of a CDB and SBD exogenously. The

[1] See Campos and Azzoni (2020) for a deeper discussion on theoretical and empirical approaches.

existence of one or more sub-centers results from the decentralization of economic activity and is taken as given (White, 1976, 1988, 1999; Romanos, 1979; Sullivan, 1986; Wieand, 1987; Sivitanidou and Wheaton, 1992; Papageorgiou, 1971; Hartwick and Hartwick, 1974; Hotchkiss and White, 1993; Yinger, 1992; Ross and Yinger, 1995; Henderson and Mitra, 1996; Zhang and Komei, 1997, 2000; Wrede, 2015). These models' focus is to evaluate how households and workers decide on where to live and work and the resulting spatial pattern of land prices, population density, and commuting. Empirical results based on these studies indicate a decay of land rents and wages as the distance to the CBD and SBD increases.

Different factors might encourage the spread of employment outside the CBD, such as taxes and land use policies (Sullivan, 1986; Zhang and Komei, 1997, 2000; Henderson and Mitra, 1996).

The empirical identification of SDB is still an open question, as the theoretical models do not provide empirically testable propositions. This study presents a new methodological approach to identify subcentralities, overcoming many of the limitations presented in most empirical studies, such as prior knowledge of the region and a predefined *cutoff approach*. Our approach is based on a bottom-up method that allows the data to determine the SBD location. To provide a means for comparison, we replicate McMillen and Smith's (2003) top-down approach with our data. The comparison of the results indicates how our method is useful to overcome the limitations of their method. Besides this contribution, we also provide essential knowledge about CDB and SDB determination at a highly disaggregated spatial scale in a major metropolitan area in a developing country.

The comparison of results reveals that our findings are more conservative: we identify 91 cells as CBD or SBD (out of 9,071), while their approach identifies 119 cells and present larger standard errors. Using a Probit model, we find that the probability that a cell belongs to an SBD is also larger in their approach. These empirical findings indicate that the spatial structure in the SPMA is concentrated around the CBD area in the city of São Paulo, and the SBD areas are localized in other municipalities. We also observe a reduction in the SDB areas in the period 2002-2014.

This paper is organized as follows. In section 2, we present a discussion on SBD identification strategies. Section 3 presents our methodological approach to SBD identification, using Geographically Weighted Regression. In section 4, we present the database. In section 5 we present the main results of both methodologies and compare the results. We present the conclusions in section 6.

2. Procedures for the Identification of Subcenters

All empirical strategies to identify subcenters face limitations. Bender and Hwang (1985), Heikkila et al. (1989), and Richardson et al. (1990) used prior knowledge of the region as a starting point, but their results were not consistent (McMillen, 2001b). The *cutoff approach* most discussed in the literature was proposed by McDonald (1987), and extended by Giuliano and Small (1991).[2] In a study about the Los Angeles metro region, a subcenter was defined as a set of contiguous areas with a minimum density of 10 employees per acre, and at least 10,000

[2] Greene (1980) had previously presented a threshold approach.

employees. Important limitations of the study have been raised, as the *ad hoc* choice of the threshold value, the sensitivity of the results to the cutoff choice (McMillen, 2001a, 2003) the size of the area (McMillen, 2001a,b), and the impossibility of generalization to other cities (Anas et al., 1998; McMillen, 2001b).

Some empirical studies have already estimated an employment density function for a monocentric city.[3] McDonald (1987) and Heikkila et al. (1989) are pioneers in the SBD identification using parametric econometric models. McDonald (1987) identifies as SBD areas with positive and statistically significant residuals. However, since no control variables are used, the residuals may contain many other effects that underestimate the predicted employment density.[4] Heikkila et al. (1989) use the land prices as the dependent variable and control for many other covariates. The estimated coefficients with decreasing gradients identify the SBD. Their approach demands previous knowledge of SBD candidates and data on land prices and amenities, which are not so easy to obtain. Despite their limitations, these studies provide a step forward, as they use statistical inference and seek to accommodate the empirical approach to theoretical models.

Craig and Ng (2001) examined the shape of the employment density function using quantile-smoothing splines as a nonparametric specification. Their non-parametric approach is not sensitive to the unit of analysis, but the estimation of a population density function and its symmetric employment around the CBD can generate biased results if no symmetry was observed (McMillen, 2001b). Since it is based on cutoff values, Redfearn (2007) points to the sensitivity of the results and the absence of statistical inference.

McMillen (2001a,b) proposed a two-stage model that requires no prior knowledge of the city. In the first stage, the logarithm of employment density is regressed on the distance to the CBD, though using locally weighted regression (LWR), as in McDonald (1987) and Craig and Ng (2001). SBD candidates are those whose estimated residuals create statistically significant clusters. In the second stage, they estimate a semiparametric model considering just the potential sub-centers previously identified. The nonparametric part of the regression uses the first-stage residuals as the dependent variable and the distance to the CBD as the explanatory variable. The negative and statistically significant coefficients are labeled as SBD. This approach's main limitation is that only the SDBs defined in the first stage are used in the second stage, leaving aside all other areas of the city. This may bias the results and create false positives due to the creation of an artificial city pattern. Redfearn (2007) also applies nonparametric models, using the employment density surface. Areas with local maxima are the candidates to subcenters since the estimates are local. LWR is used to estimate a neighborhood of employment density around the candidate for subcentrality. The estimates involve three steps: estimation of global peaks, estimation of peaks relative to neighbors, and bootstrapping for statistical inference.

McMillen (2003) presents a hybrid method using LWR, combining the approaches of Giuliano and Small (1991) and McMillen (2001a). Subcenters are contiguous areas with at least 10,000 employees and positive and significant residuals. Although such an approach brings one of the

[3] $D(x) = D_0 e^{-\beta x + u}$, where D is employment or population density, x is the distance to the CBD, D_0 is the mean employment or population density at distance zero from the CBD, β is the density gradient (decay measure), and u is the random error term.

[4] Heikkila et al. (1989) and McDonald and Prather (1994) used the same approach.

two arbitrary elements of Giuliano and Small (1991), the author states that this threshold is less arbitrary than the minimum employment density. In McMillen and Smith (2003), the fixed threshold is replaced by a cutoff value derived from the distribution of the estimated errors in the first stage: areas with residuals lower or equal to 5% are selected as SBDs. In practice, the study uses geographic coordinates of the target tracts as exploratory variables for the weighted least square regression. LWR provides an estimate of total employment by tract y_i, at latitude x_{1i} and longitude x_{2i}. A tricubic function is used as a kernel, and its window size is 50% for all cities. The authors argue that changes in the window size do not affect the number of identified SDB, even in the face of a vast literature that points to its effects on the outcome and standard errors (Fotheringham et al., 2000, 2002). In the second stage, a subcenter is defined as a group of sites from the first stage that are contiguous and for which total employment exceeds 10,000.[5]

Kane et al. (2016) applied this methodology for the identification of SBDs in the Los Angeles region. In the first stage, they used LWR and 120 nearest neighbors. They identify as an employment center a contiguous region surrounding each local maximum where each member had a higher employment density than its neighbors. Although the minimum density cutoff is avoided, the reliance on contiguity matrices of arbitrary order is a problem. None of these three studies deals with the problem of determining the best number of neighbors to consider in the process of SBD identification.

All econometric approaches seek objective criteria to identify SDBs. However, basing the choices on the estimated residuals may lead to skewed results, as not all deviations from the estimated errors may identify subcenters. Even with improvements in the choice of the cutoff, the estimated residuals may identify any deviation not controlled by the geographical smoothness. The next section exposes a new methodological approach that deals with cutoff rules and window sizes.

3. A New Approach

The standard procedure for modeling polycentric areas is the classical econometric approach, but it presents limitations. An incorrectly determined functional form tends to produce biased results (McMillen and McDonald, 1997). Parametric estimators have low power in complex metropolitan areas, due to the specification of the error term when the candidates to subcenters are spatially correlated and/or heterogeneous (Giuliano et al., 2005). Given these limitations, LWR have been repeatedly used, as shown above. Non-parametric estimators offer significant advantages over the simple linear regression method. They are flexible, allowing the estimated coefficient to vary between areas. They provide greater accuracy for the subcenters in the central area of the city and the peripheral sub-centers. Besides, they allow map visualization (McMillen and McDonald, 1997; McMillen, 2001b; Redfearn, 2007). Given these advantages, we work with

[5]The LWR provides an estimate of $y = g(x_{1i}, x_{2i})$. The spatial pattern of the model is based on a Tricubic function, such as $K_{ij} = \left(1 - \left(\frac{d_{ij}}{d_i^*}\right)^3\right)^3 I(d_{ij} < d_i^*)$, where $I(\cdot)$ is equals 1 when the condition is true, d_{ij} is the distance between observation j and the target point i, and d_i^* is the distance between observation i and the most distant observation, given the weight in the regression. The predicted value of y at the target point is the predicted value from the simple weight least squares regression: $\hat{y}_i = (\sum_{j=1}^n K_{ij} x_i' x_i)^{-1} \sum_{j=1}^n K_{ij} x_j' y_j$.

Geographically Weighted Regressions (GWR), which is one of the variants of the LWR family.

The choice of the areas' geographical dimension is crucial to the results since it influences the measured density. Mieszkowski and Smith (1991) point out that the gross density tends to overestimate the coefficients since areas further away from the historical center tend to concentrate larger undeveloped land areas. Redfearn (2007) argue that aggregate data in large areas mask concentrations of independent local jobs. To deal with such limitations, we divide the area into one-km^2 cells and allocate all workers to their centroids. At such a fine geographical scale, the above problems are much less of a concern.

Our basic equation for employment density is

$$e_i = \beta_0(u_i, v_i) + \varepsilon_i, \tag{1}$$

where e_i is the number of employees in cell i; (u_i, v_i) are the geographic coordinates (latitude and longitude, respectively) of the i^{th}cell centroid, $\beta_0(u_i, v_i)$ is the kernel-estimated employment average at point i, and ε_i is the random error term. The GWR estimator for $\beta_0(u_i, v_i)$ is given by

$$\hat{\beta}_0(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i), \tag{2}$$

$$W(i) = \begin{bmatrix} w_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{in} \end{bmatrix}, \tag{3}$$

where $\hat{\beta}_0$ is the estimated spatial mean, $W(u_i, v_i)$ is a nxn weight matrix, with null off-diagonal elements and the geographical weights in the main diagonal, X is a $nx1$ vector containing values equal to 1, and the superscript T indicates a transposed vector.

The GWR provides a non-parametric estimator for continuous localization functions (u_i, v_i) using kernels. The log-likelihood for each particular set of estimates does not provide a single solution. The way to adjust the optimization is to consider local log-likelihoods and take observations close to i (Bowman and Azzalini, 1997; Fotheringham et al., 2002). Thus, the estimation of GWR involves the selection of bands (or windows) for an isotropic kernel spatial weight function, such as the Gaussian, Tricubic, and Quadratic functions. Fixed or adaptive bands can also be used.[6] However, adaptive bands bias the identification of subcentralities because of the non-isotropic employment distribution over space. In cells surrounded by areas with zero employment or low standard deviations, the adaptive method requires larger bands to increase the sample or to reach higher variance. Adaptive bands do not permit the comparison with results from econometric estimations and jeopardize the identification of SDB. To overcome these limitations, we use fixed bands. Many researchers have chosen window sizes in a discretionary way

[6] A fixed band weight function considers a constant band size everywhere, regardless of the sample size. An adaptive weight function adjusts the band size to the sample size. Potential problems can derive from fixed weight functions, such as the calibration considering few observations, giving rise to parameter estimates with high standard errors, and, therefore, generating little soft surfaces. In extreme situations, insufficient variation in small samples may compromise the parameter estimation (Fotheringham et al., 2002). The second approach overcomes this limitation by adjusting the window size to the sample density variation.

(McMillen, 2001a; Redfearn, 2007; Kane et al., 2016). In this study, we apply a minimization method based on the Akaike Information Criterion (Fotheringham et al., 2000).

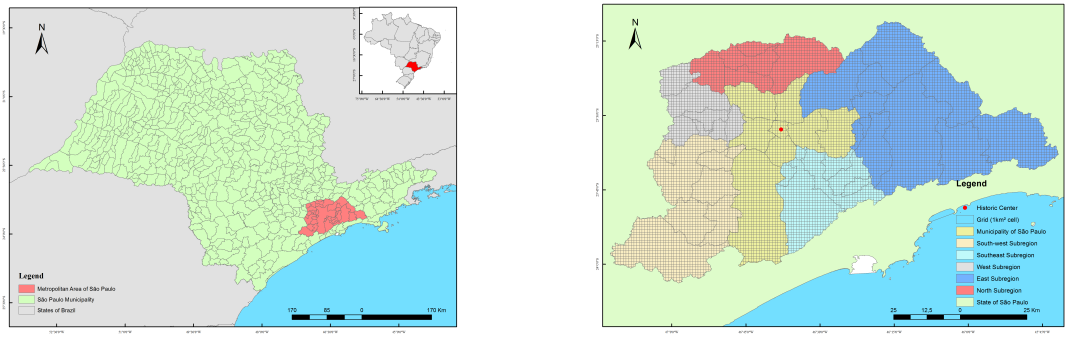
As a first step, we take as potential candidates to subcentralities areas whose spatially estimated employment is statistically significant ($\hat{\beta}_0^{**}$) and larger than or equal to an arbitrary critical value c. For these, we use the 99$^{\text{th}}$, 95$^{\text{th}}$, and 90$^{\text{th}}$ percentiles of the $\hat{\beta}_0^{*}$ distribution. Although such thresholds are arbitrary, they do not require any prior knowledge of the region and allow for changing the cutoffs in different years. Mathematically,

$$SBD_i = \hat{\beta}_0^{**} \geq c \tag{4}$$

The identification rule does not demand a pre-defined and static cutoff value. The procedure is flexible and adjustable intertemporally, differently from those presented by Giuliano and Small (1991), McMillen (2003), and Kane et al. (2016). The flexibility stems from two channels. Firstly, the estimated spatial averages are representative of the local employment agglomeration and do not involve a cutoff on total employment, as in estimates using employment density. Secondly, the procedure is based on the endogenous choice of the cutoff value, that is, the cutoffs derive from the distribution of the estimated spatial averages. It is up to the researcher to choose the level of significance for the statistical inference and the percentile. This is a step forward because it allows for the intertemporal comparability of the results. Even with the variation in the level of employment per cell, it captures the intertemporal dynamic of employment agglomeration by endogenizing the cutoff values and does not involve concerns with the optimal number of neighbors. Additionally, these potential candidates to SDB are distributed discretely in space, as proposed by Giuliano and Small (1991).

4. Database

The metropolitan area of São Paulo (SPMA), Brazil, is the country's largest metropolitan area. It is composed of 39 municipalities, spread over an area of 7,946 km^2 (Figure 1). Its population in 2017 was 21.6 million, representing 10% of the country's population. The municipality of São Paulo is the São Paulo State capital and accounts for 56% of the region's population and ranks amongst the largest cities in the world.



Source: own elaboration by using ArcGIS software.

Figure 1. Brazil, State of São Paulo and Metropolitan Area of São Paulo.

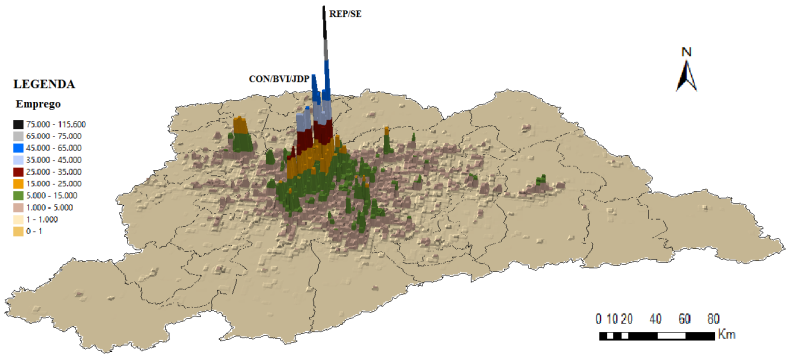
The size, location, characterization, and specialization of centers and sub-centers in municipalities in different countries around the world depend on the economic urban structure, which differs within and between countries (Duranton and Puga, 2015). European cities/regions started their development patterns in the Middle Ages, while American cities started much later, and are deeply associated with the transportation nodes (Krehl, 2016; Anas et al., 1998). Although historic factors are certainly important, government policies such as land use regulation and zoning do play a relevant role. Urban planning regulation in the US shaped the urban areas with well-defined clusters of land use: housing, commercial and industrial areas (McMillen and McDonald, 1999; Ewing et al., 2002). In Western Europe, such a characterization of land use is less clear. European cities contain multi-functional urban areas and mixes of land use (Krehl, 2016).

The spatial structure of the SPMA is closer to American cities but does present a mixed land use pattern and zoning policies. The similarities are related to the proximity to transportation nodes and land-use rules that promote verticalization. The evolution of the transportation network and government-imposed rules is important in explaining the intensification of agglomeration observed in the period analyzed. New high-capacity avenues, subway, and urban train lines in the CBD area, together with favorable construction regulations, increased verticalization and job agglomeration in the western part of São Paulo city (Campos, 2018). As a result, SPMA presents aspects of both American and European cities, with a concentration of jobs in specified areas (American dynamics) but many low concentrations of commerce and services spread around the area, although without power to overcome the attractiveness of the CDB and SDB (European dynamics).

The data comes from the mandatory Annual Social Information Report (RAIS—*Relação Anual de Informações Sociais*) produced by the Ministry of Labor. It covers all formally established (incorporated) organizations (public and private) and workers with a labor card and is considered as a census of formal workers. It leaves out informal organizations,[7] non-wage labor relations (self-employed, temporary work, etc.), and organizations of the public sector. The latter were excluded because the locational decision process for governmental activities is distinct from the private sector and could bias the results.

The addresses of the firms were geocoded, based on the street shapefile produced by the Centro de Estudos da Metropole (CEM, 2016) and World Location (online street shapefile) in ArcGIS. Private firms for which the effective place of work could not be identified, such as construction, sales, on-site maintenance services, security, etc., were dropped off as well. Since all workers are listed at the firms' headquarters in these cases, we wanted to avoid allocating to some addresses the presence of workers whose job is performed elsewhere. Finally, we work with a cross-section for 2002 and 2014. The number of employees grew from 3,084,074 in 2002 to 5,344,069 in 2014; the number of firms, from 238,918 to 348,807. To avoid the use of areas of different sizes and the questions involved in the use of gross or net employment density (McDonald, 1987; Mieszkowski and Smith, 1991; McMillen, 2001a), we have gridded the area into 9,071 cells of 1 km^2. Figure 2

[7] According to the National Household Sample Survey (PNAD, acronym in Portuguese), the share of formal employment in the SPMA was 50.1% in 2002 and 67.5% in 2014.



Source: own elaboration by using ArcGIS software, based on RAIS, 2014.

Figure 2. Spatial distribution of employment, 2014.

portrays the spatial distribution of employment in 2014. Based on the geocoded firm's addresses, we allocated employees to the cells.

5. Results

In this section, we present the results of our approach (CA). Given the relevance of McMillen and Smith (2003) in the empirical literature (MS), we apply their methodology to the same database, to provide a reasonable comparison.[8] We work with two moments, 2002 and 2014, to consider possible changes in the period. We define as subcenters contiguous areas with significant $\hat{\beta}_0$ coefficients larger than the 99^{th}, 95^{th}, or 90^{th} quantiles of the $\hat{\beta}_0$ distribution. Table 1 summarizes the selection rules for both approaches.

Table 2 presents the information criteria for the kernel functions. We opted for the tricubic function, since it presents the lowest marginal AIC, providing the optimal bandwidths. Table 3 indicates that the estimated employment cutoffs in 2014 are substantially larger than in 2002 due to the employment growth in the period. This illustrates the limitations of working with fixed thresholds.

Table 1
Main differences between the approaches.

	CA	MS
Kernel Function Selection	Test gaussian, bisquare, and tricubic functions; select the one with the lowest AIC[a]	Predefined: Tricubic
Bandwidth Size Selection	AIC minimization method	Predefined: 50% for all cities
Cutoff Selection	Quantiles applied on estimated local means	Predefined: 10,000 employees
SBD Selection	Statistically significance of local means that present estimated values larger than the cutoff	Error terms larger than pre-defined quantile and larger than the cutoff

Source: elaborated by the authors.
[a]AIC = Akaike Information Criterion.

[8]To have access to R script, see https://rodgercampos.github.io/code/.

Table 2

AIC information and optimal bandwidth.

	2002		2014	
Kernel Functions	AIC	Bandwidth (Km)	AIC	Bandwidth (Km)
Gaussian	150,453.90	0.91	159,668.70	0.871
Bisquare	150,289.60	2.271	159,429.80	2.129
Tricubic	150,197.60	2.309	159,279.60	2.063

Source: own elaboration by using R software, based on data from RAIS.

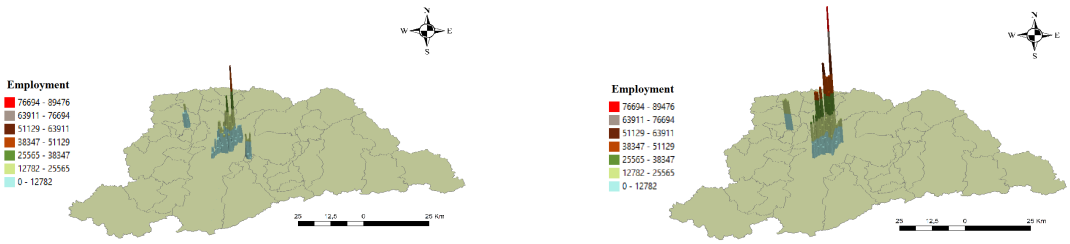
Table 3

Cutoff values and number of cells: CA approach.

Year	Quantile	Employment Cutoff	Number of cells		
			10%	5%	1%
2002	99th	5,764	91	91	91
	95th	1,876	363	363	363
	90th	819	454	454	454
2014	99th	10.07	91	91	91
	95th	3,132	363	363	363
	90th	1,380	454	454	454

Source: own elaboration by using R software. All cells identified as part of a CBD/SBD are statistically significant at 1%.

As in McMillen (2003), McMillen and Smith (2003), and Kane et al. (2016), an SBD is formed by clusters of cells identified as potential cases. Table 3 presents the numbers of cells of the SBDs for distinct cutoff values. The numbers are the same in both years: 91, 363, and 454 for the 99th, 95th, and 90th percentiles, respectively. As we use one-km^2 cells, the SBD's total extension is 91 km^2 in the more restrictive case, only 1.1% of the metropolitan area (4.6% and 5.7% in the other cases). The significance level does not influence the number of cells of the SBDs, and it does not change between years. The cutoff values have changed, but the number of cells in the SDBs remained the same, which is another form of showing that using fixed cutoff values is not appropriate for identifying SBD.



(a) 2002. **(b)** 2014.

Source: own elaboration by using ArcGIS software.

Figure 3. SBD identified (99th Percentile).

Figure 3 maps the location of the identified cells included in the SBD. We identify three SDB in 2002 and only two in 2014. The CBD and the SDB in the western part of the area have increased their importance, while the SDB in the eastern part of 2002 disappeared in 2014. There was intense growth in employment in the period, 58%, but it was spatially concentrated. Our method with changing thresholds captured this dimension of the evolution of the labor market.

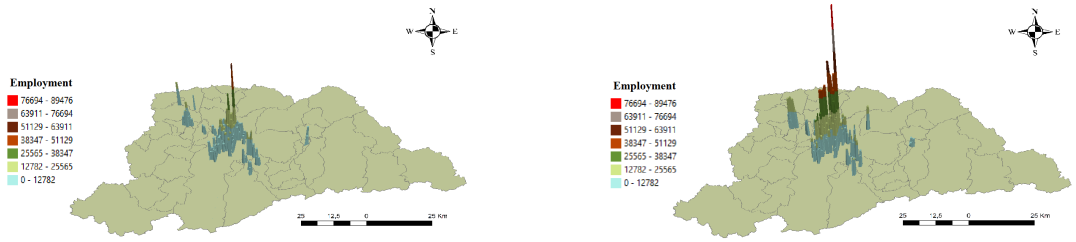
As to provide a means of comparison, we applied the method proposed by McMillen and Smith (2003) to the same database. We used the local polynomial (*lp*) model term for LWR, with smoothing parameters provided as an argument to *lp*, 50% bandwidth, tricubic kernel function, and degree 1. The identification of cells as potential SBD is performed in the first stage, and the SBD are selected in the second stage. As Table 4 shows, the numbers of cells are larger than in our approach for 99% but smaller for 95% and 90%. Moreover, the number of cells vary between years, increasing for the 99% threshold and diminishing for the other two. Figure 3 shows the spatial location of the selected cells. The MS approach identified 16 SBD in 2002 and 17 in 2014 as maps Figure 4. There are cases of SDB that disappeared and others that were created in the period. The two approaches come to different results, both in terms of area (km^2) and cells' location, but many areas are similar. Figure 5 shows the differences in the location of the areas. It shows that our approach does a better job in identifying areas closer to the CDB, while the MS procedure identifies more areas in the periphery.

Table 4

Cutoff values and total areas selected: MS approach.

Cutoff	2002	2014
99th	118	119
95th	159	144
90th	186	168

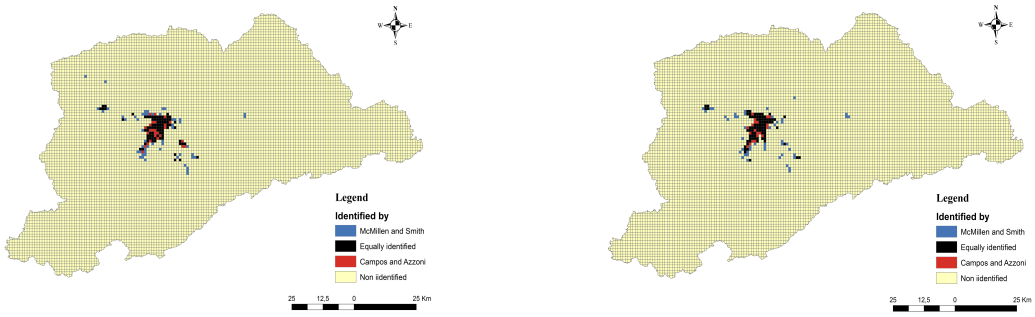
Source: own elaboration by using R software, based on RAIS database. All selected cells are statistically significant at 1%.



(a) 2002. **(b)** 2014.

Source: own elaboration by using ArcGIS software, based on RAIS database.

Figure 4. SBD by the MS approach, 2002 and 2014.



(a) 2002. **(b)** 2014.

Source: own elaboration by using ArcGIS software, based on RAIS database.

Figure 5. Differences in SBDs identified by the two methodologies.

Table 5
Descriptive statistics of residuals.

		2002		2014	
		Mean	Std Error	Mean	Std Error
MS approach[a]	Full sample	132.84	1,396.52	302.95	2,523.54
	SBD only*	9,348.41	6,485.09	16,387.50	12,347.09
CA approach[b]	Full sample	-0.013	798.32	-0.009	1,254.69
	SBD only**	955.42	5,991.02	1,395.42	10,080.35
MSCA approach[b]	Full sample	10.35	1,524.46	29.03	2,655.85
	SBD only*	9,323.89	6,361.96	16,198.93	12,121.86

Source: own elaboration by using R software, based on RAIS database.
*Number of observations is 119 (2002) and 118 (2014); **Number of observations is 91 in both years.
[a]Takes into account 50% of the sample as bandwidth.
[b]Uses the AIC criteria for bandwidth selection, as described in Table 1.

We have determined thus far that the two methodologies lead to quite different results. The next step is to explore the reasons why that happens. Table 5 brings information on the means and standard deviations of the residuals, showing that the MS procedure leads to larger values and higher variation, considering all cells, and the differences are quite high. Considering only the cells of the SDB, the residual means and standard deviations are larger, as compared to the whole sample, in both approaches.

The MS approach uses 50% of the sample as bandwidth, and our approach uses the AIC criteria. This difference could be a possible channel for the differences in the number of potential areas identified as SBD. To test this idea, we replaced the kernel distance from the MS approach with one based on the AIC information—MSCA hereafter. As shown in Table 6, it ended up finding 138 potential candidates in 2002 (189 at 5% and 231 at 10%) and 130 in 2014 (175 and 217). Table 5 shows that the means of the residuals estimated by using the MSCA approach decreased sharply. The standard deviations, however, increased in comparison to the MS approach. Our approach presents the lowest residual means and standard deviations based on such statistical exercises, indicating a better fit of the predicted SBD.

Table 6
Cutoff values and total area selected: MSCA approach.

Quantile	2002	2014
99th	138	130
95th	189	175
90th	231	217

Source: own elaboration by using R software, based on RAIS database.

We also ran probit models to measure the importance of the residuals in explaining the probability of identifying an SBD. The SBD may be seen as a latent variable, where cells take the value 1 if it is labeled as a potential candidate to be included in an SBD, and 0, otherwise. Mathematically,

$$\Pr\left(SBD = 1|X\right) = \phi(X^T\beta),$$

where Pr denotes the probability and ϕ is the Cumulative Distribution Function of the standard normal function. The parameters β are estimated by maximum likelihood and the matrix X is composed of covariates.

Table 7 brings the results. The coefficients for the variable SBD measure the probability of a cell being a potential SBD member. When estimated by MS, they are 10 to 100 times larger, as compared to our approach. These results indicate how important the error term is in each approach, and it is more relevant in MS than in our methodology. Furthermore, the R^2 is also larger in MS than in CA, reinforcing the previous conclusion. Thus, contrary to what McMillen and Smith (2003) stated, changing bandwidths does seem to affect the results. However, the differences are smaller than those found comparing the methodologies as they were presented. Thus, the estimator seems to be the most important factor for the differences in the results.

Table 7
Probit Model.

		2002			2014		
		SBD99	SBD95	SBD90	SBD99	SBD95	SBD90
MS approach[a]	SBD	0.0008604*	0.0010266*	0.0010861*	0.0003851*	0.0004268*	0.0004484*
	R^2	0.8909	0.872	0.8525	0.8296	0.8141	0.7824
CA approach[b]	SBD	0.0000769*	0.0000538*	0.00003887*	0.00004*	0.0000247*	0.0000233*
	R^2	0.0173	0.0035	0.0012	0.0133	0.0019	0.0012
MSCA approach[b]	SBD	0.0008783*	0.0010266*	0.0010869*	0.000377*	0.0004387*	0.00052*
	R^2	0.8854	0.8746	0.8572	0.8248	0.8023	0.7852
	Sample	9,071	9,071	9,071	9,071	9,071	9,071

Source: own elaboration by using R software, based on RAIS database. Statistical significance level: *1%, **5% and ***10%. [a]Takes into account 50% of the sample as bandwidth. [b]Uses the AIC criteria for bandwidth selection, as described in Table 1.

In summary, even using AIC criteria identified bandwidths, the MSCA approach diverges from the benchmark MS procedure and our approach. Our approach provides lower estimated residual means and standard deviations in comparison to the regular MS methodology. It is a relevant issue since the cutoff in the MS approach is based on the error term. This might be a reason for their approach to identifying more SBD cells than ours. Thus, at least in terms of the estimated errors, our approach is more powerful than MS for the identification of SBDs.

6. Final Remarks

Social scientists have tried to develop empirical models to identify SBDs without previous knowledge of the studied area. Besides the difficulty with databases, the definition of a cutoff value is at the center of the debate. This paper proposes a new empirical methodology that uses the quantiles of the estimated coefficients as cutoffs. This is a step towards overcoming arbitrariness since it permits the cutoff value to change temporally and spatially. It captures the labor market evolution, its spatial trend, and eventual crises. Our approach is also available for small areas since the bandwidth size is adjustable for city or country size. The comparison of the results of our model with the most referred model (MS) revealed that our methodology is more robust and avoids some of its limitations.

The findings reveal spatial disparity in job agglomeration in the area, with three business centers in 2002 and only two in 2014. The CDB is clearly defined, around the historic city center. In a period of intense job creation, the spatial trend of employment growth is highly concentrated, as the rate of growth in the CDB is higher than elsewhere. The spatial dynamics in SPMA is close to the American case and has aspects of the European experience: spatial dispersion of economic activity, multi-functional urban areas, and a mix of housing, commerce, and industry. The dispersion of commerce is also observed, as in Europe, but not enough to create business centers. Our findings indicate that SPMA is not a monocentric region, once the existence of a global business center in the region is evident. However, the spatial concentration of economic activities is intense, with only one or two sub-centers, and the predominance of the CDB increased in a period of high employment growth. This process caused the disappearance of one of the SDB observed at the beginning of the period.

The agglomeration strength of the CDB affects housing prices and location, the demand and supply of public transportation, the spatial match in the labor market, and wages. Both polycentric and monocentric theoretical models indicate higher land prices close to the CBD and SBD. Thus, increasing spatial concentration facilitates gentrification, especially in the city of São Paulo. The identification of sub-centers and their evolution is relevant for public policy. Although we use a grid of small cells to perform the analysis, it is possible to recover the area of each of the 39 municipalities, facilitating the design and implementation of public policy by local administrations. Gentrification increases the commuting time and stresses the demand for public transportation. The effects of concentration on land prices and gentrification are relevant research topics for metro regions in Brazil and other developing countries. Comparisons with metropolitan areas of developed countries are also relevant, something lacking in the Urban Economics literature.

References

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2016). The Economics of Density: Evidence from the Berlin Wall. *Econometrica 83(6)*, 2127–2189.

Alonso, W. (1964). *Location and Land Use*. Cambridge: Harvard University Press.

Anas, A., and Kim, I. (1996). General Equilibrium Models of Polycentric Urban Land Use with Endogenous Congestion and Job Agglomeration. *Journal of Urban Economics 28(1)*, 318–325.

Anas, A., Arnott, R., and Small, K. A. (1998). Urban Spatial Structure. *Journal of Economic Literature 36(3)*, 1426–1464.

Beckmann, M. J. (1974). Spatial Equilibrium in the Housing Market. *Journal of Urban Economics 1(1)*, 99–107.

Bender, B., and Hwang, H. (1985). Hedonic Housing Price Indices and Secondary Employment Centers. *Journal of Urban Economics l(17)*, 90-107.

Bowman, A. W., and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford Science Publications.

Campos, R. B. A. (2018). Subcentralidades e prêmio salarial intra-urbano na região metropolitana de São Paulo. São Paulo: Programa de Pós-Graduação em Economia; Faculdade de Economia, Administração e Contabilidade; Universidade de São Paulo. (Tese de Doutorado).

Campos, R. B. A., and Azzoni, R. C. (2020). Dispersão concentrada do emprego: uma incursão sobre os modelos teóricos e abordagens empíricas. *Revista Brasileira de Estudos Regionais e Urbanos 13(4)*, 606–627.

Castells, M., Hall, P., and Hall, P. G. (1994). Technopoles of the World: the Making of Twenty-First-Century Industrial Complexes. London, New York: Routledge.

Craig, S. G., and Ng, P. T. (2001). Using Quantile Smoothing Splines to Identify Employment Subcenters in a Multicentric Urban Area. *Journal of Urban Economics 49(1)*, 100–120.

Duranton, G., and Puga, D. (2015). Urban Land Use. In G. Duranton, J. V. Henderson and W. C. Strange (Eds.), *Handbook of Regional and Urban Economics* (pp. 467–560). Amsterdam: Elsevier.

Ewing, R. H., Pendall, R., and Chen, D. D. T. (2002). Measuring Sprawl and its Impacts. Smart Growth America (technical report).

Fotheringham, A. A., Brunsdon, C., and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage.

Fotheringham, A. A., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: the analysis of spatially varying relationship*. New York: Wiley.

Fujita, M. (1988). A Monopolistic Competition Model of Spatial Agglomeration: Differentiated Products Approach. *Regional Science and Urban Economics 18(1)*, 87–124.

Fujita, M., and Ogawa, H. (1982). Multiple Equilibria and Structural Transition of Non-Monocentric Urban Configurations. *Regional Science and Urban Economics 12(2)*, 161–191.

Giuliano, G., and Small, K. A. (1991). Subcenters in the Los Angeles region. *Regional Science and Urban Economics 21(2)*, 163–182.

Giuliano, G., Redfearn, C., Agarwal, A., Li, C, and Zhuang, D. (2005). Not all sprawl: evolution of employment concentrations in Los Angeles, 1980–2000. Lusk Center for Real Estate, University of Southern California, Working Paper 2005-1002.

Graham, S., and Marvin, S. (1996). *Telecommunications and The City: Electronic Spaces, Urban Places*. London: Routledge.

Greene, D. (1980). Recent Trends in Urban Spatial Structure. *Growth and Change 11(1)*, 29–40.

Hartwick, P., and Hartwick, J. M. (1974). Efficient Resource Allocation in a Multinucleated City with Intermediate Goods. *Quarterly Journal of Economics 88(2)*, 340–352.

Heikkila, E., Gordon, P., and Kim, J. I. (1989). What Happened to the CBD-Distance Gradient?: Land Values in a Policentric City. *Environment and Planning A 21(2)*, 221–232.

Helsley, R. W., and Sullivan, A. M. (1991). Urban subcenter formation. *Regional Science and Urban Economics 21(2)*, 255–275.

Henderson, J. V., and Mitra, A. (1996). The new urban landscape: Developers and edge cities. *Regional Science and Urban Economics 26(6)*, 13–643.

Henderson, J. V., and Slade, E. (1993). Development Games in Non-monocentric Cities. *Journal of Urban Economics 34(2)*, 207–230.

Hotchkiss, D., and White, M. (1993). A Simulation Model of a Decentralized Metropolitan Area with Two-Worker, 'Traditional' and Female-Headed Households. *Journal of Urban Economics 34(2)*, 159–185.

Kane, K., Hipp, J. R., and Kim, J. H. (2016). Los Angeles employment concentration in the 21st century. *Urban Studies 55(4)*, 1–26.

Krehl, A. (2016). Urban subcentres in German city regions: Identification, understanding, comparison. *Papers in Regional Science 97(1)*, 79–105.

Lucas, R. E., and Rossi-Hansberg, E. (2002). On the Internal Structure of Cities. *Econometrica 70(4)*, 1445–1476.

McDonald, J. F. (1987). The Identification of Urban Employment Subcenters. *Journal of Urban Economics 21(2)*, 242–258.

McDonald, J. F., and Prather, P. J. (1994). Suburban Employment Centres: The Case of Chicago. *Urban Studies 31(2)*, 201–218.

McMillen, D. P. (2001a) Nonparametric Employment Subcenter Identification. *Journal of Urban Economics 50(3)*, 448–473.

McMillen, D. P. (2001b). Polycentric urban structure: The case of Milwaukee. *Economic Perspectives 25(2)*. Federal Reserve Bank of Chicago, 15–27.

McMillen, D. P. (2003). Identifying Sub-centres Using Contiguity Matrices. *Urban Studies 40(1)*, 57–69.

McMillen, D. P., and McDonald, J. F. (1997). A Nonparametric Analysis of Employment Density in a Polycentric City. *Journal of Regional Science 37(4)*, 591–612.

McMillen, D. P., and McDonald, J. F. (1999). Land use before zoning: The case of 1920's Chicago. *Regional Science and Urban Economics 29(4)*, 473–489.

McMillen, D. P., and Smith, S. C. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics 53(3)*, 321–338.

Mills, E. S. (1967). An Aggregative Model of Resource Allocation in a Metropolitan Area. *American Economic Review 57(2)*, 197–210.

Mills, E. S. (1972). *Studies in the Structure of the Urban Economy.* Baltimore: John Hopkins University Press.

Mieszkowski, P., and Smith, B. (1991). Analyzing urban decentralization: The case of Houston. *Regional Science and Urban Economics 21(2)*, 183–199.

Muth, R. F. (1969). *Cities and Housing: The Spatial Pattern of Urban Residential Land Use.* Chicago: University of Chicago Press.

Muth, R. F. (1975). Numerical Solution of Urban Land-Use Models. *Journal of Urban Economics 2(4)*, 307–332.

Ogawa, H., and Fujita, M. (1980). Equilibrium Land Use Patterns in a Non-monocentric City. *Regional Science and Urban Economics 20(4)*, 455–475.

Papageorgiou, G. J. (1971). The population density and rent distribution models within a multicentre framework. *Environment and Planning A 3*, 267–282.

Redfearn, C. (2007). The topography of metropolitan employment: Identifying centers of employment in a polycentric urban area. *Journal of Urban Economics 61(3)*, 519–541.

Richardson, H. W., Gordon, P., Jun, M., and Heikkila, E. (1990). Residential Property Values, the CBD, and Multiple Nodes: Further Analysis. *Environment and Planning A 22(6)*, 829–833.

Romanos, M. C. (1979). Household Location in a Linear Multi-Center Metropolitan Area. *Regional Science and Urban Economics 7(3)*, 233–250.

Ross, S, and Yinger, J. (1995). Comparative static analysis of open urban models with a full labor market and suburban employment. *Regional Science and Urban Economics 25(5)*, 575–605.

Sasaki, K., and Mun, S. (1996). A Dynamic Analysis of a Multiple-Center Formation in a City. *Journal of Urban Economics 40(3)*, 257–278.

Sivitanidou, R., and Wheaton, W. C. (1992). Wage and Rent Capitalization in the Commercial Real Estate Market. *Journal of Urban Economics 31(2)*, 206–229.

Solow, R. (1973). Congestion costs and the use of land for streets. *Bell Journal 4(2)*, 602–618.

Sullivan, A. (1986). A General Equilibrium Model with Agglomerative Economies and Decentralized Employment. *Journal of Urban Economics 20(1)*, 55–74.

Wheaton, W. C. (1974). A Comparative Static Analysis of Urban Spatial Structure. *Journal of Economic Theory 9(2)*, 223–237.

White, M. J. (1976). Firm Suburbanization and Urban Subcenters. *Journal of Urban Economics 3(4)*, 323–343.

White, M. J. (1988). Location Choice Behavior and Commuting Behavior in Cities with Decentralized Employment. *Journal of Urban Economics 24(2)*, 129–152.

White, M. J. (1999). Urban Areas with Decentralized Employment: Theory and Empirical Work. In P. Chesire and E. S. Mills (Eds.), *Handbook of Regional and Urban Economics, vol. 3:*

Applied Urban Economics (pp. 1375–1412). North-Holland.

Wieand, K. (1987). An Extension of the Monocentric Urban Spatial Equilibrium Model to a Multicenter Setting: The case of the Two-Center City. *Journal of Urban Economics 21(3)*, 259–271.

Wrede, M. (2015). A continuous spatial choice logit model of a polycentric city. *Regional Science and Urban Economics 53*, 68–73.

Yinger, J. (1992). Urban Models with More Than One Employment Center. *Journal of Urban Economics 31(2)*, 181–205.

Zhang, Y., and Komei, S. (1997). Effects of subcenter formation on urban spatial structure. *Regional Science and Urban Economics 27(3)*, 297–324.

Zhang, Y. and Komei, S. (2000). Spatial structure in an open city with a subcenter. *Annals of Regional Science 34(1)*, 37–53.