UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Matemática
Curso de Bacharelado em Estatística

RELATÓRIO FINAL TCC2

ANÁLISE MULTIVARIADA DE PAÍSES DA AMÉRICA DO SUL POR MEIO DE INDICADORES SOCIOECONÔMICOS

Discente: Geovanna Dias Costa
Orientadora: Prof.ª Dra. Priscila Neves Faria

Uberlândia, Dezembro de 2019

Agradecimentos

Gostaria, e bem mais do que isso, eu devo dedicar esse trabalho as duas pessoas que foram e são o meu alicerce e apoio frente ao meus sonhos, como este da graduação, meus pais Magna Rosa Dias Costa e José Adonizete da Costa.

Vocês são os maiores responsáveis por hoje eu ter chegado onde cheguei, e digo isso em todos os âmbitos da minha vida, sempre foram os maiores incentivadores e colaboradores dos meus sonhos, o restante das pessoas os vêm apenas como patrocinadores financeiros mas só nós três sabemos que, nessa jornada de 5 anos, o dinheiro foi de suma importância mas de forma alguma o principal. Quantas me vi incapaz de concluir essa graduação e estava prestes a desistir, incontáveis eu diria, e em todas elas vocês estavam lá com todo o amor, dedicação e afeto, acreditaram em mim quando eu já não acreditava, nunca duvidaram de que eu fosse capaz e motivada por isso eu me fiz capaz, por vocês, pela luta de vocês, por todos os sacrifícios, motivada por um dia me tornar orgulho de vocês da mesma forma que sempre foram pra mim. Nem todas as palavras e gestos do mundo são capazes de demonstrar o tamanho da minha gratidão. Obrigada mamãe. Obrigada papai.

Resumo

O estudo dos indicadores socioeconômicos de uma região é indispensável quando deseja-se avaliar seu desenvolvimento bem como propor ações de melhorias. Para isso o uso de análises estatísticas pode ser um grande aliado na obtenção de bons resultados. Diante de tal contexto, esse trabalho pretende aplicar a técnica estatística Análise de Cluster afim de agrupar os países da América do Sul com certas similaridades. Tais análises terão como base os dados desses países sobre PIB, PIB PPC, IDH e Índice de Gini retirados do site Planilhas – Bit a Bit, todas as rotinas foram implementadas no software Past. Dentre os resultados obtidos na análise de agrupamentos, obtivemos a divisão dos países em três grupos, sendo um deles composto unicamente pelo Brasil, país que mais apresentou destaque em vários aspectos analisados quando comparada aos demais países em estudo.

Palavras chaves: *Indicadores socioeconômicos, Cluster, América do Sul.*

Abstract

Studying the socioeconomic indicators of a region is indispensable when evaluating its development as well as improvement actions. For this, the use of statistical statistics can be a great ally in the search for good results. Given this context, this work intends to apply a cluster statistical analysis technique to group South American countries with certain similarities. These statistics are based on data from these countries on GDP, PPP GDP, HDI and Turn Index from Spreadsheets - Bit by Bit website, all routines have been implemented in Past software. Among the results obtained in the cluster analysis, the countries are divided into three groups, one of them being composed solely of Brazil, a country that shows more prominence in several aspects analyzed when compared to the other countries under study.

Keywords: *Socioeconomic Indicators, Cluster, South America*

1. Introdução

Os indicadores socioeconômicos são medidas usadas para expressar quantitativamente um conceito social abstrato e informar sobre determinadas características de uma realidade social. Eles são de suma importância quando o objetivo é alavancar o desenvolvimento de um determinado território, pois permitem formulação de estratégias precisas.

Segundo Soligo (2012, p.14) os indicadores econômicos são um conjunto de dados estatísticos que servem para "medir e transformar essas medidas em índices utilizados para revelar e sinalizar diversos aspectos da sociedade". Eles são comumente utilizados também na comparação de um país em relação ao seu passado ou de sua posição em relação a outros países.

Os principais indicadores são Produto Interno Bruto (PIB), renda per capita, Índice de Desenvolvimento Humano (IDH), Coeficiente de Gini, taxa de desemprego e a oferta de serviços públicos.

O Produto Interno Bruto (PIB) quantifica em valores monetários a soma dos bens e serviços finais produzidos numa determinada região (país, estado, cidade) durante um período determinado (mês, trimestre, ano, etc). O PIB se diferencia em real e nominal, sendo o primeiro calculado com base nos preços do ano em que o produto foi produzido e comercializado (preços correntes) e o segundo calculado considerando o ano escolhido pelo pesquisador (preços constantes), o que elimina o efeito da inflação.

No entanto, em países em desenvolvimento os salários e os preços de alguns bens e serviços tendem a ser menores. Com isso, uma unidade de uma moeda local qualquer tende a ter um poder de compra maior nesses países do que nos países desenvolvidos (BOURNOT et al., 2011). Consequentemente, o PIB de um país em desenvolvimento e o consumo de seus residentes será subestimado caso as taxas de câmbio de mercado sejam utilizadas em conversões para comparar seus valores com aqueles dos países mais ricos. A alternativa que se propõe frente às taxas de câmbio de mercado é a utilização das estimativas de paridade do poder de compra (PPC) (SILVA, 2003).

A PPC é uma taxa de câmbio alternativa que procura refletir o poder de compra das moedas locais e são as taxas nas quais a moeda de um país deve ser convertida para a de outro para que se torne possível comprar a mesma quantidade de bens e serviços em ambas as economias. As PPCs são estimadas via coletas de preços em países ao redor do mundo, e formação de números índices de preços com esses dados. Segundo Deaton e Heston (2010), da mesma forma que índices de preços dentro de países, as PPCs podem ser vistas como uma média estatística de preços ou uma dada interpretação do custo de vida.

A fim de eliminar desigualdades oriundas de comparações entre economias que possuem diferentes moedas, foi utilizada a Paridade de Poder de Compra (PPC), o PIB PPC corrige o cálculo original, ou seja, sinaliza o quanto uma determinada moeda pode comprar em termos internacionais, aproximando seu valor da capacidade econômica real do país.

Criado em 1990 pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), a partir do trabalho de dois economistas, o paquistanês Mahbub Ul Haq e o indiano Amartya, o Índice de Desenvolvimento Humano (IDH) tem por objetivo avaliar a qualidade de vida e o desenvolvimento econômico de uma população pautado em três critérios: saúde, educação e renda. O IDH pode assumir valores entre 0 e 1, sendo que quanto mais próximo de 1 for o IDH de um país significa que mais desenvolvido ele é. A titulo de curiosidade, o Brasil ocupou o 75° lugar no Relatório de Desenvolvimento Humano de 2015, com IDH de 0.755. O IDH é apresentado como um contraponto ao PIB, que considera apenas a dimensão econômica do desenvolvimento.

O Índice de Gini leva esse nome em homenagem ao seu criador, o matemático italiano Conrado Gini, e trata-se de um medidor de concentração de renda. Na prática, ele compara os 20% mais pobres de um grupo com os 20% mais ricos desse mesmo grupo, assumindo valores de zero a um (alguns apresentam de zero a cem) em que o valor zero representa igualdade (mesma renda) e um (ou cem) o extremo oposto (um único grupo detém toda a riqueza).

Os quatro medidores supracitados foram os alvos de estudo principais nesse trabalho. Em termos conceituais, estudos a respeito de indicadores socioeconômicos visam comparar a realidade de regiões e analisar seus desenvolvimentos a fim de subsidiar informações sobre as atividades de planejamento público, possibilitando o monitoramento das condições de vida e bem-estar da população, o que é fundamental, principalmente, em termos de gestão pública.

Kubrusly (2001) utilizou duas técnicas de análise multivariada em seu trabalho sobre índices econômicos: a Análise de Agrupamento e a Análise de Componentes Principais. Em muitos casos pode ser interessante aproveitar o poder altamente descritivo dessas duas técnicas para interpretação do índice resultante. Em particular, a Análise de Componentes Principais analisa a matriz de correlação das variáveis, e por seu resultado é possível saber se um único índice é adequado para a ordenação, ou se o conjunto de variáveis fornece duas ou mais dimensões igualmente importantes (KUBRUSLY, 2001).

Além disso, a importância de classificar e aglomerar países em grupos com características em comum gera extração de informações e assim é possível alcançar resultados mais assertivos. Para isso é possível contar com o uso da técnica estatística Análise de Agrupamentos (AA), em que se classifica elementos em grupos (*clusters*), de uma forma em que elementos dentro de um mesmo cluster sejam muito parecidos, e elementos em clusters diferentes sejam distintos entre si.

Perante o exposto, os objetivos dessa pesquisa são estudar as diferenças das condições socioeconômicas da população dos países da América do Sul mediante um conjunto de indicadores socioeconômicos; aplicar a Análise de Agrupamentos nos dados a fim de obter o gráfico Dendrograma; analisar o impacto do uso de indicadores socioeconômicos como PIB, renda per capita, IDH e Coeficiente de Gini na posição de um país (em termos econômicos) em relação a outros países; avaliar o desenvolvimento socioeconômico de países que compõem a América do Sul; agrupar países com características semelhantes por meio de clusterização e, por fim, retratar a importância de se utilizar a técnica de Componentes Principais para destacar quais características econômicas sob análise concentram a maior parte da variabilidade contida nos dados.

2. Material e Métodos

2.1. Análise de Componentes Principais (ACP)

Segundo Morrison (1976), o primeiro componente principal (Y_1) de um conjunto de p variáveis, $X_1, X_2, \dots, X_p$, contidas no vetor $X' = (X_1, X_2, \dots, X_p)$ é definido como a combinação linear $Y_1 = b_{11}X_1 + b_{21}X_2 + \cdots + b_{p1}X_p = \boldsymbol{b'_1 X}$, cujos coeficientes b_{i1} são elementos do vetor característico b_1 , associado à maior raiz característica (λ_1) da

matriz de covariância amostral, S, das variáveis $X_{i's}$. Os autovalores (ou raízes características) ordenados, ou seja, $\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_p \geq 0$, são as variâncias amostrais dos componentes principais.

A técnica envolve a matriz de covariância amostral S ou a matriz de correlação R, uma vez que a maioria das aplicações envolve esta última, pois, frequentemente, as variáveis têm escalas diferentes, apresentando então uma necessidade de padronização. O grau de influência que cada variável X_j tem sobre a componente Y_i é determinado por sua correlação, definida por:

$$Corr(X_j, Y_i) = R_{X_j.Y_i} = b_{1j}.\frac{\sqrt{Var(Y_i)}}{\sqrt{Var(X_j)}}$$

$$= \sqrt{\lambda_1}.\frac{b_{1j}}{\sqrt{Var(X_j)}}$$

As variáveis padronizadas, $Z_1, \dots, Z_p$, são expressas por $Z_j = \frac{(X_j - \mu_j)}{\sqrt{\sigma_{jj}}}$, com $j = 1, \dots, p$. A contribuição do $j - ésimo$ componente na explicação da variação total pode ser expressa por

$$Contr(Y_j) = \frac{Var(Y_j)}{\sum_{j=1}^{p} Var(Y_j)} = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

Para a determinação do número de Componentes Principais a serem utilizados na análise pode-se utilizar diversos critérios de escolha, sendo os principais: o Critério de Kaiser, o Diagrama dos Autovalores e o critério da porcentagem da variância acumulada, salientando que a escolha de um critério específico pode levar a diferentes resultados, constituindo-se então de uma decisão subjetiva e explanatória a critério do pesquisador (SILVA e PADOVANI, 2006).

O Critério de Kaiser determina que o número de componentes deve ser igual ao número de autovalores maiores ou iguais à média das variâncias das variáveis analisadas. Quando a análise é feita sobre a matriz de correlação (variáveis padronizadas), esse critério corresponde à exclusão de componentes com autovalores inferiores a um, segundo Artes (1998).

2.2. Análise de Agrupamento (Cluster Análise)

A Análise de Agrupamentos (AA) tem por objetivo reorganizar objetos (indivíduos, elementos) em grupos de forma que exista homogeneidade dentro do grupo

(e heterogeneidade entre os grupos), ou seja, elementos de um mesmo grupo possuem semelhanças entre si e se diferem dos elementos dos outros grupos (SOUZA; VICINI, 2005).

A AA constitui uma metodologia numérica multivariada, com o objetivo de propor uma estrutura classificatória, ou de reconhecimento da existência de grupos, objetivando, mais especificamente, dividir o conjunto de observações em um número de grupos homogêneos, segundo algum critério conveniente de similaridade (ou dissimilaridade) (TOMAZ et al., 2010).

O termo dissimilaridade surgiu em função da relação da distância entre dois pontos P e Q, definida como d(P,Q), pois, à medida que ela cresce, diz-se que a divergência entre os pontos (unidades amostrais) P e Q aumenta, ou seja, tornam-se cada vez mais dissimilares. Os valores de distâncias são geralmente obtidos a partir de informações de “n” unidades amostrais, mensurados em relação a “p” caracteres (variáveis).

É necessário especificar um coeficiente de semelhança que indique a proximidade entre os indivíduos sendo importante considerar, em todos os casos semelhantes a este, a natureza da variável (discreta, contínua, binária) e a escala de medida (nominal, ordinal, real ou razão).

Conforme descrito por Johnson e Wichern (1998), cita-se como medida de dissimilaridade a Distância Euclidiana, que é insatisfatória para algumas situações estatísticas, como por exemplo quando há diferentes unidades de medida. Isso ocorre devido à contribuição de cada coordenada ter o mesmo peso para o cálculo da distância.

A "distância euclidiana" preconiza que a "distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j), para todas as p variáveis" (FÁVERO et al., 2009, p. 201), conforme a fórmula $d_{ii'} = \sqrt{\sum_j \left(Y_{ij} - Y_{i'j}\right)^2}$, considerando Y_{ij} a observação no i-ésimo indivíduo para a j-ésima característica. Após a construção da matriz de similaridade, ou dissimilaridade, existem vários métodos para o agrupamento hierárquico onde cada um formará um tipo diferente de agrupamento. Os mais comuns são: Ligação Simples, Ligação Completa, Centroides, Ligação Média e método Ward, cuja explicação pode ser vista na literatura (MINGOTI, 2005; JOHNSON &WICHERN, 1982).

Basicamente, os Métodos Hierárquicos dividem os indivíduos em grupos sendo este processo repetido até a formação do gráfico conhecido como Dendrograma.

No método da Ligação Simples, a similaridade entre dois conglomerados (grupos) é definida pelos dois elementos mais parecidos entre si (MINGOTI, 2005). Já o agrupamento por Ligação Completa é exatamente o oposto do Método da Ligação Simples. Nesse caso, os elementos são agrupados considerando a distância máxima (ou similaridade mínima). No método do Centroide (ou UPGMC), a distância entre dois grupos é definida como sendo a distância entre os vetores de médias, também chamados de centroides, dos grupos que estão sendo comparados. De acordo com Mingoti (2005), o método do centroide é direto e simples, porém, para fazer o agrupamento, é necessário em cada passo voltar-se aos dados originais para o cálculo da matriz de distâncias, exigindo um tempo computacional maior comparado com outros métodos. O método do centroide não pode ser usado em situações nas quais se dispõe apenas da matriz de distâncias entre os n elementos amostrais.

O método de agrupamento proposto por Ward (1963) é fundamentado na mudança de variação entre os grupos e dentro dos grupos que estão sendo formados em cada passo do agrupamento. Cada elemento é considerado como um único conglomerado, e em cada passo do algoritmo de agrupamento é calculado a soma de quadrados dentro de cada conglomerado No método de Ward, também conhecido como "Mínima Variância", a formação dos grupos se dá pela maximização da homogeneidade dentro dos grupos ou a minimização total da soma de quadrados dentro dos grupos (MINGOTI, 2005). Em cada passo, são formados grupos de modo que tenham a menor soma de quadrados, em que essa soma é o quadrado da distância Euclidiana de cada elemento amostral pertencente ao grupo em relação ao vetor de médias do grupo, ou seja, $SS_i = \sum_{i=1}^{ni}(X_{ij}-\bar{X}_i)'(X_{ij}-\bar{X}_i)$, em que ni é o número de elementos no conglomerado Ci quando se está no passo k do processo de agrupamento, Xij é o vetor de observações do j-ésimo elemento amostral que pertence ao i-ésimo conglomerado, $\bar{X}_i$ é o centroide do conglomerado C_i, e SS_i representa a soma de quadrados correspondente ao conglomerado C_i . No passo k, a soma de quadrados total dentro dos grupos é definida como $SSR = \sum_{i=1}^{gk} SS_i$, sendo SSR a Soma de Quadrados Residual, em que gk é o número de grupos existentes quando se está no passo k. A distância entre os grupos C1 e C2, é então definida como $d(C_l,C_i) = \frac{nl*ni}{nl+ni} (\bar{X}_l - \bar{X}_i)' (\bar{X}_l - \bar{X}_i)$ em que nl é o número de elementos do grupo C_l e ni é o número de elementos do grupo C_i . Em cada passo do algoritmo de agrupamento, os dois conglomerados que minimizam a distância são combinados.

A ligação média entre grupos, também conhecida por UPGMA (Unweighted Pair-Group Method using the Average approach), é um método não-ponderado de agrupamento aos pares, utilizando médias aritméticas das medidas de dissimilaridade, que evita caracterizar a dissimilaridade por valores extremos (máximo ou mínimo) (CRUZ; CARNEIRO, 2003). Este método trata a distância entre dois conglomerados como a média das distâncias entre todos os pares de elementos que podem ser formados com os elementos dos dois conglomerados que estão sendo comparados. Portanto, se um Grupo D_1 tem n elementos e outro grupo D_2 tem m elementos, a distância entre eles é dada por $d(D_1,D_2) = \sum_{l\in D2}\sum_{k\in D1}\frac{1}{mn} d(X_1,X_k)$.

No final do processo de agrupamento, após a aplicação de algum método hierárquico, é obtido um gráfico denominado dendrograma ligando os objetos, ou indivíduos, segundo seus níveis de similaridade. O eixo y desse gráfico corresponde a distância de ligação entre os objetos que estão dispostos no eixo y.

2.3.Correlação Cofenética (CCC)

A Correlação Cofenética mede o grau de ajuste entre a matriz de similaridade original (matriz S) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Segundo Bussab et al. (1990), esta correlação é calculada usando:

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(c_{ij}-\bar{c}\right)\left(s_{ij}-\bar{s}\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(c_{ij}-\bar{c}\right)^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(s_{ij}-\bar{s}\right)^2}}$$

em que c_{ij}: valor de similaridade entre os indivíduos $i\ e\ j$, obtidos a partir da matriz cofenética; s_{ij}: valor de similaridade entre os indivíduos $i\ e\ j$, obtidos a partir da matriz de similaridade; sendo $\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}$ e $\bar{s} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}$.

Nota-se que essa correlação equivale à correlação de Pearson entre a matriz de similaridade original e aquela obtida após a construção do dendrograma. Assim, quanto mais próxima de 1, menor será a distorção provocada no dendrograma, originado pelo agrupamento dos indivíduos com algum método hierárquico escolhido.

2.4. Critério para seleção do número de agrupamentos

Uma questão de grande importância é como se deve proceder na escolha do número final de grupos que define a partição do conjunto de dados analisado, ou em qual passo o algoritmo de agrupamento deve ser interrompido (MINGOTI, 2005).

O critério mais simples utilizado para decidir qual o número de grupos a adotar é o corte do dendrograma pela análise subjetiva dos diferentes níveis do mesmo, o que torna esse procedimento naturalmente enviesado pelas necessidades e opiniões dos analistas e pesquisadores (MARTINS; PEDRO e ROSA, 2004).

Timm (2007) destaca o resultado obtido por vários autores de que o índice pseudo F é o mais útil para identificar o número de clusters.

2.4.1. Estatística Pseudo F

O pseudo-F, proposto por Calinski e Harabasz (1974), é semelhante ao seu homônimo da econometria clássica e está baseado na variância entre os grupos, a cada nível de agregação. Assim, um valor alto do teste é desejável e implicaria na rejeição à hipótese de homogeneidade entre os grupos criados.

Calinski e Harabasz (1974) sugerem para cada passo do agrupamento, o cálculo da estatística chamada Pseudo F, definida por:

$$F = \frac{SSB/(g*-1)}{SSR/(n-g)} = \left(\frac{n-g*}{g*-1}\right)\left(\frac{R^2}{1-R^2}\right)$$

Em que SSB é a soma de quadrados total entre os g* grupos da partição, e SSR a soma de quadrados total dentro dos grupos da partição (Soma de Quadrados Residual), R^2 é o coeficiente da partição, dado por $R^2 = \frac{SSB}{SST_c}$, em que SST_C é a soma de quadrados total corrigida para a média global em cada variável e g* é o número de grupos relacionado com a partição do respectivo estágio de agrupamento e n é o número de elementos amostrais.

Segundo Calinski e Harabasz (1974), se F é monotonicamente crescente com g*, os dados sugerem que não existe qualquer estrutura natural de partição dos dados. Porém, se isso não ocorrer e a função F apresentar um valor de máximo, o número de

conglomerados e a partição referente a esse valor máximo corresponderão à partição ideal dos dados.

2.5. Construção do Banco de Dados

Todas as informações supracitadas foram obtidas diretamente do site Planilhas – Bit a Bit (http://radames.manosso.nom.br/bitabit/) que fornece planilhas com dados livremente disponíveis para todos utilizarem e redistribuírem como desejarem, sem restrição de licenças, patentes ou mecanismos de controle, além de cursos online de Excel e dicas do mundo digital.

Os dados desta pesquisa correspondem aos países da América do Sul e as covariáveis os próprios indicadores socioeconômicos, nesse caso PIB, PIB PPC, IDH e Índice de Gini. Para cada um dos indivíduos foram analisadas as covariáveis citadas a fim de organizá-los em grupos (clusters) em que elementos de um mesmo cluster sejam similares. Para definir tal semelhança é usada uma função de distância Euclidiana, que foi definida levando em conta o contexto do problema em questão. As análises preliminares dos dados foram realizadas no software Excel.

2.6. Variáveis Analisadas

Foram consideradas para o estudo 4 variáveis relacionados a cada um dos 12 países da América do Sul. Tais variáveis representam indicadores socioeconômicos, sendo eles PIB, PIB PPC, IDH e Índice de Gini, traduzindo assim possíveis futuras classificações dos países.

O Produto Interno Bruto (PIB) refere-se ao total dos bens e serviços valorados a preço de mercado e produzidos pelas unidades produtoras residentes (IBGE, 2018). O PIB equivale, portanto, à soma dos valores adicionados pelas diversas atividades econômicas acrescida dos impostos, líquidos de subsídios, sobre produtos, podendo ser expresso de três formas:

a) da produção – o produto interno bruto é igual ao valor bruto da produção, a preços básicos, menos o consumo intermediário, a preços de consumidor, mais os impostos, líquidos de subsídios, sobre produtos;

b) da despesa – o produto interno bruto é igual à despesa de consumo das famílias, mais o consumo do governo, mais o consumo das instituições sem fins de lucro a serviço das famílias (consumo final), mais a formação bruta de capital fixo, mais

a variação de estoques, mais as exportações de bens e serviços, menos as importações de bens e serviços;

c) da renda – o produto interno bruto é igual à remuneração dos empregados, mais o total dos impostos, líquidos de subsídios, sobre a produção e a importação, mais o rendimento misto bruto, mais o excedente operacional bruto.

O mais comum é que sejam considerados os valores do PIB per capita que é, basicamente, o valor do Produto Interno Bruto dividido pelo número de habitantes do país , ou seja, ele mede, dentro de um determinado espaço geográfico, o valor médio agregado por indivíduo dos bens e serviços finais produzidos neste local no ano considerado. Este é capaz de indicar a qualidade da produção econômica de um território em relação ao seu contingente populacional, sendo que valores muito baixos indicam segmentos sociais com precárias condições de vida.

Não é regra, entretanto, espera-se que países desenvolvidos possuam renda per capita maior do que países subdesenvolvidos. Um contraexemplo disso é a China (segunda maior economia do mundo), com renda per capita de 7,4 mil dólares em 2014, valor considerado baixo quando se fala em um país com tal economia, isso se deve ao tamanho da população chinesa (1,3 bilhões de pessoas), pois quando realizada a média de renda por pessoa o valor se tornará pouco expressivo.

Ainda diferencia-se o PIB em nominal e real, O PIB nominal refere-se ao valor do PIB calculado a preços correntes, ou seja, no ano em que o produto foi produzido e comercializado. Já o PIB real é calculado a preços constantes, onde é escolhido um ano-base, eliminando assim o efeito da inflação, e é de suma importância quando é desejada uma analise mais consistente na variação do PIB, já que não leva em conta as alterações nos preços de mercado dos bens produzidos e sim as variações nas quantidades produzidas.

A Paridade do Poder de Compra (PPC) é um parâmetro capaz de corrigir o cálculo do Produto Interno Bruto, relacionando a capacidade aquisitiva de uma economia com o custo de vida local, para assim representar o seu real poder de compra (REIS, 2018) . Muito utilizado em comparações internacionais, o PIB PPC remove as distorções causadas pelas diferentes taxas de câmbio, custo de vida e rendimentos da população, tornando-se mais assertivo quanto a produção total econômica do país. O fator de conversão de paridade de poder de compra (PPC) é o número de unidades da moeda de um país necessárias para comprar a mesma quantidade de bens e serviços no

mercado interno como uma quantidade de dólares compraria nos Estados Unidos (REIS, 2018).

O Índice de Desenvolvimento Humano (IDH) é um indicador utilizado para medir e comparar padrões de vida de diferentes populações. É uma maneira padronizada de avaliação e medida do bem-estar de uma população (SIMONE, EDINILSA E MARIA CECILIA, 2007). O IDH é composto por três dimensões: longevidade (esperança de vida ao nascer), educação (taxa de analfabetismo e número de anos de estudo) e renda (renda familiar per capita). O índice varia de 0 a 1, dado que quanto mais próximo de 1 melhor é o padrão de vida da população em estudo. De acordo com o Pnud o IDH é classificado da seguinte forma:

IDH entre 0 e 0,5 – Baixo Desenvolvimento Humano;

IDH entre 0,5 e 0,8 – Médio Desenvolvimento Humano;

IDH entre 0,8 e 1 – Alto Desenvolvimento Humano.

O Índice (ou Coeficiente) de Gini é uma medida numérica que representa o afastamento de uma dada distribuição de renda (Curva de Lorenz) da linha de perfeita igualdade, variando de "0" (situação onde não há desigualdade) a "1" (desigualdade máxima) (IBGE, 2018).

Todos os valores considerados no banco de dados são referentes as coletas realizadas no ano de 2017, ou seja, as variáveis demostram a classificação dos países neste ano em questão podendo haver variações nos anos anteriores ou subsequentes.

As variáveis consideradas no modelo são quantitativas contínuas.

3. Resultados

3.1.Caracterização Geral dos Países da América do Sul

Previamente, a fim de se obter informações e entender melhor o comportamento das varáveis em relação ao contexto do estudo, foram realizadas algumas análises descritivas da soma de todos os países considerados no trabalho (Tabela 1).

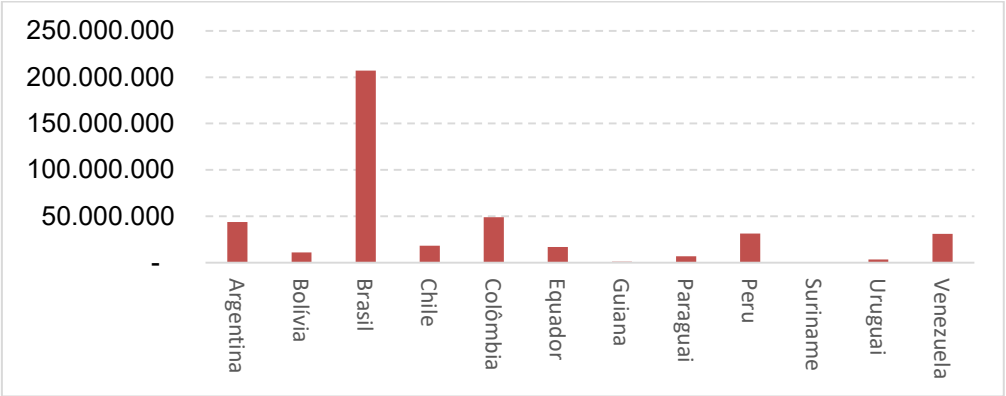
Tabela 1. Estatísticas descritivas referente aos países da América do Sul.

	População	***Área em Km2***	***Densidade Populacional***	***PIB PPC***
Média	35.043.997,50	1.475.738,25	23,78	553.167,17

Mediana	17.445.150,00	836.697,50	21,89	308.593,50
Desvio padrão	56.639.487,20	2.335.850,92	17,31	890.268,64
C.V. (%)	161.62	158.28	72.79	160.94
Intervalo	206.683.362	8.352.497	61,82	3.211.513
Mínimo	541.638	163.270	3,32	6.477
Máximo	207.225.000	8.515.767	65,13	3.217.990

	PIB nominal	***PIB nominal per capita***	***PIB PPC per capita***	***GINI***	***IDH***
Média	324.527,08	9.044,86	14.836,69	47,92	0,74
Mediana	151.363,00	8.121,18	14.736,05	47,60	0,73
Desvio padrão	495.369,16	4.840,25	5.680,86	4,72	0,06
C.V. (%)	152.64	53.51	38.29	9.85	8.11
Intervalo	1.769.505	13.509,89	18.182,95	16,00	0,20
Mínimo	3.086	3.003,72	6.792,12	41,60	0,64
Máximo	1.772.591	16.513,60	24.975,07	57,60	0,84

A média populacional dos países da América do Sul é de cerca de 35 milhões de habitantes, sendo que a população do Brasil é a única que se destaca entre os países estudados, pois sua população é muito maior do que a de todos os outros países do continente, conforme observado na Figura 1. Em contrapartida, a Guiana e o Suriname são os países que se destacam por ter a população muito pequena, sendo, inclusive, menor que a de muitas cidades brasileiras.



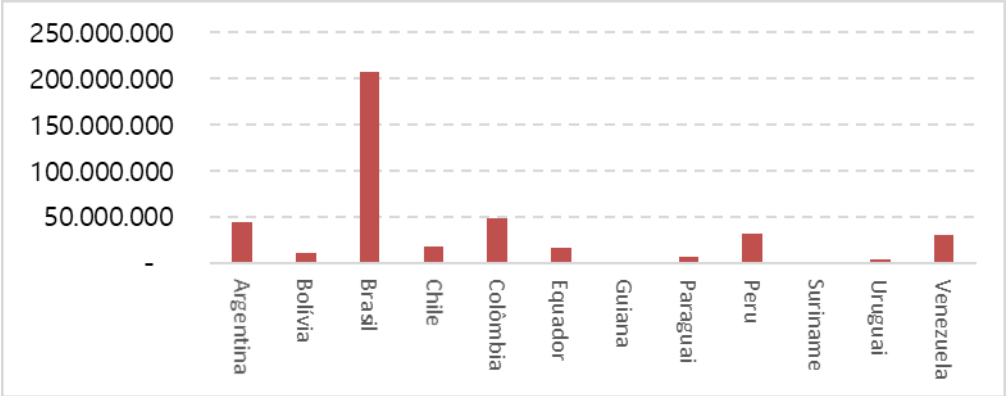


Figura 1. Gráfico de barras da população dos países da América do Sul considerados no estudo.

É importante mencionar que o tamanho da população é uma variável importante quando estuda-se a renda per capita, pois um país pode ser bastante desenvolvido socioeconomicamente e ter uma população expressiva, mas ainda assim ter renda per capita baixa. A China é um ótimo exemplo disso, apesar de se tratar de um país com uma das maiores economias do mundo, não apresenta uma renda per capita tão alta quando comparado ao seu PIB, isso se deve ao gigantesco número de habitantes.

Em relação ao PIB PPC, percebe-se pela Figura 2 que os países mais populosos são também os que possuem "maior riqueza". Novamente, o Brasil é o país que mais se destaca entre os países da América do Sul, com um PIB PCC bem maior que os demais. A Guiana e o Suriname figuram entre os países mais "pobres" do continente.

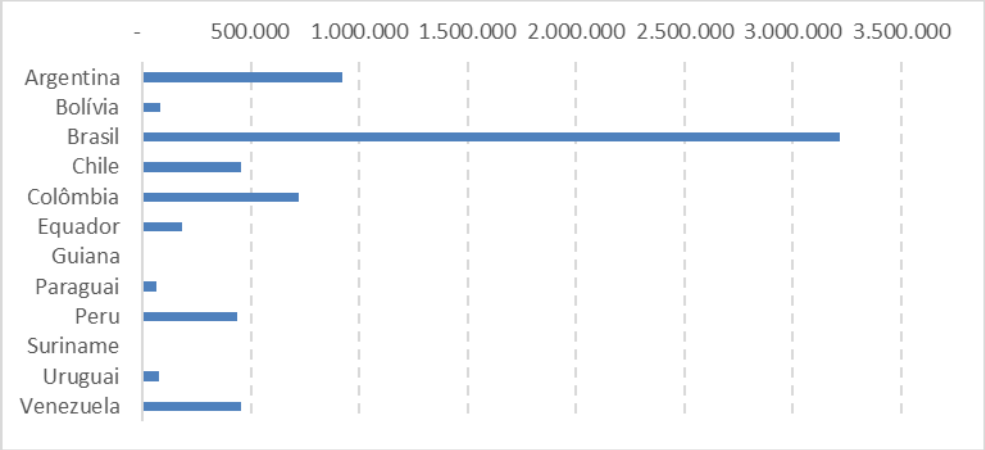


Figura 2. Gráfico de barras do Produto Interno Bruto PPC (PIB PPC) dos países da América do Sul considerados no estudo.

O IDH mede não só o desenvolvimento econômico da população de uma determinada região mas também a sua qualidade de vida, pela Figura 3 percebe-se que em relação a essas duas características ficam em evidência, frente aos demais países do estudo, a Argentina e o Chile, apresentando IDH em torno de 0.8, destacam-se de forma inversa a Bolivia, a Guiana e o Paraguai com IDH próximo de 0.6.

Figura 3. Gráfico de radar do Índice de Desenolvimento Humano (IDH) dos países da América do Sul considerados no estudo.

Na Figura 4 é possível observar que em relação ao Índice de Gini não existe uma diferença marcante entre os países, todos possuem valores entre 40 e 60. O Suriname apresenta um sutil destaque, o que nos sugere que, entre os países estudados, ele é o que possui distribuição de riquezas menos igualitária.

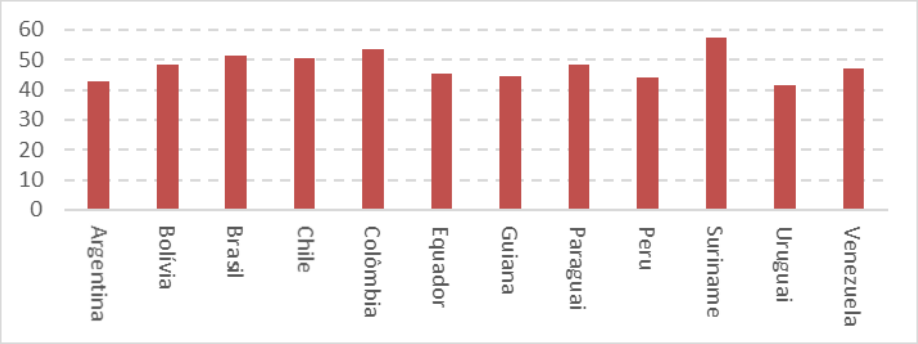


Figura 4. Gráfico de barras do Índice de Gini dos países da América do Sul considerados no estudo.

3.2. **Análise de Agrupamento**

Foi realizada no software Past uma análise de agrupamentos, em que a primeira decisão deu-se em relação ao método hierárquico que melhor se adequa aos dados, considerando que quanto maior o Coeficiente de correlação cofenético (CCC), melhor o método utilizado. Foram obtidos os resultados a seguir:

Tabela 2. Coeficiente de correlação cofenético obtidos.

Ligação Simples	0,9760
Ligação Média	0,9819
Método de Ward	0,9565

De acordo com os resultados, o método mais indicado é o da Ligação Média com CCC igual a 0,9819, que corresponde a aproximadamente 98,2% de consistência no agrupamento obtido pelo dendrograma (Figura 1). É importante ressaltar que para todos os métodos aplicados foram obtidos bons valor para o CCC.

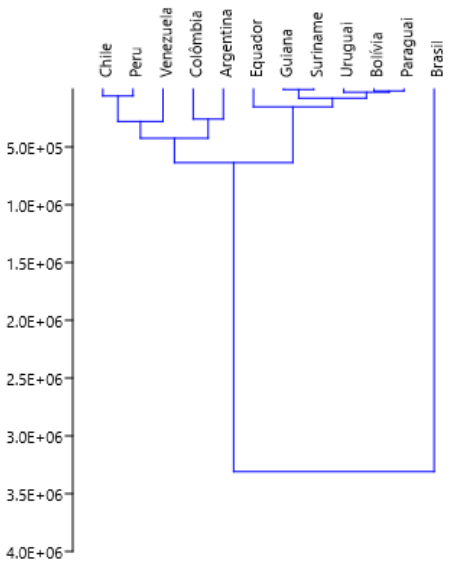


Figura 5. Dendrograma gerado com distância Euclidiana e método da Ligação Média.

Pelo dendrograma, observa-se a formação de três grupos distintos, sendo um deles um grupo unitário formado pelo Brasil. A fim de confirmar o ponto de corte, foi realizada uma análise estatística levando em consideração o critério Pseudo F (com auxílio do software R) que indicou de fato a divisão de três grupos. Foi então realizado um ponto de corte no gráfico onde seriam considerados três grupos, o mesmo pode ser melhor vizualizado na Figura 6.

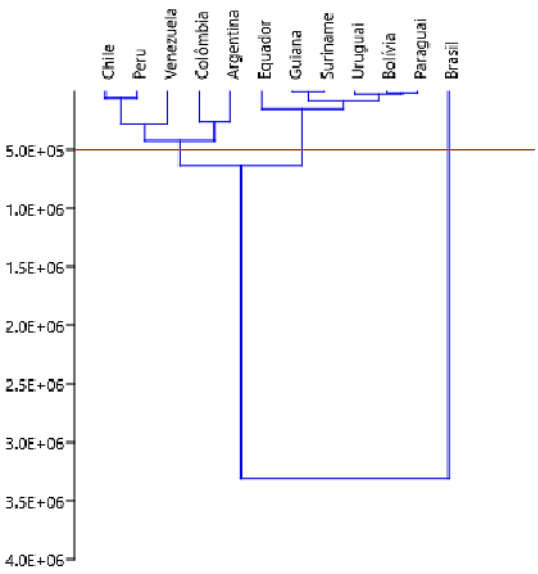


Figura 6. Dendrograma da Figura 5 acrescido do ponto de corte.

Como visto na Figura 6, obteve-se pelo dendograma os seguintes grupos:

Tabela 3. Grupos formados na análise de agrupamentos.

Grupo 1	Grupo 2	Grupo 3
Chile	Equador	Brasil
Peru	Guiana	
Venezuela	Suriname	
Colômbia	Uruguai	
Argentina	Bolivia	
	Paraguai	

Apesar de importantes, comparações tendo como base proximidades geográficas podem ser traiçoeiras, principalmente tratando-se da América do Sul, um grupo nada homogêneo, sendo assim devem ser feitas com bastante cuidado para que não gerem distorções. Em relação a AA, para um resultado mais nítido uma boa tática é a análise individual de cada um dos grupos.

Perante todos os resultados obtidos nas análises feitas nesse estudo, desde as estatísticas descritivas, não é surpresa que a o Brasil tenha sido agrupado em um grupo unitário. O país de fato se sobressai em vários aspectos quando comparado aos demais países sul americanos, começando pela sua população que chega a ser milhares de vezes maior do que outras localidades, como a Guiana, por exemplo.

Se analisado de forma superficial, pode-se chegar a conclusões errôneas, como, por exemplo, baseado no IDH, e sabendo que o Brasil é o mais bem classificado no continente nesse aspecto (ocupou o 75° lugar no Relatório de Desenvolvimento Humano de 2015), afirmar que o mesmo foi isolado em um grupo sozinho por se tratar "do melhor lugar para se morar no continente". Deve ser levado em conta que o IDH não abrange todos os aspectos de desenvolvimento humano e portanto uma conclusão como esta não deve ser considerada.

De fato o país se destaca positivamente em vários aspectos, além do que já foi dito, poderíamos citar o seu PIB PPC com valor bem superior aos demais países, o que é facilmente explicado por se tratar de um país com uma das maiores riquezas naturais do mundo. Em contrapartida, o Índice de Gini sinaliza que o país é tão desigual em termos de distribuição de renda, quanto comparado aos seus vizinhos de continente.

No Grupo 2 foram aglomerados os países mais pobres do continente, concentrando-se nele as localidades com menor número de habitantes, os menores valores de PIB PPC e IDH, reafirmando a pobreza e fraca economia, os elevados Índices de Gini, indicador de uma população com distribuição de renda desigual.

O Grupo 1 aparentemente inclui as localidades que não possuem valores dos indicadores socioeconômicos altos o suficiente para serem equiparados ao Brasil mas que também não estão entre os mais pobres do continente.

3.3.Análise das Componentes Principais

A Tabela 4 fornece os autovalores, a proporção da variação explicada por cada Componente Principal (CP), bem como a porcentagem acumulada da variabilidade, valor este obtido pela adição das proporções sucessivas da variação explicada para obtenção do valor total de 100%.

Tabela 4. Autovalores e seus respectivos percentuais de variação para cada Dimensão (Componente).

	Autovalor	**% da variância**	**% acumulada**
CP1	2,20485	55,121	55,121
CP2	1,17451	29,363	84,484
CP3	0,60919	15,230	99,714
CP4	0,01145	0,2862	100,00

Por exemplo, aproximadamente 55,12% da variação total é explicada pelo primeiro Autovalor. Pode-se notar que os 2 primeiros componentes principais possuem autovalor maior do que 1 e são capazes de explicar mais de 84% da variabilidade contida nos dados. Optando por reter estes 2 componentes principais na análise, tem-se uma redução de 4 variáveis originais para 2 variáveis latentes, perdendo menos de 16% da informação acerca da variabilidade dos dado s. De acordo com Rencher (2002), pelo menos 70% da variância total devem ser explicadas pelos primeiros e o segundo componentes principais.

Para a criação de cada variável latente, observamos os coeficientes gerados na análise de componentes principais, formando cada componente principal por uma combinação linear das variáveis originais. A partir destes coeficientes é possível compreender o sentido de cada componente extraída na análise (Tabela 5).

Tabela 5. Tabela dos coeficientes gerados na análise de componentes principais

	CP1	C2
PIB PPC	0.6540	0.0847
PIB	0.6580	0.0523
Índice de Gini	0.1461	0.7945
IDH	0.3435	-0.6000

A correlação entre determinada variável e o Componente Principal é utilizada como as coordenadas do indivíduo no CP. Sua representação difere do gráfico de observações: as observações são representadas por suas projeções, mas as variáveis são projetadas por suas correlações (JOHNSON & WICHERN, 1982).

Tabela 6. Coordenadas dos países analisados dentro dos Componentes Principais.

	PC1	**PC2**	**PC3**	**PC4**
Argentina	0,6510	-1,6636	0,4072	0,2914
Bolívia	-0,7706	0,6920	-0,7092	0,1322
Brasil	2,7717	0,7314	-1,0265	0,2914
Chile	0,2884	-0,4339	1,8655	0,4683
Colômbia	0,1836	1,0415	0,5337	0,3832

Equador	-0,4642	-0,3784	-0,2302	0,2673
Guiana	-0,9906	0,2544	-1,5812	-0,0401
Paraguai	-0,7079	0,4977	-0,4212	0,1244
Peru	-0,2751	-05480	-0,5990	0,8180
Suriname	-0,4553	1,6672	1,5493	-0,0175
Uruguai	-0,4116	-1,5130	0,1205	0,3719
Venezuela	0,1806	-0,3475	0,0909	-3,0904

No primeiro componente principal ficou evidente o contraste entre Brasil e Guiana, não por coincidência, logo que a CP1 pode ser chamada de PIB e esses dois países representam os valores extremos (maior e menor) desse indicador em relação aos países em estudo.

Já no segundo componente principal destacaram-se Suriname e Argentina, sendo que quando comparado aos demais países em estudo, a Argentina apresenta um dos maiores valores para o IDH e um dos menores para o Índice de Gini e o Suriname se mostra na direção oposta, ou seja, apresenta um dos menores valores para o IDH e um dos maiores para o Índice de Gini. Sendo assim, a CP2 poderia ser renomeada de forma a indicar uma combinação dos valores de IDH e Índice de Gini.

Por definição, a correlação entre os principais componentes é zero, isto é, a variação explicada em CP1 é independente da variação explicada em CP2 e assim por diante (SAVEGNAGO, 2015).

Uma outra maneira de se observar a relação entre os elementos de cada componente é através do gráfico Biplot (Figura 6).

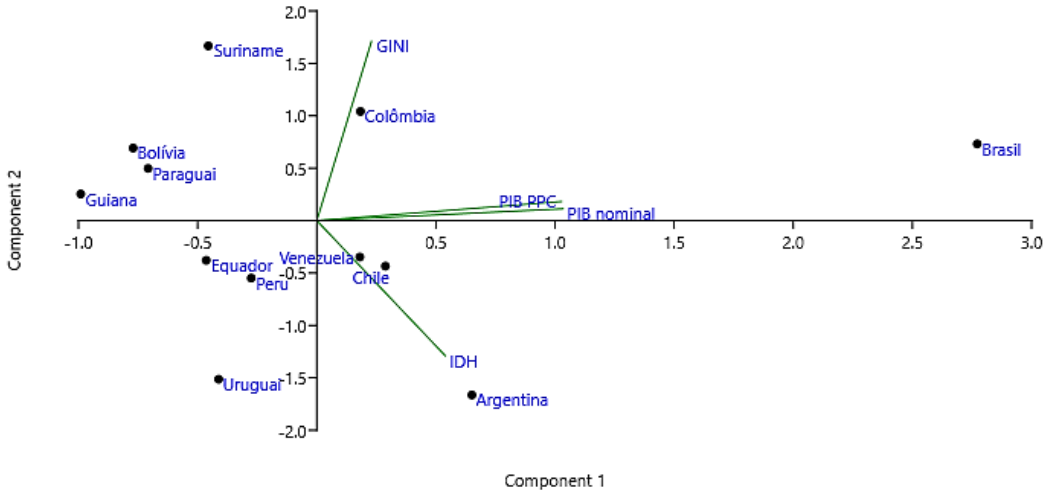


Figura 6. Gráfico Biplot referente a relação entre os países da América do Sul considerados no estudo e as componentes 1 e 2.

Pelo gráfico é observada a alta correlação entre PIB e PIB PPC, logo que se observa que os vetores de ambos quase se sobrepõem. Também é de fácil visualização, quando observado o grau formado pelo vetores do IDH e do Índice de Gini, a explicação dada anteriormente sobre a CP2. Vale ressaltar que o Brasil encontra-se significante isolado no gráfico, o que reafirma o seu destaque em várias características.

4. Conclusões

Comparações tendo como base proximidades geográficas podem ser traiçoeiras, principalmente tratando-se de localidades com pouca homogeneidade, como é o caso da América do Sul, deste modo devem ser feitas de forma criteriosa. Pensando nisso e de acordo com os resultados obtidos, a análise de agrupamentos se mostrou eficaz, entretanto a sua interpretação deve vir acompanhada de algumas observações.

Quanto ao Brasil, por exemplo, seus valores para a maioria dos indicadores se destacaram, sendo assim o seu agrupamento em um grupo unitário não trouxe grandes surpresas, todavia, dentre os países analisados, ele é o que apresenta maiores riquezas naturais além de possuir o maior número de habitantes, alguns outros países do estudo possuem população inferior a de cidades brasileiras. Logo, quando desejar-se uma análise mais aprofundada é interessante considerar todas essas variáveis, que provavelmente se mostrarão muito significativas no resultado final.

A formação dos outros dois grupos também não se mostrou diferente do esperado, concentrou-se no Grupo 1 aqueles países, em relação aos demais da pesquisa,

que apresentaram valores melhores para os indicadores analisados, mas que também não atingiram resultados que pudessem os equiparar ao grande destaque, no caso o Brasil, e no Grupo 2 restaram as localidades que possuem valores para os indicadores inferiores ao restante do grupo, ou seja, os países mais pobres do continente.

A análise de componentes principais se mostrou efetiva e permitiu a retirada ou descarte de duas variáveis, o que representa metade das consideradas inicialmente, que apresentaram baixa variabilidade ou foram redundantes por estarem correlacionadas com as de maior importância para dois componentes principais. Assim, um menor número de variáveis foram necessárias para explicar a variação total resultando em economia de tempo e de recursos em futuros trabalhos que utilizarão essa mesma base de dados, sem perda significativa de informação. Um dos objetivos da ACP, neste caso, foi atingido, pois um número relativamente pequeno de componentes foi extraído (CP1 e CP2) com a capacidade de explicar a maior variabilidade nos dados originais (84,48%).

5. Referências Bibliográficas

BOURNOT, Sophie; KOECHLIN, Francette; SCHREYER, Paul. 2008 benchmark PPPs - measurement and uses. Statistics brief: Organisation for economic co-operation and development. Washington, n. 17, p. 1-8, 2011.

DEATON, Angus; HESTON, Alan. Understanding PPPs and PPP-based national accounts. American economic journal: macroeconomics. Pittsburgh, v. 2, n. 4, p. 1-35, 2010.

JOHNSON, R.A.; WICHERN, D.W. Applied Multivariate Statistical Analysis. New Jersey-USA: Englewood Cliffs, 642p. 1998.

KUBRUSLY, Lucia Silva. Um procedimento para calcular índices a partir de uma base de dados multivariados. Pesqui. Oper. [online]. 2001, vol.21, n.1, pp.107-117.

SILVA, César R. L. da. Comparações internacionais e a paridade de poder de compra da moeda. Informações econômicas, São Paulo, v. 33, n.1, p. 35-37, 2003.

SOLIGO, V. Indicadores: conceito e complexidade do mensurar em estudos e fenômenos sociais. 2012. Disponível em: <http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/1926/3184> Acesso: 05 de maio de 2019.

SOUZA, A. M.; VICINI, L. Análise multivariada da teoria à prática. Santa Maria: Departamento de Estatística UFSM. 2005.

TOMAZ, F. S. C.; PETERNELLI, L. A.; MARTINS FILHO, S. Avaliação da eficiência do método de Ward para comparação de modelos logísticos. In: 19º Simpósio Nacional de Probabilidade e Estatística, 2010, São Pedro - SP. Anais do 19º SINAPE, 2010. Disponível em: < http://www.ime.unicamp.br/sinape/19sinape/node/255>. Acesso: 13 de maio de 2019.

JOHNSON, R. e WICHERN, D.: Aplied multivariate statistical analysis. Englewood Cliffs: Prentice-hall, 1982.

FLECK, M. P. A. e Bourdel, M. C.: Método de simulação e escolha de fatores na análise dos principais components. Rev. Saúde Pública vol. 32, 1988, São Paulo – SP.

ARTES, R.: Aspectos estatísticos da análise fatorial de escalas de avaliação. Revista de Psiquiatria Clínica, v. 25, n. 5, 1998.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE: Síntese de Indicadores Sociais: Uma análise das condições de vida da população brasileira. Estudos&Pesquisas, informação demográfica e socioeconômica, v. 39, 2018.

REIS, T.: Como funciona o PIB por Paridade de Poder de Compra (PPC)?. Suno Research. Disponível em: <a href="https://www.sunoresearch.com.br/artigos/paridade-de-poder-de-compra-ppc/"> Acesso: 25 de novembro de 2019.

FAVERO, L.P.; BELFIORE, P.; SILVA, F.L.; CAHN, B.L.: Análise de dados – modelagem multivariada para tomada de decisão. 8ª ed. Rio de Janeiro: Ed. Elsevier Ltda.. 646p, 2009.

SAVEGNAGO, R.P., CAETANO, S.L., RAMOS, S.B., NASCIMENTO, G.B., SCHMIDT, G.S., LEDUR, M.C. MUNARI, D.P. Estimates ofgenetic parameters, and cluster and principal components analysesof breeding values related to egg production traits in a White Leghorn population, Poultry Science, 90, p.2174-2188. 2011.

RENCHER, A.C. Methods of Multivariate Analysis.A JOHN WILEY & SONS, INC. PUBLICATION.p.727. 2ed. 2002.

HONGYU, K., SANDANIELO, V. L. M., JUNIOR, G. J. O.: Análise de Componentes Principais: resumo teórico, aplicação e interpretação Scientific Journal of FAET and ICET UFMT, Hongyu, et al, E&S - Engineering and Science, 2015.

ASSIS, SG., SOUZA, ER., and MINAYO, MCS. Caracterização dos municípios: desenvolvimento social, econômico e contexto demográfico. In: MINAYO, MCS., and DESLANDES, SF., orgs. Análise diagnóstica da política nacional de saúde para redução de acidentes e violências [online]. Rio de Janeiro: Editora FIOCRUZ, 2007.

MINGOTI, Sueli Aparecida. Análise de Dados Através de Métodos de Estatística Multivariada: Uma abordagem Aplicada. Belo Horizonte: Editora UFMG, 2005.

MARTINS, M. do R. F. de O., SOFIA P., SOFIA, R. Escolha do número de grupos e validação da solução em análise classificatória: da teoria à prática. Disponível em: <https://run.unl.pt/handle/10362/7686>. Acesso em: 17 set. 2015.

TIMM, N. H. Applied multivariate analysis. Ney York: Springer, 2002. 695 p.

CALINSKI, T.; HARABASZ, J. A Dendrite Method for Cluster Analysis. Communications in Statistics, v. 3, n. 1, p. 1-27, 1974.