International Journal of Statistics and Applications 2022, 12(3): 77-82
DOI: 10.5923/j.statistics.20221203.03

Using Principal Component Analysis to Build Socioeconomic Status Indices

Wilson da C. Vieira*, José A. Ferreira Neto, Mariane P. B. Roque, Bianca D. da Rocha

Department of Agricultural Economics, Federal University of Viçosa, Brazil

Abstract This article presents a simple and effective procedure for the construction of socioeconomic status indices using principal component analysis. The methodological approach consists of obtaining principal components of the correlation matrix from a sample of random variables. For the calculation of the index, a weighted average of selected principal components is used. The proposed method is sufficiently general and can be applied to obtain other types of composite indices. To illustrate the versatility of the method, we provide in this article the calculation of a social vulnerability index for the municipalities of an area of the São Francisco river basin, Brazil, based on data from the demographic census.

Keywords Socioeconomic status index, Principal component analysis, Methodology

1. Introduction

In this article we propose a simple and effective procedure for the construction of socioeconomic status indices using principal component analysis. This type of index has aroused great interest in recent years, mainly for use in public policy design. It allows ranking and spatializing the socioeconomic status of a given locality, municipality, state, region or even an entire country. Once the socioeconomic status is ranked and spatialized, policies can be designed to target specific groups of individuals.

Although the method we propose can be used to build other types of composite indices, the focus of this article is on the construction of socioeconomic status indices. According to [1], "socioeconomic status is the social standing or class of an individual or group. It is often measured as a combination of education, income and occupation" and "examinations of socioeconomic status often reveal inequities in access to resources, plus issues related to privilege, power and control".

According to the above definition, the socioeconomic status of individuals involves multiple dimensions. This characteristic allows the creation of different socioeconomic status indices, with specific purposes, in addition to the possibility of using different methods to construct them. Regarding these indices, [2] discuss what they are, what they are for and how they are constructed. [3], in turn, review different methods used to build socioeconomic status indices.

Regardless of the method used to build composite indices, it is important to consider issues related to choosing variables, preparing data or problems such as data clustering. However, the methods used in the construction of these indices in which the choice of the weights of the variables or sub-indices is made subjectively are subject to strong criticism. The principal component analysis method is not subject to this type of criticism. In fact, when applying this method, the weights of the variables or sub-indices emerge naturally. This feature and the ease of working with multiple variables has contributed to the increasing use of principal component analysis in the construction of composite indices.

Composite indices that are constructed using principal component analysis are based on principal components drawn from the sample of variables, with each principal component being a linear combination of the original variables. Most authors who use principal component analysis to build socioeconomic status indices consider only the first principal component and its relationship to the original variables as a composite index (see, for example, [4] or [5]). Others consider only the first two principal components, but interpret them as two distinct composite indices. This is the case, for example, of [6], who developed the Institut National de Santé Publique du Québec (INSPQ) index; in their work, the first principal component comprises the weights of a "material-based" deprivation index and the second principal component comprises the weights of a "social-based" deprivation index.

The purpose of this article is twofold: first, to propose the construction of socioeconomic status indices using principal component analysis that consider not only the first principal component, but a weighted average of the first principal components. As mentioned earlier, current literature on socioeconomic status indices generally considers only the

* Corresponding author:
wvieira@ufv.br (Wilson da C. Vieira)

Received: Aug. 2, 2022; Accepted: Aug. 17, 2022; Published: Aug. 30, 2022
Published online at http://journal.sapub.org/statistics

first principal component as a composite index; and second, to illustrate the proposed method with the calculation of a social vulnerability index for the municipalities of an area of the São Francisco river basin, Brazil, using data from the demographic census.

We give the following reasons to justify constructing a socioeconomic status index as a weighted average of more than one principal component: i) in general, the first principal component explains only a small part of the variance of the original data. A composite index with more than one principal component would explain a greater portion of the variance of the original data; ii) the socioeconomic status of individuals involves multiple dimensions and hardly a single principal component could capture all these multiple dimensions.

2. Principal Components and Socioeconomic Status Indices

Principal component analysis is a statistical method that transforms a set of correlated variables into another set of uncorrelated variables called principal components (for more details on this method, see, for example, [7], [8], or [9]). These principal components are linear combinations of the original variables and must satisfy certain properties. In this transformation, information on data variability is preserved and their complexity is reduced. To obtain the principal components, the variance-covariance matrix or correlation matrix of a sample of random variables is used.

Formally, suppose the vector $x' = [x_1, x_2, \dots, x_n]$ represents a set of *n* random variables with mean $\mu' = [\mu_1, \mu_2, \dots, \mu_n]$ and variance-covariance matrix Σ. Let $z' = [z_1, z_2, \dots, z_n]$ be the random vector of the corresponding standardized variables, that is,

$$z_j = \frac{x_j - \mu_j}{\sigma_j},\ j = 1, 2, \dots, n,$$

where $\sigma_j^2 = var(x_j)$ represents the variance of variable x_j, $j = 1, 2, \dots, n$. Note that the covariance between variables z_k and z_j, $cov(z_k, z_j)$, is related to the covariance between variables x_k and x_j, $cov(x_k, x_j)$, as follows

$$cov(z_k, z_j) = \frac{1}{\sigma_k \sigma_j} cov(x_k, x_j),\ k, j = 1, 2, \dots, n,$$

that is, the variance-covariance matrix of *z* corresponds to the correlation matrix of *x*. In this article, the correlation matrix of *x* will be denoted by *C*.

Although the principal components can be obtained from the variance-covariance matrix of *x* or the correlation matrix of *x*, they are not necessarily the same. This implies that the interpretation of the results must take into account the choice of the matrix that will be used to extract the principal components. [10] recommend using the correlation matrix to extract principal components when the scales of variables vary widely or they have very different variances. In this article, the analysis will be carried out with the correlation matrix, since the variables generally used to obtain socioeconomic status indices are diverse and with very different variances.

In this sense, the principal components $p_1, p_2, \dots, p_n$ are associated with the random vector *z*, such that

$$p_j = a_{1j} z_1 + a_{2j} z_2 + \dots + a_{nj} z_n,\ j = 1, 2, \dots, n,$$

where a_{ij}, $i, j = 1, 2, \dots, n$, are constants that satisfy certain conditions. It can be shown that the mean of p_j is equal to zero, $\mu_{p_j} = 0$, and its variance is given by $var(p_j) = a_j' C a_j$, where $a_j' = [a_{1j}, a_{2j}, \dots, a_{nj}]$.

The principal components are obtained sequentially: first, p_1 is selected to capture as much of the variation in the original data as possible amongst all linear combinations of *z* such that $a_1' a_1 = 1$. Then p_2 is selected to account for a maximum proportion of the remaining variance subject to not being correlated with the first principal component, $a_2' a_1 = 0$, and $a_2' a_2 = 1$. Subsequent principal components are obtained in a similar manner. Formally, the *j*th principal component is the linear combination $p_j = a_j' z$ that has the greatest variance subject to the following conditions

$$a_j' a_j = 1,$$
$$a_k' a_j = 0\ (k > j).$$

As it is an optimization problem with equality constraints, the Lagrange method can be used to obtain the solution (see, for example, [7]). The results of applying this method show that the vector of coefficients that defines the *j*th principal component, a_j, is the eigenvector of the matrix *C* associated with its *j*th largest eigenvalue. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the *n* eigenvalues of *C*. It can be shown that $var(p_j) = \lambda_j$, that is, the variance of the *j*th principal component is equal to the eigenvalue λ_j. It can also be shown that $\sum_{j=1}^{n} var(p_j) =$ *j=1nvarzj=n*. Thus, the proportion of the total variance of the standardized variables explained by the *j*th principal component is given by

$$\frac{\lambda_j}{n},\ j = 1, 2, \dots, n,$$

and the percentage of the total variance explained by the *m* first principal components, $1 < m \leq n$, is given by

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_m}{n} \times 100\%.$$

According to [11], "applied principal component analysis consists most often of a mere computation of eigenvectors and eigenvalues of a sample covariance matrix or correlation matrix" (p. 606). That's largely what we are going to do in this article. To start, we summarize the main results of the principal component analysis related to eigenvalues and eigenvectors that will be useful for the construction of socioeconomic status indices in the following properties:

(i) $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$;
(ii) $\sum_{j=1}^{n} \lambda_j = \sum_{j=1}^{n} var(z_j) = n$;
(iii) $a_j' a_j = 1,\ j = 1, 2, \dots, n$;
(iv) $a_k' a_j = 0,\ j \neq k,\ k, j = 1, 2, \dots, n.$

As mentioned in the introduction, some authors consider only the first principal component, p_1, as a socioeconomic status index and others consider the first two main components, p_1 and p_2, but as two distinct indices. In this article we propose the construction of a socioeconomic status index as a weighted average of the first m, $1 < m \leq n$, principal components. The idea behind this proposal is that the first few principal components will represent a substantial proportion of the variation in the original variables and can therefore be used to provide a convenient lower-dimensional summary of these variables.

In this sense, we can construct a socioeconomic status index (SSI) as a linear combination of all the principal components as follows

$$SSI = b_1 p_1 + b_2 p_2 + \cdots + b_n p_n,$$

where the weight vector, $b' = [b_1, b_2, \dots, b_n]$, with $\sum b_j = 1$, is given by

$$b'_{(n\times1)} = \left(\frac{a_{ij}}{\sum a_j}\right)_{(n\times n)} \cdot \left(\frac{\lambda_j}{\sum \lambda_j}\right)_{(n\times1)}, \ i,j = 1,2,\dots,n. \quad (1)$$

Note that the only information needed to construct the socioeconomic status index indicated above are the scores of the principal components and the values of the eigenvalues and eigenvectors obtained from the sample of random variables. From a practical point of view, it is usual to consider the first few principal components as long as they satisfy some criterion. In practical applications in the area of social sciences and humanities, [8] suggests choosing the first principal components until reaching at least 60% of the total variation of the original data. Assume that the first m^*, $1 < m^* < n$, principal components satisfy the criterion of [8], then the socioeconomic status index is given by

$$SSI = b_1^* p_1 + b_2^* p_2 + \cdots + b_{m^*}^* p_{m^*},$$

where $b^{*'} = [b_1^*, b_2^*, \dots, b_{m^*}^*]$ represents the vector of corrected weights, such that $\sum_{j=1}^{m^*} b_j^* = 1$. Note that if we were to use the original weights $b_1, b_2, \dots, b_{m^*}$, we would have $\sum_{j=1}^{m^*} b_j < 1$, $1 < m^* < n$, a result whose sum of weights is not equal to 1. To correct the weights, we use the following expression:

$$b_j^* = \frac{1+(m^*-1)b_j - \sum_{q=1,q\neq j}^{m^*} b_q}{m^*}, \ j = 1,2,\dots,m^*.$$

This correction of the weights is fundamental to obtain a socioeconomic status index as a weighted average of the first principal components. To illustrate this correction of weights, suppose that the first three principal components were selected to construct a socioeconomic status index and that the original weights are b_1, b_2 and b_3. Applying the correction formula, knowing that $m^* = 3$, we have

$$b_1^* = \frac{1+2b_1-b_2-b_3}{3};$$

$$b_2^* = \frac{1+2b_2-b_1-b_3}{3};$$

$$b_3^* = \frac{1+2b_3-b_1-b_2}{3}.$$

Note that, after correction, we have $\sum_{j=1}^{3} b_j^* = 1$. It is important to keep in mind that this procedure for constructing a socioeconomic status index must take into account all variables and all observations of each variable. Suppose you want to build a socioeconomic status index from a sample of 10 variables $(j = 1,2,\dots,10)$ and each variable contains 100 observations $(l = 1,2,\dots,100)$. In this case, the principal components of each observation are calculated, that is,

$$p_{j,l} = a_{1j} z_{1,l} + a_{2j} z_{2,l} + \cdots + a_{10j} z_{10,l},$$
$$j = 1,2,\dots,10; \ l = 1,2,\dots,100,$$

where $p_{j,l}$ represents the jth principal component of the lth observation. After calculating the 10 principal components associated with each to the 100 observations, the weights $b_1, b_2, \dots, b_{10}$ corresponding to the sample of variables can be obtained (see expression for determining the vector of weights b above). Suppose further that three principal components were selected to build the socioeconomic status index. In this case, this index is constructed for each observation of the sample of variables as follows:

$$SSI_l = b_1^* p_{1,l} + b_2^* p_{2,l} + b_3^* p_{3,l}, \ l = 1,2,\dots,100,$$

where b_j^*, $j = 1,2,3$, represents the corrected weight, and $p_{j,l}$ denotes the principal component j associated with observation l of the sample of variables. After carrying out all the calculations, one obtains, as a result, an interval composed of 100 socioeconomic status indices (one index for each observation of the sample of variables) that can divided equally or using some other criterion to form the socioeconomic status classes (levels). This number of classes is defined according to the purpose of the study or to meet public policy interests. Commonly used arbitrary cut-off points classify the lowest 40% of individuals as 'poor', the highest 20% as 'rich' and the remainder as the 'average' group (see, for example, [12]).

To avoid an eventual negative component in the weight vector, b, another possibility to define weights is to use the expression (2) given in the following proposition

Proposition. If the $n \times n$ correlation matrix $C = [c_{ij}]$ is positive definite and the eigenvectors associated with C are such that $a_j' a_j = 1$, $a_i a_i' = 1$ and $a_i a_j' = 0$, $i \neq j$, $i,j = 1,2,\dots,n$, then the weight vector b given by

$$b'_{(n\times1)} = \left(\frac{(a_{ij})^2}{a_j' \cdot a_j}\right)_{(n\times n)} \cdot \left(\frac{\lambda_j}{n}\right)_{(n\times1)}, \ i,j = 1,2,\dots,n. \quad (2)$$

satisfies the properties $\sum_{j=1}^{n} b_j = 1$, $b_j > 1$, $j = 1,2,\dots,n$, and $b_1 = b_2 = \cdots = b_n = 1/n$.

Proof. From the expression $Ca_j = \lambda_j a_j$, we have $a_{ij} + \sum_{k=1,k\neq i}^{n} c_{ik} a_{kj} = \lambda_j a_{ij}$, $i,j = 1,2,\dots,n$, and the eigenvalues are given by $\lambda_j = \left(a_{ij} + \sum_{k=1,k\neq i}^{n} c_{ik} a_{kj}\right)/a_{ij}$, $i,j = 1,2,\dots,n$. Substituting the expression for λ_j in (2), we have $b_j = \frac{\sum_{i=1}^{n} c_{ji}(a_i a_i')}{n} = \frac{1}{n}$. Since $j = 1,2,\dots,n$ is arbitrary, the proof is complete.

If we use expression (2) to define the weight vector b, no correction is needed if $1 < m^* < n$, where m^* is the number of principal components used to build the composite index. In this case, $b_1 = b_2 = \cdots = b_{m^*} = 1/m^*$.

Standard statistical software (such as STATA or SPSS) can be used to perform the necessary calculations and build composite indices. In the illustration of the use of the proposed method in the next section, the statistical analysis was performed with the R software ([13]) and ArcGis® ([14]) was used to spatialize the results (vulnerability classes) of the study area.

3. Results: Social Vulnerability Index

This section presents the calculation of a social vulnerability index for the municipalities of an area of the São Francisco River Basin as un illustration of the method proposed in this article. This basin has a drainage area of about 630 thousand square kilometers and includes 521 municipalities belonging to six different states (Minas Gerais, Goiás, Bahia, Pernambuco, Alagoas e Sergipe) plus the Federal District. The basin has a population of over 16 million inhabitants, with a large part of it, around 77% of the total, inhabiting urban areas. The São Francisco River is often called the "river of national integration" because it unites different physiographic regions of the country, especially the Southeast and Northeast.

Due to its large extension and diversity, the São Francisco River Basin is divided into Upper, Middle, Sub-medium and Lower São Francisco. According to [15], taking 2014 as a reference, the shares of these regions in the Gross Domestic Product (GDP) of the basin are as follows: Upper São Francisco (86.6%), Middle São Francisco (4.9%), Sub-medium São Francisco (5.4%) and Lower São Francisco (3.1%). This distribution of GDP in the basin highlights the economic discrepancy between these regions, showing that most of the wealth is generated in the Upper São Francisco.

To calculate the social vulnerability index, we considered only part of the Upper São Francisco region. This study area is equal to 58,204.65 square kilometers and comprises 105 municipalities. Figure1 locates the São Francisco River basin in Brazil and also shows this study area. The São Francisco River originates in this selected area, more specifically in Serra da Canastra, in the central-western part of the state of Minas Gerais. This is a predominantly urban area and is home to various economic activities, such as steel production, mining, textile industry, chemical industry and industrial equipment.

To calculate the social vulnerability index, we selected 12 variables from the last demographic census ([16]). This selection of variables took into account the international literature and, in the Brazilian context, the work of [17]. These variables were classified into three categories: human capital, urban infrastructure and occupation/income and are listed below.

Human capital

- % of illiterate children from five to 14 years old
- % of illiterate female heads of households
- % of people aged 15 and above
- Infant mortality (per 1,000 live births)

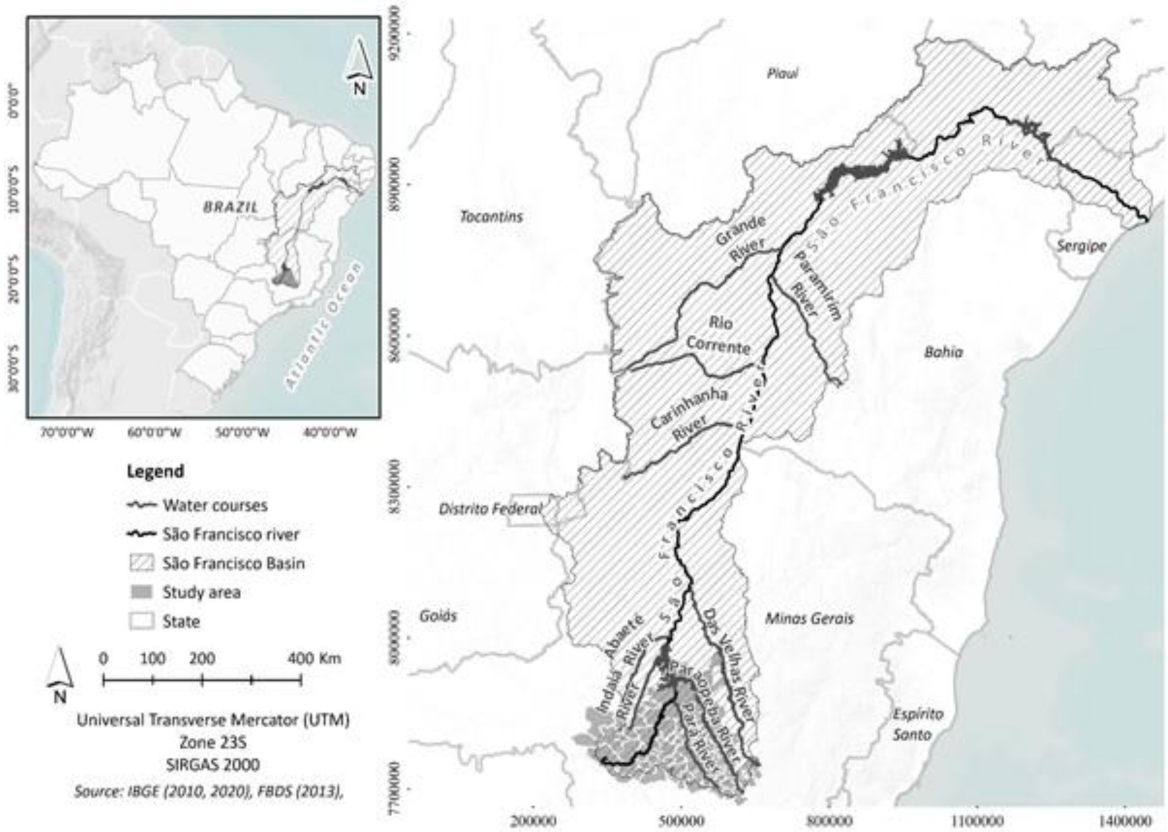

Figure 1. São Francisco River basin and study area

Urban infrastructure

- % of dwellings with inadequate sewage disposal systems
- % of dwellings without access to the general supply water network
- % of families with an income of less than one minimum wage and with inadequate housing conditions
- % of dwellings without access to the general electricity grid

Occupation/income

- % of people earning up to minimum wage
- % of responsible people with no monthly nominal income
- % of people without incomes of their own
- % of families that depend on the income of older adults

All variables are in percentages, except infant mortality. The percentages were obtained by dividing the value of each variable in the municipality by the population of the respective municipality. Note that for the calculation of the social vulnerability index we have 12 variables and 105 observations for each variable. After preparing the data, the principal component analysis was carried out, first obtaining the correlation matrix of the variables. From the principal components extracted from the correlation matrix, we selected the first four that explain 70.89% of the variance of the original data and we used expression (2) to define the vector of weights in the construction of the social vulnerability index.

The values of the social vulnerability index obtained from the sample of variables ranged from a minimum of 0.05244 (Belo Horizonte) to a maximum of 0.313853 (Piedade dos Gerais). We divided the range of variation of this social vulnerability index equally into three classes, with break values equal to 0.140, 0.227 and 0.314. These vulnerability classes (or levels) were called “low” (variation from 0.052 to 0.140), “moderate” (variation from 0.141 to 0.227), and “high” (0.228 to 0.314). The vulnerability results obtained for these different classes were spatialized for better visualization (see Figure 2).

The municipalities classified as less socially vulnerable, according to our classification, are Belo Horizonte, Itaúna, Contagem, Nova Serrana, Pará de Minas and Sete Lagoas. On the other hand, the municipalities classified as the most vulnerable in the study area are Piedade dos Gerais, Itaverava, Rio Manso, Esmeraldas, Moeda, Desterro de Entre Rios, Felixlândia, Jeceaba, Quartel Geral and Serra da Saudade. The other municipalities in the study area belong to the moderate class of social vulnerability. Note that only ten municipalities out of the 105 considered in our study (9.52%) were classified as the most socially vulnerable. This result is consistent with the socioeconomic information available for the São Francisco River basin, especially those for the Upper São Francisco region.

Once the social vulnerability of a given area or region is ranked and spatialized, public policies can be designed to target (directly or indirectly) specific groups of individuals. Another alternative way of using this type of information is to design differentiated policies considering simultaneously different groups of individuals. In this case, considering more than three classes of social vulnerability would allow a more detailed and accurate ranking.

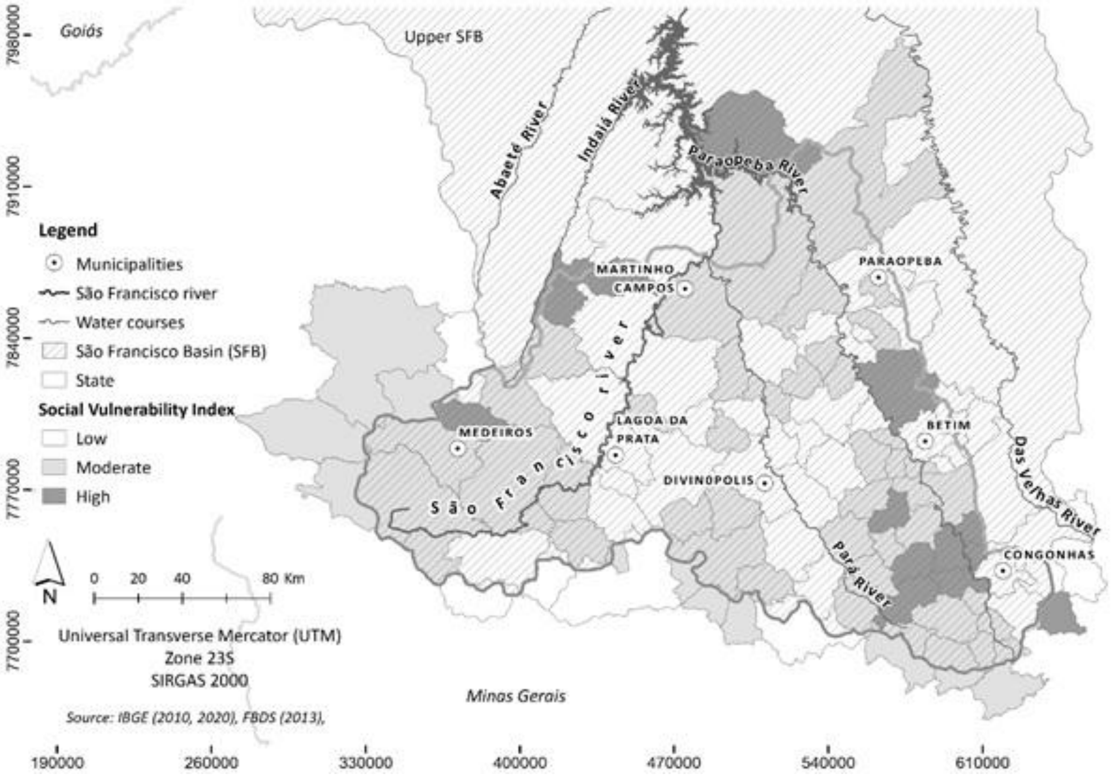

Figure 2. Social vulnerability for the municipalities of an area of the São Francisco River basin

4. Conclusions

In this article we present a simple and effective method for building socioeconomic status indices based on principal component analysis. The index is calculated as a weighted average of principal components selected from among those extracted from the correlation matrix of a sample of random variables. The calculation of a social vulnerability index for municipalities in an area of the São Francisco river basin, Brazil, using data from the demographic census was used to illustrate the method.

The proposed method is sufficiently general and can be applied to obtain other types of composite indices and not just socioeconomic status indices. As the index obtained by the principal components analysis is based on a set of random variables, the choice of these variables is fundamental to obtain a reliable and useful index to meet the objectives that motivated its construction. In this sense, the set of selected variables must be sufficient to represent with some accuracy the indicator being measured, such as the socioeconomic status of individuals. In most cases, it is also necessary to prepare the data or do some transformation of variables before obtaining the correlation matrix. [18] discuss these and other issues associated with the construction of indices using principal component analysis.

Finally, although the presentation of the proposed method was based on the correlation matrix to extract the principal components, the same procedure can be done considering the covariance-variance matrix of the original data. In this case, it is convenient that the variances of the original variables are close to each other. It is also possible to use other criteria for the selection of the first principal components in the construction of the composite index and not just the criterion presented in this article.

REFERENCES

[1] APA - American Psychological Association, Socioeconomic Status. Available at: http://www.apa.org Accessed: April, 2022.

[2] Figueiredo Filho, D. B., Paranhos, R., Rocha, E. C., Silva Jr., J. A., and Maia, R. G., 2013, Análise de componentes principais para construção de indicadores sociais, Revista Brasileira de Biometria, 31, 61-78.

[3] Vincent, K. and Sutherland, J. M., A review of methods for deriving an index for socioeconomic status in British Columbia, Vancouver, British Columbia: UBS Centre for Health Services and Policy Research (Technical Report), 2013.

[4] Messer, L. C., Laraia, B. A., Kaufman, J. S. et al., 2006, The development of a standardized neighborhood deprivation index, Journal of Urban Health: Bulletin of the New York Academy of Medicine, 83, 1041-1062.

[5] Avila-Vera, M., Rangel-Blanco, L. and Picazzo-Palencia, E., 2020, Application of principal component analysis as a technique to obtain a social vulnerability index for the design of public policies in Mexico, Open Journal of Social Sciences, 8, 130-145.

[6] Pampalon, R. and Raymond, G., 2000, A deprivation index for health and welfare planning in Quebec, Chronic Diseases in Canada, 21, 104-113.

[7] I. Jolliffe, I., Principal Component Analysis, New York: Springer-Verlag, 2002.

[8] J. F. Hair, C. W. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis, 7th ed., New York: Person, 2009.

[9] R. A. Johnson. and D. W. Wichern, Applied Multivariate Statistical Analysis, 6th ed., New Jersey: Prentice-Hall, 2007.

[10] B. Everitt and T. Horthorn, An Introduction to Applied Multivariate Analysis with R, New York: Springer, 2011.

[11] B. Flury, A First Course in Multivariate Statistics, New York: Springer, 1997.

[12] Filmer, D. and Pritchett, L. H., 2001, Estimating wealth effect without expenditure data – or tears: an application to educational enrollments in states of India, Demography, 38, 115-32.

[13] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna: Austria. URL: http://www.r-project.org, 2020.

[14] ESRI. Mapping Products – GIS Software Products – Esri. https://www.esri.com/enus/arcgis/products/index?medium=www_esri_com_EtoF&rsource=/enus/arcgis/products, 2020.

[15] C. N. Castro and C. N. Pereira, Revitalização da Bacia Hidrográfica do Rio São Francisco: Histórico, Diagnóstico e Desafios. Brasília: IPEA, 2019.

[16] IBGE – Instituto Brasileiro de Geografia e Estatística, Base de Informações do Censo Demográfico, Rio de Janeiro: IBGE, 2010.

[17] M. A. Costa and B. O. Marguti, Atlas da Vulnerabilidade Social nos Municípios Brasileiros. Brasília: IPEA, 2015.

[18] Vyas, S. and Kumaranayake, L., 2006, Constructing socio-economic status indices: how to use principal components analysis, Health Policy and Planning, 21, 459-468.