Analytics and Spatial Statistics: Market Expansion as a Growth Strategy for a Health Products Business in Brazil

ABSTRACT

This study aimed at exploring which criteria are important to support the expansion strategy of a Brazilian healthcare distribution company focused on Orthopedics. The investigated level of granularity was city (within the whole Brazilian territory), and hence, the idea was to identify what makes a city interesting in terms of potential market. We also intended to answer whether there was a "market score" that could help the company's executives to choose the best places to expand. To do that, a dataset was built with information gathered from public datasets including health and socioeconomic indicators. The adopted analytical approaches included the use of logistic regression to compute city scores, and a series of regression models including linear (OLS) and the innovative Spatial Auto-regressive models (SAR) and Geographically Weighted Regression (GWR). Cluster analyses (k-means and SKATER algorithms) were also used to find non-disjointed territorial groups of cities, according to a set of variables. Main results confirmed that the use of the spatial analytical approaches showed utility and significantly improvements. The spatial regression models showed increasingly better fit to the data, with the SAR model presenting a lower Akaike information criterion (AIC) (42635.8) and greater R^2 (0.85) than the OLS model (42761.3 and 0.84, respectively); and the GWR, a very lower AIC (25748.72) and greater R^2 (quasi-global R^2 of 0.99) than both the OLS and SAR models. Findings provided useful insight for management, relationship segmentation and prediction that can support the development of a solid expansion strategy.

KEYWORDS: Geographic Expansion Strategy, Spatial Statistics, Cluster Analysis, SKATER

INTRODUCTION

Data science accomplishes a process of deeply inspecting and understanding facts generated by multiple data sources that can be used to support predictive models targeting decision-making (PROVOST; FAWCETT, 2013). As result, such analytical based insights can highly benefit business decision-making (BARTON; COURT, 2012). This work describes how data analysis was used in a real case to support a healthcare distribution company based in Brazil on its decision to geographically expand its business to highly promising regions. It focuses on the development of predictive models to more accurately answer the following business question: What priority territories should be targeted to support the company's geographic expansion strategy?

1. OBJECTIVES AND STUDY OBJECT

1.1 Target: The Market Expansion of a Healthcare Distribution Company in Brazil

The target company of this study focuses on distributing products from an American brand manufacturer specialized on medical devices to treat intra-articular lesions (market dominated by a restricted number of manufacturers, which makes the business well centralized). Specifically, the company represents, imports and distributes medical products and supplies for the video-arthroscopy segment (i.e., a minimally invasive surgical procedure in which internal joint damage is assessed with the use of a specific endoscope). A product example is a set of titanium flying plates applied to shoulders and knees. In addition, the company offers hospitals and doctors services associated with the products, such as renting the video-arthroscopy device to perform the surgical procedure and providing a range of surgical instruments for loan (e.g., screws and plates of different sizes).

The company is currently concentrating its activities in two states in the Southern Brazil (Rio Grande do Sul and Paraná), being the leader in the region. The Southern region is the second most important region for the health sector in Brazil – with the second largest concentration of orthopedists – and presents the highest life expectancy (Santa Catarina (SC), 78.7 years, ranks first in the ranking of Brazilian states; Rio Grande do Sul (RS), 77.5 years, the fifth; and Paraná (PR), the seventh, 76.8 years). As characteristics of the company's distribution operations, some entry barriers are observed: (a) technical (knowledge), due to the strong product specialization; (b) delivery and logistics in each location, and (c) network (relationship with potential clients). These barriers end up restricting the performance of such companies in general. More recently, there has been an increasingly demand for professionalization of such distribution companies which, on one hand, has reduced the number of competitors of the company, but on the other hand, has increased costs and required the development of a more sophisticated (and more expensive) structure. Consequently, the local and current market share (customers of the Southern region) is no longer sufficient, thus generating the need to expand not just in the Southern region but also especially to other cities of the country. To be able to decide which areas have potential for expansion, the company executives need an answer about what makes a place (city) interesting in terms of potential market. Enabling operations in other locations requires having a minimum billing that justifies investment, available equipment and maintenance of employees.

Currently, the company's target is on the private health market. Some relevant indicators in this context are: (a) the sociodemographic ones, such as the Gross Domestic Product (i.e., *per capita* GDP of the place) and *per capita* income, (b) those related to the healthcare network structure, such as the ones related to hospitals (e.g., orthopedic surgical beds), human resources (e.g., number of orthopedists), and (c) private health care indicators, such as the number of beneficiaries of private health insurances per municipality.

The company already operates in municipalities in Southern Brazil in which market volume is known thus making operations feasible, like Porto Alegre, Passo Fundo, Bento Gonçalves, Novo Hamburgo, Lajeado, and Curitiba. There are also municipalities in which despite not operating, the company knows there is market potential as, for example, Joinville and Florianópolis, in Santa Catarina (SC) State. These municipalities can help better understand which combination of variables is important to the success of the company's operations. What common points exist between possible new market scenarios and the markets in which the company already operates successfully? The answer to this question can help to establish a set of characteristics that maximize the decision-making by new markets. Is there a "market score" that helps executives to choose the best places? This is one of the questions this work aims to explore.

1.2 Data Analytical Approaches: How to tackle the Research Question

There are several factors that must be considered when deciding where to locate a business, being from a simple new branch in a neighborhood to a business expansion in new regions and cities with promising markets. In the case of this research, geospatial analysis is key to determine where to expand the company business, particularly when the granularity is city. By analyzing trends and spatial patterns, it is possible to develop predictive models that help the decision of which city the business should expand to. One goal of geospatial analyses is to derive practical insight from customer and market data. Such obtained insight may represent an advantage and help businesses to find a way towards profitability and growth (MILLER, 2011). Geospatial analyses can therefore provide decision makers relevant information on the dynamics of their marketplace (e.g., establishing areas of untapped potential, predicting new business catchment areas). In addition, the use of Spatial Statistics incorporated spatial

distribution of the variables in the prediction models (FRANCISCO, 2010) and usually improve the variance explanation (e.g., MORANDI *et al.*, 2016, FRANCISCO *et al.*, 2010).

The analytical strategy applied to this study included, on one hand, to develop logistic regression models to allow the prediction of the probability of a city being worth expanding to. More specifically, logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories (HOSMER; LEMESHOW, 2000). The dependent variable is categorical and binary. For example, a company that wants to predict the chances that a client is a good client can use logistic regression. The company can use customers' historical data to compute a binary dependent variable such as good vs. bad clients and add independent variables in the model that help to compute a credit scoring (i.e., the probability of a client to be good or bad). In this study, a logistic regression model was built to predict the probability of a city to be good (i.e., worth expanding to). An Ordinary Least Squares (OLS) regression model to predict number of beds in orthopedic postoperative units was also carried out. The OLS equation model is expressed as $y = X\beta + \varepsilon$ (SWEENEY; WILLIAMS; ANDERSON, 2013).

Aiming to consider spatial variables that may influence these predictive regression models, spatial regressions were performed. Namely, a Spatial Autoregressive model (SAR) (ANSELIN, 1988) and a Geographically Weighted Regression model (GWR) (FOTHERIGHAM; BRUNSDON; CHARLTON, 2002). The SAR method includes the global *spatial autocorrelation* indicator, *Moran's I*, and the Local Indicator of Spatial Autocorrelation (LISA). Simply put, a pure spatial autoregressive model consists of a spatially lagged version of the dependent variable, y, such that $y = \rho Wy + X\beta + \varepsilon$. The SAR equation model is similar to a standard linear regression model where the first term is constructed from a predefined *n* by *n* spatial weighting matrix, W, applied to the observed variable, y, together with a spatial autoregression parameter, ρ, which typically has to be estimated from the data. Essentially a spatial lag model is expressing the notion that the value of a variable at a given location is related to the values of the same variable measured at nearby locations, reflecting some kind of interaction effect (DE SMITH, 2015).

Regarding the GWR regression model, it is used to investigate how the relationship between a dependent variable (y) and one or more independent variables $(X1, X2, \dots, Xn)$ may vary geographically, $y(gi) = X(gi)x\,\beta(gi) + \varepsilon$. What GWR does is to provide a local model of the variable being predicted (y) by fitting a regression equation to every feature in the dataset. Separate equations that incorporate the dependent and explanatory variables of features falling within the bandwidth of each target feature are computed. The shape and size of the bandwidth is dependent on user input for the Kernel Type, Bandwidth Method, Distance, and Number of Features (FOTHERINGHAM *et al.*, 2002).

Finally, spatial analysis was also performed through a Spatial 'K'luster Analysis by Tree Edge Removal procedure (SKATER) (ASSUNÇÃO; LAGE; REIS, 2002). SKATER is a spatial cluster approach that is in a way similar to the usual cluster analytical approach in which a set of objects are grouped in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). The main difference between SKATER and the usual cluster techniques is that SKATER takes into account geographical regions, aiming to find spatial clusters within contiguous areas.

2. METHODOLOGY

2.1 Database Creation

The study's database includes data for 5565 cities within Brazil. It was created with variables from different public databases (see Table 1). It includes sociodemographic variables (e.g., Human Development Index, Gini coefficient, Gross Domestic Product, and Population),

variables that inform on registered health establishments (e.g., number of beds in postoperative units and in orthopedic postoperative units), and health related variables from the public (e.g., number of Orthopedists, number Orthopedic Surgical Procedures) and the private health systems (e.g., number Orthopedic Surgical Procedures, and number of health insurance beneficiaries). Except for City and State, all variables in the dataset are numeric and continuous. A health insurance beneficiaries´ average ticket index was computed based on the average ticket per insurance company weighted by the number of beneficiaries in each state. Then, for each city, an index was computed based on the state ticket index weighted by the number of beneficiaries in the city. Diagram 1 describes the steps for computing the health insurance beneficiaries´ average ticket index.

Table 1: Description of Study Database Variables

Variable	Source	Period
City	Brazilian Institute of Geography and Statistics (IBGE)	-
State	Brazilian Institute of Geography and Statistics (IBGE)	-
IBGE Code (for city)	Brazilian Institute of Geography and Statistics (IBGE)	-
Human Development Index (HDI)	Human Development Atlas in Brazil	2010
Per Capita Income	Brazilian Institute of Geography and Statistics (IBGE)	2010
Gini Coefficient	Brazilian Institute of Geography and Statistics (IBGE)	2010
Gross Domestic Product (GDP)	Brazilian Institute of Geography and Statistics (IBGE)	2013
Population	Brazilian Institute of Geography and Statistics (IBGE)	2012
Male Population	Brazilian Institute of Geography and Statistics (IBGE)	2012
Female Population	Brazilian Institute of Geography and Statistics (IBGE)	2012
Age in years*	Brazilian Institute of Geography and Statistics (IBGE)	2012
Beds in Postoperative Units	National registration of health establishments (CNES/DATASUS)	10/2016
Public Beds in Postoperative Units	National registration of health establishments (CNES/DATASUS)	10/2016
Private Beds in Postoperative Units	National registration of health establishments (CNES/DATASUS)	10/2016
Beds in Orthopedic Postoperative Units	National registration of health establishments (CNES/DATASUS)	10/2016
Public Beds in Orthopedic Postoperative Units	National registration of health establishments (CNES/DATASUS)	10/2016
Private Beds in Orthopedic Postoperative Units	National registration of health establishments (CNES/DATASUS)	10/2016
Private Health Insurance Beneficiaries	National Regulatory Agency for Private Health Insurance and Plans (ANS)	10/2016
Orthopedic Surgical Procedures (Public)	Basic Health Indicators and Data - Brazil (DATASUS)	01-06/2016
Value spent in Orthopedic Surgical Procedures (Public)	Basic Health Indicators and Data - Brazil (DATASUS)	01-06/2016
Orthopedic Surgical Procedures (Private)	National Regulatory Agency for Private Health Insurance and Plans (ANS)	01-06/2016
Orthopedists	Basic Health Indicators and Data - Brazil (DATASUS)	11/2016
Health insurance beneficiaries average ticket index	National Regulatory Agency for Private Health Insurance and Plans (ANS)	11/2016

Source: The authors

*:Age is divided in the following categories (in years): <1, 1 to 4, 5 to 9, 10 to 14, 15 to 19, 20 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, 70 to 79, >=80.

Diagram 1: Construction of the health insurance beneficiaries average ticket index

Step 1: Health insurance company average ticket (*x*)	**Step 2:** Weight of each health insurance company per state (*w*)
$x = \frac{Health\ insurance\ company\ revenue\ in\ Brazil}{Health\ insurance\ company\ beneficiaries\ in\ Brazil}$	$w = \frac{Health\ insurance\ company\ beneficiaries\ in\ the\ state}{Total\ health\ insurance\ beneficiaries\ in\ the\ state}$
Step 3: State health insurance average ticket	**Step 4:** City weight (*CW*)
$\bar{x} = \frac{\sum_{i=1}^{n}(xi * wi)}{\sum_{i=1}^{n} wi}$	$CW = \frac{Health\ insurance\ beneficiaries\ in\ the\ city}{Health\ insurance\ beneficiaries\ in\ the\ state}$
Step 5: Health insurance beneficiaries average ticket index	
$State\ health\ insurance\ average\ ticket\ x\ City\ weight$	

Source: The authors.

2.2 Data Analysis

Data analyses included three main approaches: computation of city scores by using a logistic regression approach to investigate which cities were good cities to expand and which were not (see Results for the complete description on how these scores were computed). Descriptive maps showing the distribution of cities by score were created. Regression models were carried out to determine which variables predicted the number of beds in orthopedic postoperative units (the overall number includes both the public and the private beds). Regression model approaches included linear models (Ordinary Least Square, OLS), and spatial models (Spatial Autoregressive, SAR, and Geographically Weighted Regression, GWR). At last, cluster analyses including K-Means clustering and spatial cluster analysis with the use of Spatial 'K'luster Analysis by Tree Edge Removal (SKATER).

3. RESULTS

3.1 City Scores

Previously to running the regression models to explore scores, the sample was divided into learning (70%) and test (30%) samples. The first step for computing city scores was to carry out a backward logit model within the learning sample to select the significant variables to be included in the binary logistic regression model (i.e., $p < 0.10$). All the independent variables were continuous and were standardized before the analysis (i.e., $M = 0$, $SD = 1.0$). The independent variables included in the model were: *per capita* income, Gross Domestic Product (GDP), population, number of private health insurance beneficiaries, orthopedic surgical procedures (public), orthopedic surgical procedures (private), health insurance beneficiaries average ticket index, and number of orthopedists. Because there is no historical data indicating whether a city presents a good market for orthopedic health care products, we operationalized it as follows: if the number of available beds in orthopedic postoperative units was < 5, the city took the value 0, whereas if the number of available beds in orthopedic postoperative units was >=5, the city took the value 1. This analysis indicated that *per capita* income, population, number of private health insurance beneficiaries, health insurance beneficiaries' average ticket index, and number of orthopedists were significant considering an *alpha* < 0.10.

A logistic regression analysis was then conducted to predict whether a city presented a good market for expansion (i.e., number of beds in orthopedic postoperative units >=5) using *per capita* income, population, number of private health insurance beneficiaries, health insurance beneficiaries average ticket index, and number of orthopedists as predictors. A test of the full model against a constant only model was statistically significant, indicating that the predictors as a set reliably distinguished between cities with at least five beds in orthopedic postoperative units and cities with less than five beds in orthopedic postoperative units (chi square = 1076.1791, $p < .001$ with $df = 5$). In this case, as shown in Table 2, lower income, smaller population, smaller health insurance beneficiaries average ticket index, and less orthopedists were associated with cities in which the number of beds in orthopedic postoperative units were less than five. On the other hand, the greater the number of private health insurance beneficiaries was, the higher was the chance that the city had less than five beds in orthopedic postoperative units. The error rate of the model in the learning sample was 4.62% and in the test sample, 5.75%.

From the logistic regression equation, scores were obtained to indicate the probability that a city had less than five beds in orthopedic postoperative units. The regression equation obtained was:

$$\boldsymbol{Z = 3.17 - 0.74\ x\ per\ capita\ income - 9.78\ x\ population + 8.55\ x\ private\ health\ insurance\ beneficiaries - 1.49\ x\ health\ insurance\ beneficiaries\ average\ ticket\ index - 3.45\ x\ orthopedists}$$

To convert the resulting Z value into scores (probabilities), the following formula was applied:

$$\boldsymbol{City\ score}\ (\mathbf{bad}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{Exp}(-\boldsymbol{Z})}$$

To find the probability of a city representing a good market to expand to, the above probability (i.e., "bad") was then subtracted from 1.

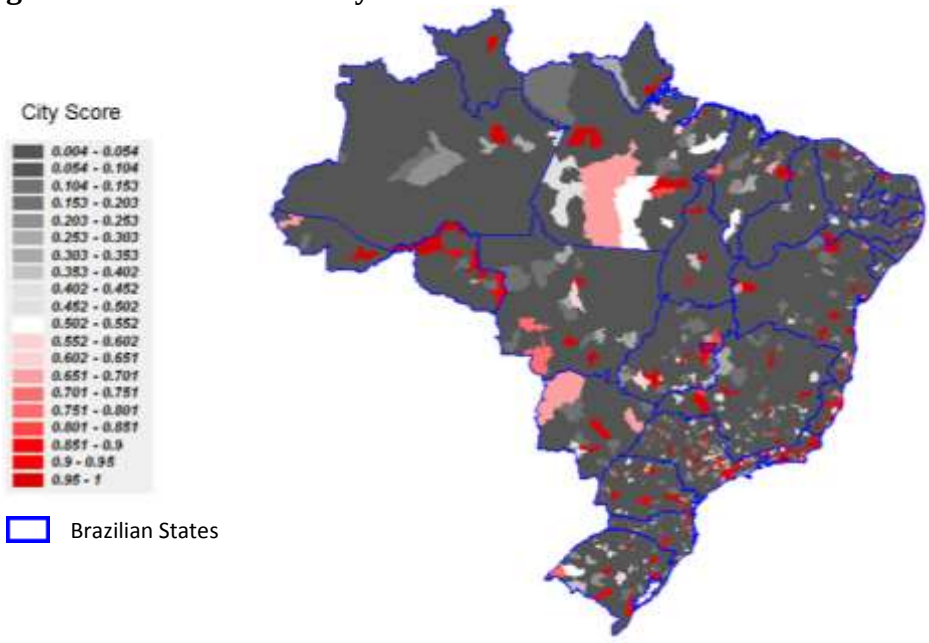
Table 2: Binary Logistic Regression on Number of Beds in Orthopedic Postoperative Units (< 5)

Variable	Coefficient	S.D.	Wald	Signif.
Constant	3.17	0.12	684.22	0.0000
Per capita income	-0.74	0.09	60.34	0.0000
Population	-9.78	0.76	164.81	0.0000
Private health insurance beneficiaries	8.55	1.02	69.69	0.0000
Health insurance beneficiaries average ticket index	-1.49	0.54	7.65	0.0057
Orthopedists	-3.45	0.78	19.64	0.0000

Source: The authors.

To help visualizing the city scores, a map was created (Figure 2). As can be seen, the scores vary from 0.004 to 1.00 (the greater the number, the higher the probability of a city being a good market for expansion).

Figure 1: Distribution of city scores within Brazil



Source: The authors, using ArcView GIS 3.2

3.2 Regression Models

Ordinary Least Square Model (OLS). The OLS regression model was carried out in R 3.0.1 and GeoDA 1.6.7. As both software showed the same results, we describe the GeoDA results here. Per capita income, Gross Domestic Product, population, number of health insurance beneficiaries, number of orthopedic surgical procedures in the public and the private system, health insurance beneficiaries average ticket index, and number of orthopedists were regressed on the overall number of beds in orthopedic postoperative units. This model was significant, *F-statistic* = 3858.44, $p < 0.001$, *df* = 5551. Detailed results are shown in Table 3. Per capita income, population, number of orthopedic surgical procedures in the public system, health insurance beneficiaries average ticket index, and number of orthopedists were positively associated to the overall number of beds in orthopedic postoperative units, whereas Gross Domestic Product, number of private health insurance beneficiaries, and number of orthopedic surgical procedures in the private system were negatively associated to it. The total variance explained by the model was $R^2 = 0.84$. The Log-likelihood was equal to -21317.7, the Akaike info criterion (AIC) was 42761.3, and the Schwarz criterion was 42820.9. The multicollinearity condition number was 21.56, which is below 30, thus indicating no multicollinearity problem. Diagnostics for spatial dependence showed that both the LM (lag) and the LM-robust (lag) were significant, while the LM (error) was not. This indicates that a Spatial Autoregressive Model (SAR) based on lag and not error is recommended to improve the model fit.

Table 3: OLS Regression Results on Number of Beds in Orthopedic Postoperative Units

Variable	Coefficient	S.E.	t-statistic	Probability
Constant	-2.415304	0.3611998	-6.686892	0.00000
Per capita income	0.001714109	0.0007419369	2.310316	0.02091
Gross Domestic Product	-0.00002190709	0.000008470546	-2.586266	0.00973
Population	0.0001273737	0.000005059834	25.1735	0.00000
Private health insurance beneficiaries	-0.00007879721	0.00001068373	-7.375443	0.00000
Orthopedic surgical procedures (public)	0.01813244	0.001673563	10.83463	0.00000
Orthopedic surgical procedures (private)	-0.005187848	0.001144254	-4.533826	0.00001
Health insurance beneficiaries average ticket index	0.02386661	0.001906661	12.51749	0.00000
Orthopedists	0.237412	0.03215602	7.383129	0.00000

Source: The authors.
Note. Coefficients are small due to the high values in the variables and high variance

Spatial Autoregressive Model (SAR). The SAR regression model was carried out in R 3.0.1 and GeoDA 1.6.7. As both software showed the same results, we describe the GeoDA results here. *Per capita* income, Gross Domestic Product, population, number of health insurance beneficiaries, number of orthopedic surgical procedures in the public and the private system, health insurance beneficiaries´ average ticket index, and number of orthopedists were regressed on the overall number of beds in orthopedic postoperative units. This model showed a better fit than the OLS model as can be seen by a greater Log- likelihood (-21307.9), and smaller AIC info (42635.8) and Schwarz info (42702) criteria. The total variance explained by the model was greater than in the OLS model, $R^2 = 0.85$. Detailed results are shown in Table 4. *Per capita* income, population, number of orthopedic surgical procedures in the public system, health insurance beneficiaries´ average ticket index, and number of orthopedists were positively associated to overall number of beds in orthopedic postoperative units, whereas Gross Domestic Product, number of private health insurance beneficiaries, and number of orthopedic surgical procedures in the private system were negatively associated to it.

In the maps on Figure 2, the Local Indicator Spatial Auto-correlation maps as well as the Moran's I are shown for the dependent variable (beds in orthopedic postoperative units) and the variables that presented higher coefficients in the model. The significance for local and global indicators are lower than 0.05.

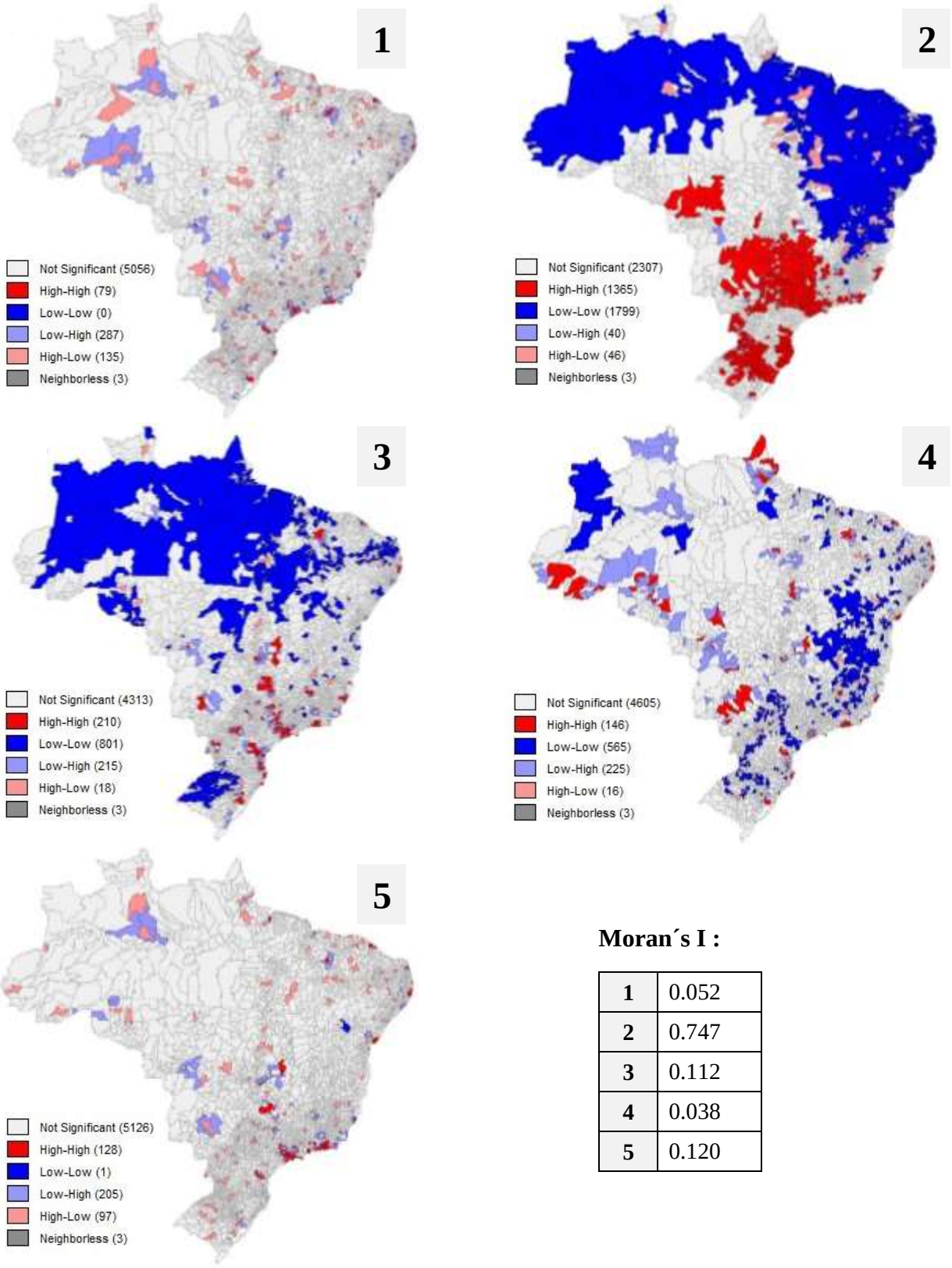
Table 4: SAR Regression Results on Number of Beds in Orthopedic Postoperative Units

Variable	Coefficient	S.E.	z-value	Probability
W-OrthoRooms	-0.07938045	0.006556688	-12.10679	0.00000
Constant	-2.559616	0.3568609	-7.172588	0.00000
Per capita income	0.002267229	0.0007345107	3.08672	0.00202
Gross Domestic Product	-0.00001718137	0.000008375679	-2.051341	0.04023
Population	0.000137281	0.000005050955	27.17922	0.00000
Private health insurance beneficiaries	-0.00009605651	0.000001062984	-9.036499	0.00000
Orthopedic surgical procedures (public)	0.01765671	0.001653043	10.68133	0.00000
Orthopedic surgical procedures (private)	-0.006089429	0.001132606	-5.376475	0.00000
Health insurance beneficiaries average ticket index	0.02066851	0.001901241	10.87106	0.00000
Orthopedists	0.2604981	0.03183282	8.183318	0.00000

Source: The authors.
Note. Coefficients are small due to the high values in the variables and high variance

Geographically Weighted Regression Model (GWR). The GWR spatial regression model included the following independent variables: *per capita* income, Gross Domestic Product, population, orthopedists, health insurance beneficiaries, health insurance beneficiaries' average ticket index, and orthopedic surgical procedures in the public and private systems. The dependent variable was the number of beds in orthopedic postoperative units. The kernel function used was the gwr.Gauss, and each of the 5560 cities had about 8 neighbors (five cities had no neighbors as for example, islands, like Fernando de Noronha). Table 5 shows the summary of GWR coefficient estimates at data points. Most important, the GWR model showed a quasi-global R^2 of 0.99, and was, therefore, the model that best predicted the number of beds in orthopedic postoperative units. The AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21) was 28367.79 and the AIC (GWR p. 96, eq. 4.22) was 25748.72, with a Sigma (model: traces) of 2.506735 and Sigma (ML), 2.141505. The lower AIC value as compared to the OLS and the SAR models indicate that the GWR model presented a better fit to the data.

Figure 2: Local Moran's I and Moran´s I (in the table below) for (1) number of beds in orthopedic postoperative units, (2) per capita income, (3) orthopedic surgical procedure (public), (4) health insurance beneficiaries average ticket index, and (5) number of Orthopedists.



Moran´s I :

1	0.052
2	0.747
3	0.112
4	0.038
5	0.120

Source: The authors, using GeoDA 1.6.7.

Table 5: GWR Coefficient Estimates at Data Points

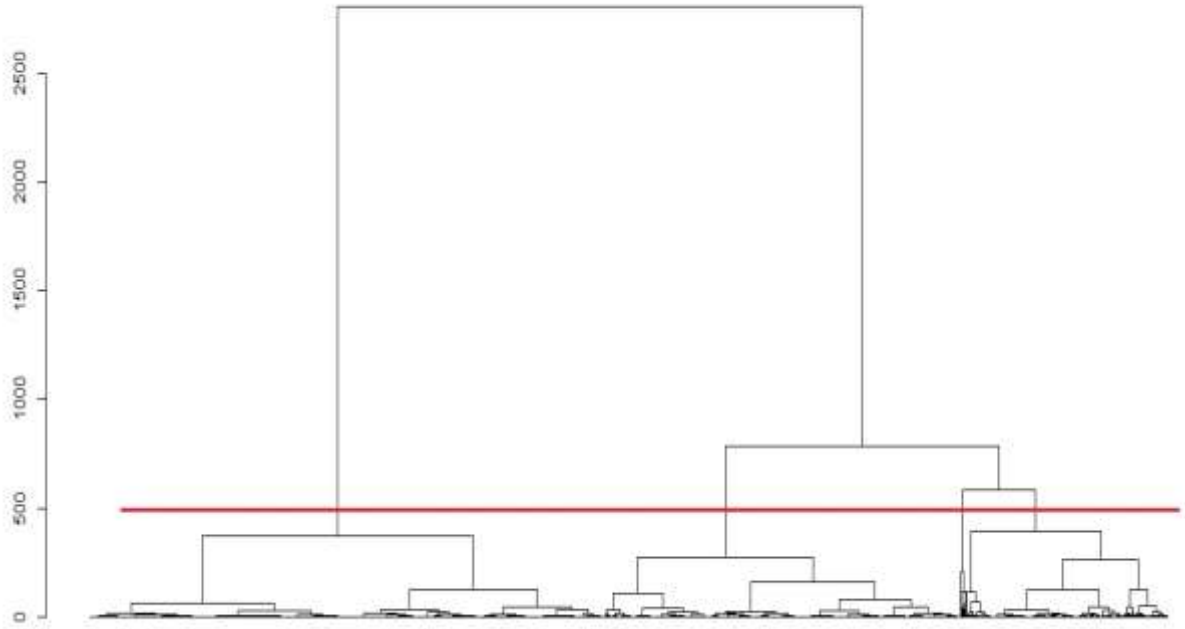
Variables	Min.	1st Q.	Median	3rd Q.	Max.	Global
X.Intercept	-36.980	-1.264	-0.107	0.751	17.740	-2.4153
Per capita income	-0.042	-0.002	0.000	0.002	0.047	0.0017
Gross Domestic Product	-0.001	0.000	0.000	0.000	0.001	0.0000
Population	-0.001	0.000	0.000	0.000	0.001	0.0001
Orthopedists	-8.699	-0.051	0.578	1.315	10.520	0.2374
Health insurance beneficiaries	-1.063	-0.001	0.000	0.001	1.162	-0.0001
Health insurance beneficiaries average ticket index	-910.500	-0.356	0.034	0.516	934.700	0.0239
Orthopedic surgical procedures (public)	-0.422	-0.014	0.001	0.024	1.242	0.0181
Orthopedic surgical procedures (private)	-334.100	-0.038	0.027	0.227	65.370	-0.0052

Source: The authors.
Note: 5560 local regressions were performed, and the distribution (5 numbers) of coefficients are shown and compared with global ones (OLS)

3.3 Cluster Analysis: K-means and Skater

K-Means. Prior to clustering data, missing data were removed and variables were rescaled for comparability. To determine the number of clusters that existed within the data, hierarchical cluster analyses were also conducted using Ward's method with Euclidean distance as a measure of distance between cases. A dendrogram from this hierarchical cluster analysis was examined to identify the cluster solutions that provided good differentiation between clusters and minimized the distance at which subsequent clusters were combined, as shown in Figure 3. Results from the hierarchical cluster analyses also indicated that four clusters existed within the data. Therefore, a four-cluster solution was opted for the data.

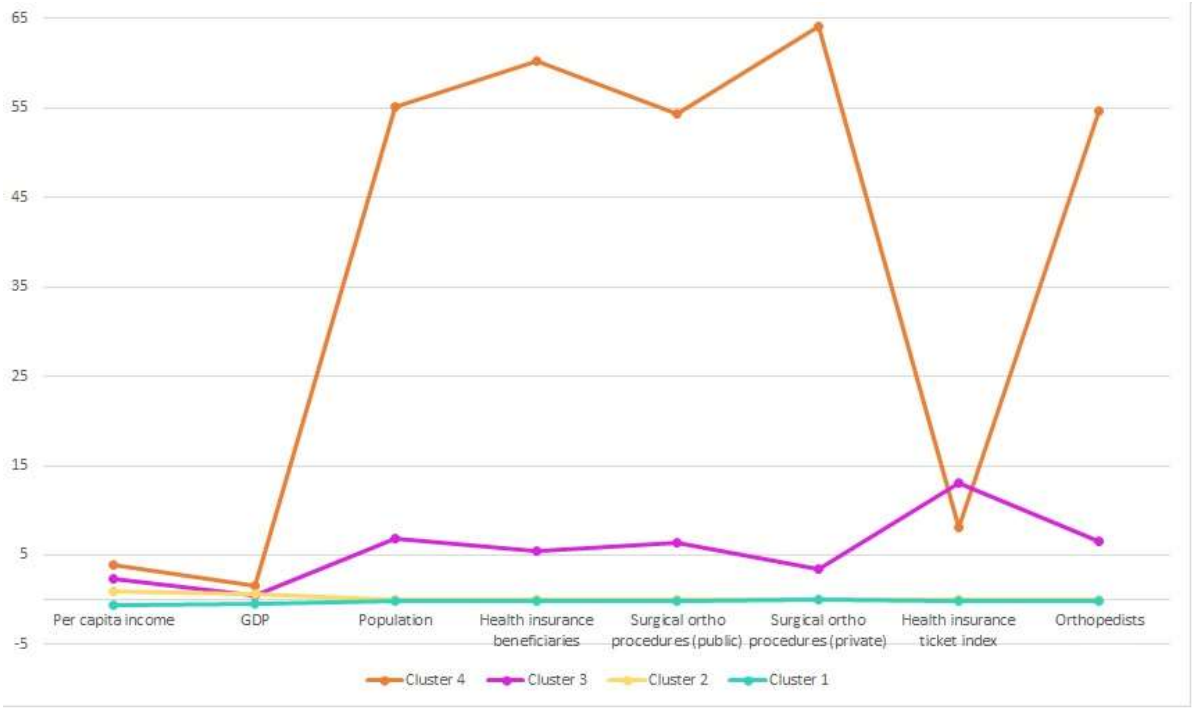
Figure 3: The dendrogram indicating number of clusters for the K-Means approach



Source: The authors, using R 3.0.1.

Figure 4 shows the mean variables for the four clusters. As shown, *per capita* income and GDP were the variables that better differentiated the four clusters, with cluster 4 and 3 showing the highest means and clusters 2 and 1, the lowest ones.

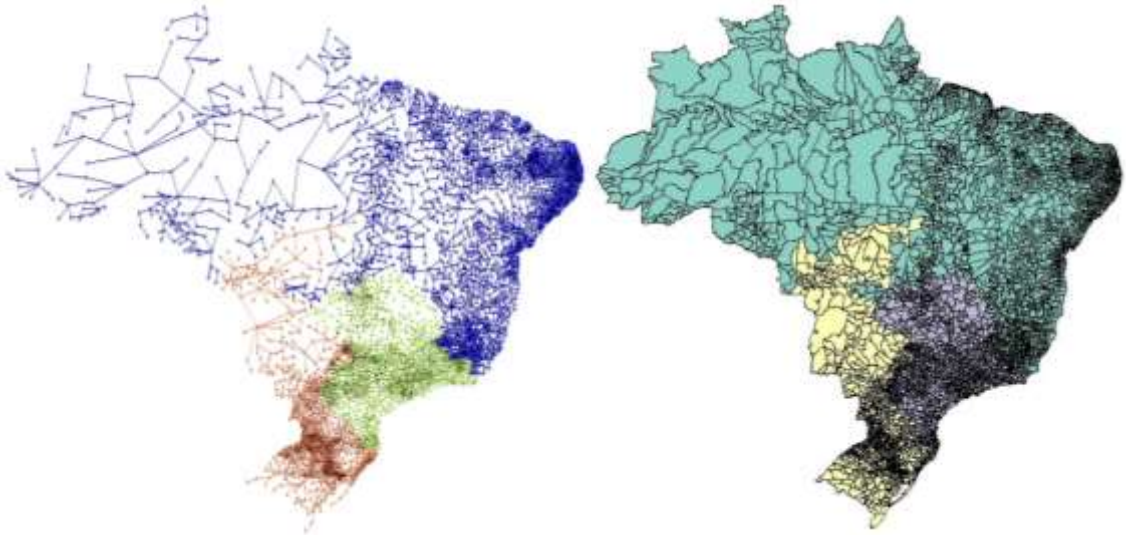
Figure 4: K-means clusters and their means on the clustering variables



Source: The authors.

'K'luster Analysis by Tree Edge Removal - Skater. As a second cluster approach, a Skater cluster analysis was performed using the same variables that were used in the K-Means approach. A four-cluster solution was applied using a contiguity matrix approach. The one-city group included Sao Paulo, which is indeed very distinct from other Brazilian cities. The Skater plot and a spatial polygon plot are shown in Figure 5. In both plots, the four-cluster solution in indicated.

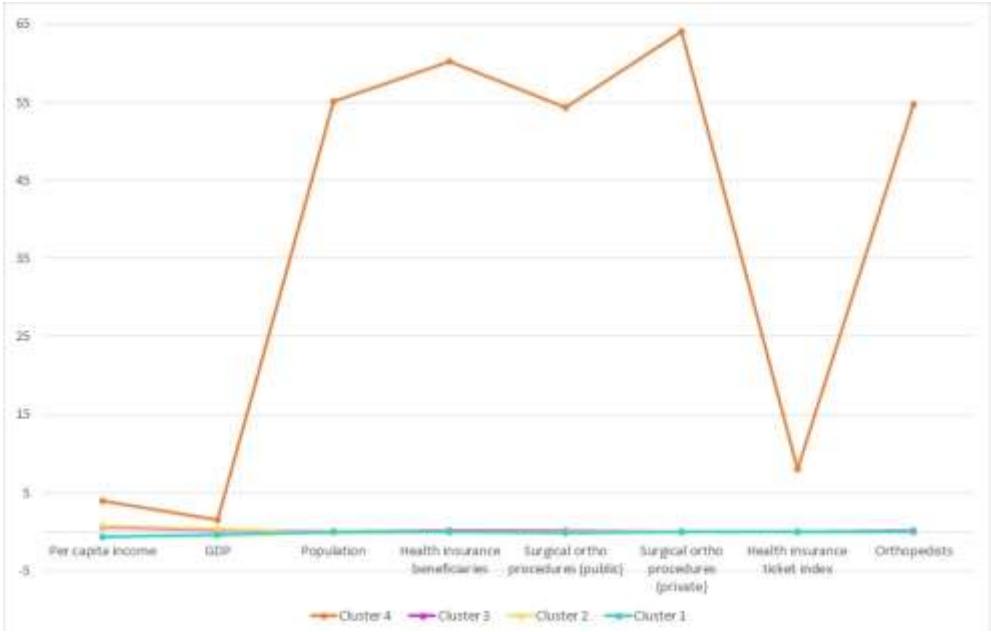
Figure 5: The Skater plot (network) (left) and Skater spatial polygon plot (right).



Source: The authors, using R 3.0.1.

Figure 6 shows the mean variables for the four Skater clusters. Confirming the K-Means approach, *per capita* income and GDP were the variables that better differentiated the four clusters, with cluster 4 showing the highest means and clusters 3, 2 and 1, the lowest ones.

Figure 6: Skater clusters and their means on the clustering variables.



Source: The authors.

An Ordinary Least Square regression model was carried out in R 3.0.1 to explore whether the obtained four skater clusters predicted number of beds in orthopedic postoperative units. To do that, the cluster variable was recoded into a dummy variable such that Cluster 4 (Sao Paulo city) was always the cluster of comparison. The model showed to be significant, $F(3,5551) = 772.10$, $p < 0.0001$. The cluster variables explained 29.44% (Adjusted R^2) of the variance in number of beds in orthopedic postoperative units. In all cases, as compared to cluster 4, the clusters predicted less number of beds in orthopedic postoperative units (i.e., cluster 4 (São Paulo city) was associated with higher number of beds in orthopedic postoperative units, see Table 6).

Table 6: OLS Regression on Number of Beds in Orthopedic Postoperative Units

Variable	Coefficient	S.E.	t-value	Signif.
Constant	1173.00	24.33	48.20	0.000
Cluster 1	-1170.34	24.34	-48.09	0.000
Cluster 2	-1171.06	24.34	-48.10	0.000
Cluster 3	-1169.14	24.34	-48.03	0.000

Source: The authors.
Note. Cluster 1, 2, and 3 as compared to Cluster 4

4. DISCUSSION

This study aimed at exploring which criteria are important to support the expansion strategy of a Brazilian healthcare distribution company focused on Orthopedics. The investigated level of granularity was city (within the whole Brazilian territory), and hence, the idea was to answer what makes a city interesting in terms of potential market. We also intended to answer whether there was a "market score" that could help the company's executives to choose the best places to expand.

The adopted analytical approaches included the use of logistic regression to compute city scores, and a series of regression models (linear and spatial ones) to allow the understanding of which variables were predictors of number of beds in orthopedic postoperative units, set as a proxy for a good market. Cluster analyses were also used to check how cities grouped according to a series of variables.

The study findings are insightful. With regards to the computation of city scores, by using a variable that served as a proxy to a good market (i.e., number of beds in orthopedic postoperative units), we were able to carry out a logistic regression model that presented a good discrimination between the supposed "good" and "bad" cities, with error rates for learning and test samples of 4.62% and 5.75%. One limitation with respect to the obtained city scores was that the within discrimination (i.e., differentiation of scores among the good (bad) cities) was not ideal. This is likely due to how the dependent variable (i.e., number of beds in orthopedic postoperative units) was dichotomized. As there are no historical data on orthopedic market indicators that can precisely index a city as having a good (bad) market, we used the strategy of choosing one variable that was considered important to a good market (i.e., number of beds in orthopedic postoperative units). This variable was dichotomized such that a certain number of beds in orthopedic postoperative units was set as a threshold for splitting the cities into potentially good or bad. Because, the number of beds in orthopedic postoperative rooms present high variance within the sample of cities (the average number of beds in orthopedic postoperative units was 3, with a variance of 837, ranging from 0 to 1173) it is possible that the used strategy to split the sample put together very distinct cities, for instance Sao Paulo, the city with the highest number of beds in orthopedic postoperative units (1173) and any other city with five or more beds. Ideally, the dichotomization strategy should have been more

conservative such that differentiation among cities was stricter. Despite this limitation, because the model showed a good differentiation between cities, the computed city scores provide a useful insight into which cities the company should look at if they want to expand their business. The model indicated that income, population, health insurance beneficiaries average ticket index, number of orthopedists, and of health insurance beneficiaries are important indicators to pay attention when considering an expansion. While income, population, health insurance beneficiaries average ticket index, and number of orthopedists are associated with more beds in orthopedic postoperative units, the number of health insurance beneficiaries is associated with less. This is interesting as it shows that what is more important is not the absolute number of health insurance beneficiaries but the health insurance beneficiaries average ticket index. A city may have many beneficiaries but their average ticket index may be low.

We carried out three regression models to determine what the important predictors of beds in orthopedic postoperative units were. The three models consistently showed that income, population, number of orthopedic surgical procedures in the public health system, health insurance beneficiaries average ticket index, and number of orthopedists were positively associated with number of beds in orthopedic postoperative units. On the other hand, health insurance beneficiaries and number of orthopedic surgical procedures in the private system were negatively associated to beds in orthopedic postoperative units. Hence, we are confident that these indicators are indeed relevant and need to be considered in the company's expansion strategy. One indicator that showed itself differently in the three models was GDP. The OLS and the SAR regression models showed that GDP was negatively associated with number of beds in orthopedic postoperative units, whereas the GWR regression model showed that it was positively associated with it.

The study findings confirm that the use of the spatial analytical approach showed to be useful. The spatial regression models showed increasingly better fit to the data, with the SAR model presenting a lower AIC and greater R^2 than the OLS model, and the GWR, a lower AIC and greater R^2 than both the OLS and SAR models.

Two cluster analytical approaches were used in this study. The idea was to explore how the 5565 cities gathered into groups. The KS cluster procedure indicated that four cluster were present in the data and therefore the SKATER clustering procedure considered a four-cluster solution. The four clusters found with SKATER spread into the Brazilian territory in a way that Southern and Central West Brazil grouped together, Northeastern and Northern Brazil grouped together, Sao Paulo city standed alone as one group, and the rest of Eastern Brazil standed as one group. Sao Paulo city is the most outstanding group presenting the higher mean values in all the clustering variables. Income and GDP were the variables that discriminated better the groups with Eastern and Southern/Central West Brazil presenting higher values than the Northeastern/Northern Brazil cluster. Importantly, the cluster variable alone predicted almost 30% of the total variance in the number of beds in orthopedic postoperative units. It is therefore an important geographical indicator for an expansion.

As mentioned earlier on, we recognize some limitations in this study. One is related to the strategy that was to dichotomize the dependent variable to be used in the logistic regression that allowed the computation of the city scores (discussed before). Another limitation has to do with the fact that the dataset built for this study included data from public datasets. Public datasets are fed by a notification system that need constant input to be updated. These public datasets might therefore not be updated with the latest information available. Another aspect is that public datasets may be fed within certain intervals of time. The census data, for example, are from 2010, which was the latest census study in Brazil. This limitation does not invalidate the study findings and, we must recognize, that public datasets are important sources of information that can offer a good opportunity for researchers to gather relevant data.

REFERENCES

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR, ANS. Dados consolidados de saúde suplementar [Consolidated data of the National Regulatory Agency for Private Health Insurance and Plans], 2017. Retrieved from http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor. Accessed February 8, 2017.

ANSELIN, L. *Spatial Econometrics, Methods and Models*. Boston: Kluwer Academic. 1988.

ASSUNÇÃO, R. M., LAGE, J. P., REIS, E. A. Análise de conglomerados espaciais via árvore geradora mínima [Spatial cluster analysis using minimum spanning trees]. *Revista Brasileira de Estatística, 63* (220), 7-2, 2002.

BARTON, D., COURT, D. Making Advanced Analytics Work for You. *Harvard Business Review*, 2012. Retrieved from https://hbr.org/2012/10/making-advanced-analytics-work-for-you. Accessed on February 8, 2017.

DE SMITH, M. J. *STATSREF: Statistical Analysis Handbook - A web-based statistics resource.* The Winchelsea Press, Winchelsea, UK, 2015.

FOTHERIGHAM, A., BRUNSDON, C., CHARLTON, M. *Geographically Weighted Regression: the analysis of spatially varying relationships.* John Wiley & Sons, 2002.

FRANCISCO, E. R. *Indicadores de renda baseados em consumo de energia elétrica: abordagens domiciliar e regional na perspectiva da estatística espacial*. Tese de Doutorado. FGV-EAESP, 2010.

FRANCISCO, E. R., FAGUNDES, E. B. ; PONCHIO, Mateus ; ZAMBALDI, F. Development of an indicator of propensity to energy commercial losses using geospatial statistical techniques and socio-economic data: the case of AES Eletropaulo. RAM. *Revista de Administração Mackenzie*, v. 11, n. 4, p. 178-197, 2010.

HOSMER, D. W., LEMESHOW, S. *Applied logistic regression.* John Wiley & Sons, Inc, 2000.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA IBGE. Sinopse do Censo Demográfico 2010 [Synopsis of the 2010 Demographic Census], 2011. Retrieved from http://www.ibge.gov.br/home/estatistica/populacao/censo2010. Accessed on January 29, 2017.

MILLER, J. *Spatial analysis: An Experian white paper*. Retrieved from http://www.experian.co.uk/assets/business-strategies/white-papers/wp-spatial-analysis.pdf. p. 2-13, 2011.

MORANDI, E. L. ; RIBEIRO, R. V. ; HERNANDES, E. G. S. ; CAMARA, B. S. ; SPINOLA, L. ; FRANCISCO, E. R. . *Análise Geoespacial da Relação entre Transporte Público sobre Trilhos, Renda e Tempo Médio de Deslocamento*. In: EnANPAD - Encontro da ANPAD, Costa do Sauípe - BA. Anais do XL EnANPAD. Costa do Sauípe - BA, 2016. v. 1. p. 1-13, 2016.

PROVOST, F., FAWCETT, T. *Data Science for Business.* O'Reilly Media, 2013.

SWEENEY, D. J., WILLIAMS, T. A., ANDERSON, D. R. *Estatística Aplicada à Administração* [Statistics applied to Business Administration] (565-613). 3.ed. São Paulo: Cengage Learning, 2013.