**“Análise fatorial e PCA”: Desenvolvimento social e econômico estadual brasileiro utilizando o fator étnico-racial**

Vinicius Faria Baessi dos Santos;Luiz Fernando Ribeiro

**“Análise fatorial e PCA”: Desenvolvimento social e econômico estadual brasileiro utilizando o fator étnico-racial (TCC)**

**Resumo**

O **desenvolvimento** **social** e **econômico** é algo abordado de forma constante pois interfere diretamente na vida das pessoas, como indivíduos presentes em uma sociedade. O fator **étnico-racial** infere grande peso, no que tange à estados brasileiros, cuja a carga histórica do país carrega uma pluralidade cultural e racial, sendo necessário considerá-la durante o entendimento e aplicação de projetos políticos. Neste cenário, o estudo fez uma análise de **indicadores** de evolução em estados brasileiros, onde foram separados por fatores e o objetivo entornaria no entendimento da variação e **comportamento** desses dados em conjunto, para realização de inferências na realidade atual do Brasil de forma parcial. A modalidade do estudo foi quali-quantitativa e os dados foram coletados através do portal do IBGE, que faz o levantamento através de pesquisas sazonais em todo o país. Como resultado, foi pontuada as **correlações** positivas, negativas e suas **variações** entre as observações impostas. Houve uma **exploração** concisa sobre as áreas de foco para aplicação de projetos, em que trariam impactos reais no âmbito de desenvolvimento humano, com uma visibilidade igualitária no meio social e racial.

**Palavras-chave:** Desenvolvimento; Social; Econômico; Indicadores; Comportamento; Variações; Correlações; Exploração.

**Introdução**

A análise de desenvolvimento econômico e social de determinada região brasileira é de extrema importância e uma das principais pautas de noticiários locais, devido a sua relevância como um todo. Podendo ser avaliado por meio de diversas perspectivas, dispondo de parâmetros para aspectos essenciais, presentes na nossa sociedade (Bezerra, 2019).

Há meios variados para capturar esses fenômenos, com base em fatores específicos, existentes em cada região ou em uma escala temporal. Sendo alguns deles o Produto Interno Bruto [PIB], que avalia a evolução econômica e em contrapartida, o Índice de Desenvolvimento Humano [IDH], combinado com os levantamentos de felicidade, onde trazem características o qual a progressão econômica é impossibilitada de refletir (Bezerra, 2019).

O primeiro cálculo do PIB foi efetivamente realizado em 1953 e durante um longo período, foi utilizado como medida comum de bem-estar humano, embasado em Produto Nacional Bruto [PNB] per capita e PIB per capita. Ambos foram criticados pois representam a valorização de apenas transações monetárias, fazendo alusão de que recursos naturais são ilimitados e livres, descartando o quesito de liberdade e direitos humanos, tais com a distribuição de renda e outros motivadores (Stanton, 2007)

No início da década de 1990, a Organização das Nações Unidas [ONU] lançou o IDH, que propôs validar as melhorias de um país, utilizando-se de indicadores de desempenho, tornando-o conhecido em sua época, como cálculo de desenvolvimento humano (Torres et al.; 2003).

Definido pelo Programa das Nações Unidas para Desenvolvimento [PNUD], o IDH engloba três componentes imprescindíveis para mensurar o progresso humano, tais como, longevidade, educação e renda. Observa-se que esses pilares são fortemente presentes até os dias atuais, devendo-se levar em conta, as questões culturais, sociais e históricas de cada localidade, evitando ponderações arbitrárias ao se basear em outras regiões. Avalia-se que este seja um guia para os cidadãos, organismos públicos e privados para atribuição de investimentos entre nações e políticas públicas. Compete a ele, portanto, considerar esforços que envolvem a busca de resultados positivos, embasando-se em bons indicadores que focam principalmente na distribuição de renda, gastos com educação e saúde, referindo a setores básicos e importantes nacionalmente, aumentando a eficiência de gastos públicos e impactando diretamente no cálculo, alterando a vista da nação aos olhares do mundo inteiro. Não se limitando somente à atribuição de uma conjuntura econômica-social (Belchior, 2011).

Na condução da história, a discussão referente às especificidades da economia, foram direcionadas para a qualidade de vida e as condições fundamentais para a população, colocando-se contra às antigas ponderações voltadas para o globo financeiro do indivíduo, que se sobressaia diante dos aspectos sociais inerentes a ela. Com isso o IDH obteve um peso indeclinável para a definição e conceito de desenvolvimento, apontado à uma região (PNUD; IPEA; FJP, 2013). Voltada à igualdade socioeconômica e racial, é importante frisar o impacto positivo que as políticas nacionais e internacionais trazem para esses indicadores, porém ainda assim, é notório que existe um longo caminho pela frente no âmbito brasileiro (Belchior, 2011).

Construir pontes para aproximação de pessoas com diferentes raças, dentro de uma realidade é um imenso desafio de engenharia social e econômica. Foram aplicadas políticas públicas em diferentes níveis do governo ao longo das últimas décadas, impulsionando uma base para tal feito, onde indicadores dissemelhantes apresentam avanços na condição da vida populacional, bem como acesso à renda, educação, direitos e serviços. Por meio de análises é visto que os componentes presentes no IDH, avançam de forma mais abrupta para negros, comparando-a com a população branca em território brasileiro, destacando uma assimetria com base no fator racial. É verossímil que exista um abismo de desigualdade étnica-racial e é algo que precisa ser tratado com políticas afirmativas de forma consistente (Charão, 2011).

O presente trabalho propõe uma análise de indicadores subjacentes de desempenho presentes no Brasil, por meio econômico, social e racial, visando avaliá-los através de fatores ortogonais entre si. Ao ressaltar as áreas deficitárias, bem como setores com potencial de crescimento positivo, nota-se um norte para a igualdade no país.

**Implementação de Algoritmo(s) de Machine Learning**

O processo metodológico para a aplicação exploratória será dividido em 10 partes, sendo elas: i) Pesquisa sobre desenvolvimento humano e etnia; ii) Coleta de dados para aplicação do estudo; iii) Escolha do software RStudio para aplicação da linguagem R para estudo; iv) Data Wrangling viabilizando a aplicação da análise fatorial e PCA; v) Visualização do comportamento conjunto das variáveis entre si; vi) Teste de esfericidade de Bartlett, para validar hipótese nula; vii) Extração dos autovalores e autovetores; viii) Extração dos scores fatoriais e comunalidades; ix) Definição de fatores usados para análise; x) Insights e validações de acordo com a pesquisa. O diagrama de processos abaixo, descreve os passos.

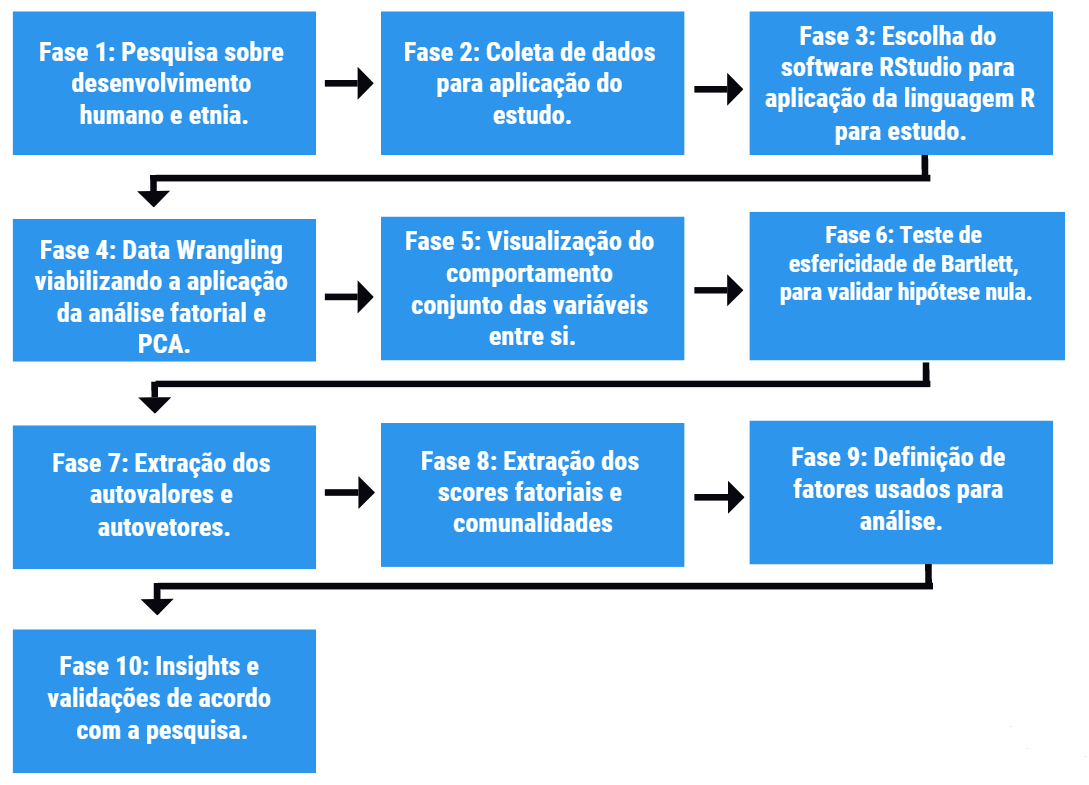


Figura1. Diagrama de processos – etapas para aplicação do tema ao trabalho.

Fonte: Dados originais da pesquisa

Serão utilizadas informações e orientações presentes no artigo “Análise Fatorial” para estruturação lógica durante a implementação (Enap, 2019). Utiliza-se a metodologia qualitativa e quantitativa, sendo um misto das duas e a base de dados aplicada em todo o processo de estudo do caso, foi divulgada de forma pública, como método de transparência através do portal do Instituto Brasileiro de Geografia e Estatística [IBGE], tratando-se de populações com diferentes aspectos dentro do contexto de desenvolvimento humano. Totalizadas 27 observações (linhas da tabela) e 10 variáveis (colunas da tabela), com início em 2010 e término em 2021, sobre indicadores, inseridos dentro dos três componentes principais para análise de desempenho em relação a todos os estados brasileiros.

(Continua)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estados | População estimada - 2021 | Densidade demográfica hab/ km² - 2010 | Matrículas no ensino fundamental – 2021 | Receitas realizadas – 2017 |
| Acre | 906876 | 4.47 | 153015 | 663.288.310.836 |
| Alagoas | 3365351 | 112.33 | 458782 | 1.195.043.846.015 |
| Amapá | 877613 | 4.69 | 133839 | 539.641.714.471 |
| Amazonas | 4269995 | 2.23 | 702763 | 1.732.845.943.402 |
| Bahia | 14985284 | 24.82 | 1946957 | 5.019.100.324.052 |
| Ceará | 9240580 | 56.76 | 1161434 | 2.842.022.247.191 |
| Distrito Federal | 3094325 | 444.66 | 369128 | 2.381.221.127.074 |
| Espírito Santo | 4108508 | 76.25 | 503003 | 1.968.561.674.376 |
| Goiás | 7206589 | 17.65 | 855021 | 3.788.533.516.848 |
| Maranhão | 7153262 | 19.81 | 1112636 | 1.850.326.135.491 |
| Mato Grosso | 3567234 | 3.36 | 486568 | 2.395.852.883.588 |
| Mato Grosso do Sul | 2839188 | 6.86 | 391975 | 1.639.665.577.012 |
| Minas Gerais | 21411923 | 33.41 | 2407107 | 9.719.982.315.616 |
| Pará | 8777124 | 6.07 | 1389983 | 2.584.944.610.454 |
| Paraíba | 4059905 | 66.7 | 540919 | 1.309.700.531.928 |
| Paraná | 11597484 | 52.4 | 1348296 | 6.016.357.612.246 |
| Pernambuco | 9674793 | 89.62 | 1249850 | 3.574.602.897.102 |
| Piauí | 3289290 | 12.4 | 459871 | 1.212.421.561.511 |
| Rio de Janeiro | 17463349 | 365.23 | 1945408 | 7.848.814.078.862 |
| Rio Grande do Norte | 3560903 | 59.99 | 447692 | 1.352.755.273.159 |
| Rio Grande do Sul | 11466630 | 37.96 | 1257992 | 6.639.746.817.915 |
| Rondônia | 1815278 | 6.58 | 244815 | 912.231.072.305 |
| Roraima | 652713 | 2.01 | 103123 | 426.631.854.785 |
| Santa Catarina | 7338473 | 65.27 | 900240 | 3.469.677.282.078 |
| São Paulo | 46649132 | 166.23 | 5396803 | 23.282.249.656.706 |
| Sergipe | 2338474 | 94.36 | 320638 | 1.014.504.695.355 |
| Tocantins | 1607363 | 4.98 | 227743 | 1.030.509.901.288 |

(Conclusão)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Despesas empenhadas – 2017 | Rendimento domiciliar mensal – 2021 | Branca – 2021 | Preta – 2021 | Parda – 2021 |
| 60.844.168.063 | 888 | 133 | 42 | 684 |
| 1.046.063.491.711 | 777 | 801 | 182 | 2320 |
| 422.446.408.829 | 855 | 136 | 44 | 659 |
| 1.532.489.655.705 | 800 | 552 | 92 | 3273 |
| 455.701.600.038 | 843 | 2626 | 3398 | 8716 |
| 2.460.835.218.276 | 881 | 2423 | 500 | 6160 |
| 219.904.646.849 | 2.513 | 1208 | 308 | 1496 |
| 1.439.233.800.188 | 1.295 | 1476 | 433 | 2104 |
| 2.424.838.034.233 | 1.276 | 2360 | 554 | 4117 |
| 1.762.717.075.574 | 635 | 1177 | 849 | 4971 |
| 1.818.736.327.009 | 1.362 | 987 | 329 | 2115 |
| 1.450.691.537.404 | 1.471 | 1132 | 182 | 1356 |
| 9.839.166.916.323 | 1.325 | 8275 | 2366 | 10533 |
| 2.253.347.004.547 | 847 | 1467 | 675 | 6412 |
| 1.007.470.004.266 | 876 | 1244 | 266 | 2471 |
| 5.553.440.297.483 | 1.541 | 7401 | 432 | 3499 |
| 3.332.048.644.445 | 829 | 2847 | 639 | 5957 |
| 967.673.631.835 | 837 | 585 | 313 | 2361 |
| 6.796.554.869.757 | 1.724 | 7790 | 2366 | 7101 |
| 1.133.095.755.333 | 1.109 | 1282 | 292 | 1944 |
| 6.247.627.934.364 | 1.787 | 9055 | 736 | 1557 |
| 70.855.300.168 | 1.023 | 509 | 121 | 1129 |
| 386.509.427.139 | 1.046 | 98 | 41 | 373 |
| 2.559.510.337.918 | 1.718 | 5826 | 210 | 1135 |
| 23.198.224.369.176 | 1.836 | 26683 | 3762 | 14966 |
| 849.492.719.996 | 929 | 442 | 204 | 1649 |
| 892.945.643.836 | 1.028 | 303 | 182 | 1065 |

Fonte: (Instituto Brasileiro de Geografia e Estatística, 2022).

Tabela 1 Base de dados aplicada ao estudo

Por questões estatísticas, as variáveis selecionadas serão quantitativas, exceto a primeira variável qualitativa nominal presente na tabela para identificar as demais. Vale ressaltar que o ano contido nas variáveis, referem-se ao último censo divulgado para tal informação. Sendo elas: Estados (atribuição de cada distrito); População estimada - 2021 (referindo a quantidade estimada de pessoas em determinada região); Densidade demográfica hab/ km² - 2010 (número de habitantes por quilômetro quadrado); Matrículas no ensino fundamental – 2021 (representando o quantitativo de crianças matriculadas neste nível de ensino); Receitas realizadas – 2017 (Arrecadação do valor disponível nos cofres públicos, disponíveis para uso do estado); Despesas empenhadas – 2017 (Refere-se ao valor reservado e comprometido para gastos e atribuições de políticas públicas); Rendimento domiciliar mensal – 2021, ( Indica a média de renda mensal por pessoa dentro de uma residência; Branca – 2021 (aborda a quantidade estimada de pessoas denominadas pela raça branca); Preta – 2021 (aborda a quantidade estimada de pessoas denominadas pela raça preta); Parda – 2021 (aborda a quantidade pessoas de pessoas denominadas pela raça parda). Todos os dados são divulgados de forma conjunta pelo portal (Ibge, 2022).

Será utilizado o algoritmo de Machine Learning não supervisionado para Análise Fatorial por Componentes Principais (Principal Components Analysis - PCA), esta categoria faz diagnósticos de interdependências existentes dentro da base de dados. Caracteriza-se como uma técnica exploratória, atribuída de novas variáveis métricas, que são chamadas de fatores e possuem coeficientes de correlação significativamente relevantes entre si, onde capturam o comportamento conjunto das variáveis originais. Portanto, entende- se que a (PCA), visa o agrupamento de variáveis através de critérios, previamente estabelecidos e em alguns casos, propõe a redução estrutural da dimensionalidade dos dados, facilitando e viabilizando a validação de constructos (Fávero; Belfiore, 2017). A visualização das correlações possibilita a notoriedade das variáveis que impactam diretamente no resultado final, trazendo insights para um direcionamento às áreas deficitárias para políticas já existentes de forma eficiente (Abud, 2019).

Para execução da análise fatorial por componentes principais, será utilizado o software RStudio, versão 4.0.5, que interpreta a linguagem R, possibilitando a execução do estudo e algoritmos que nele estão. A função prcomp(), será responsável por executar toda a análise fatorial usando as fórmulas descritas abaixo para obtenção dos fatores, porém serão usadas outras formas além, para ter insights; afinal, trata-se de uma técnica exploratória. O pacote plotly() para elaboração de gráficos, tidyverse() manipulação de dados, kableExtra() e knitr() para apresentação de tabelas, PerformanceAnalytics() para análise correlacional, factorextra() e psych() para auxílio na análise fatorial.

Será usada a função chart.Correlation() do pacote PerformanceAnalytics() para plotar um relatório gráfico completo sobre o comportamento e as correlações entre determinadas variáveis, afim de representar de forma breve a distribuição inicial dos fatores que serão gerados nos estágios finais da metodologia e ajudar a compreender como serão dados esses comportamentos, auxiliando na didática.

Para iniciar a análise, será extraída a matriz de correlações de Pearson da base de dados original. Abaixo será representada a aplicação da formula com suas devidas descrições, pressupondo que para cada observação referida por n, será dada a equação; i (i = 1, ... , n), o valor (k) refere-se ao quantitativo de variáveis métricas representadas por X.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Observação *i*** | **X 1*i*** | **X 2*i*** | **…** | **X *ki*** |
| **1** | **X 11** | **X 21** | **…** | **X *k*1** |
| **2** | **X 12** | **X 22** | **X *k*2** |
| **3** | **X 13** | **X 23** | **X *k*3** |
| **…** | **X 14** | **X 24** | **…** |
| ***n*** | **X 1*n*** | **X 2*n*** | **X *kn*** |

Figura 2. Modelo estrutural da base de dados para aplicação da análise fatorial.

Fonte: (Fávero; Belfiore, 2017).

Para extração de fatores partindo de k quantidade de variáveis X, será definida a matriz de correlações ρ que refletem os valores das correlações lineares de Pearson dentre cada par de variável conforme as figuras abaixo. A matriz ρ é simétrica no sentido diagonal principal com valores iguais a 1. Na formula, é dado o exemplo de uma base com duas variáveis, sendo elas X1 e X2, a média das variáveis X1 e X2.

Figura (3)

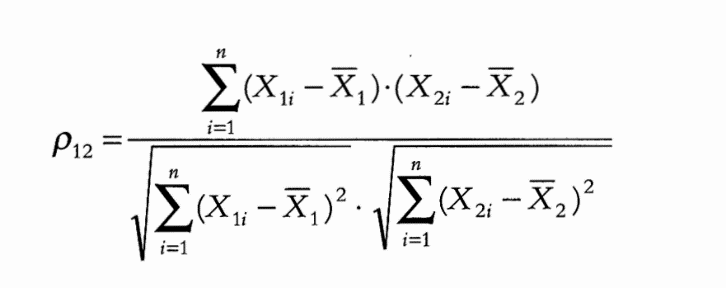


Figura 3. Expressão utilizada para o cálculo da correlação de Pearson.

Fonte: (Fávero; Belfiore, 2017).

Figura (4)

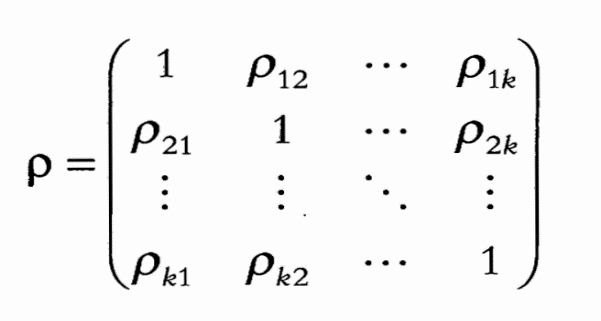


Figura 4. Matriz de correlações.

Fonte: (Fávero; Belfiore, 2017).

A correlação de Pearson pode variar entre -1 e 1, o valor estando próximo dessas extremidades, indica determinado grau de relação linear viabilizando a construção de um fator. Em contrapartida, o valor próximo de zero, pode indicar baixo grau de correlação. Em outras palavras, o atributo analisado, pode não impactar de forma significativa para a alteração de outras variáveis durante a análise, resultando na adequação em outros fatores de forma isolada. Os fatores de forma única, são entendidos como representações de dimensões latentes que evidenciam o comportamento das variáveis originais. A figura abaixo, ilustra o nível de correlação que é identificada a partir da matriz ρ de forma bidimensional, analisando os dados em pares. A figura cinco (5) nos mostra um nível de correlação positivo, ou seja; se o valor de determinada variável aumenta, uma outra semelhante sobe em conjunto. Todavia, a figura seis (6) indica uma correlação de Pearson próxima de 0, indicando que uma variável não impacta a outra. Entende-se que as escolhas dos fatores utilizados para a análise são de livre escolha pelo pesquisador, visando adequá-las para o estudo dentro da área qualificada. Para adequação global da análise fatorial será levada em consideração a hipótese nula. Em outras palavras, nossa matriz ρ é estatisticamente significante, dado o nível de significância de 95%.

Figura (5) Figura (6)

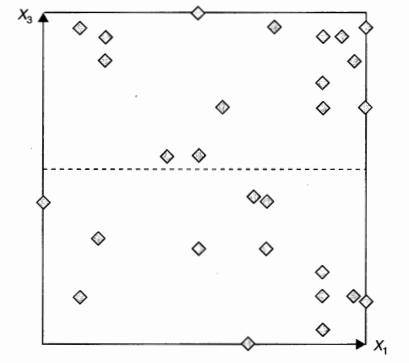
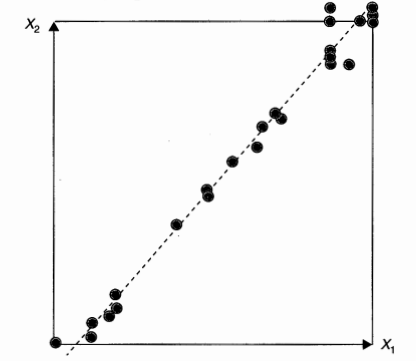


Figura 5 & 6. Projeção dos pontos em cada plano, formado por determinados pares de variáveis.

Fonte: (Fávero; Belfiore, 2017).

Um fator representa a combinação de forma linear das variáveis originais, seguindo este conceito, definimos variáveis K e uma subsequente quantia de K fatores. Para seguirmos com o estudo, precisamos extrair os autovalores de determinada matriz ρ que são obtidos a partir da solução do cálculo presente na figura abaixo e nos dará uma matriz K x K de dimensões (Fávero; Belfiore, 2017).

Figura (7) Figura (8)

Figura 7. Representa o cálculo para obter os autovalores.

Figura 8. Indica os autovalores propriamente ditos.

Fonte: (Fávero; Belfiore, 2017).

A formula será resolvida através de cálculo matricial onde será gerada uma nova matriz nomeada com matriz de autovalores conforme a figura abaixo.

Figura (9)

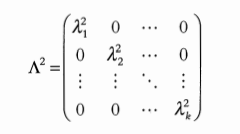


Figura 9. Matriz de autovalores.

Fonte: (Fávero; Belfiore, 2017).

Ao darmos continuidade com o processo, será necessária a extração dos autovetores que representam a parcela da variância total das variáveis de estudo e são advindas da matriz de autovalores.

Figura (10)

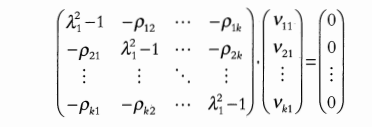


Figura 10. Expressão para extração do um (1) autovetor, partindo de um (1) autovalor.

Fonte: (Fávero; Belfiore, 2017).

As equações representam a extração de um único autovetor a partir do primeiro autovalor e o cálculo será reproduzido de acordo com autovalores, vindos da quantidade de variáveis para análise e de forma subsequente para a continuidade, seguindo a dimensão K x K.

Figura (11)

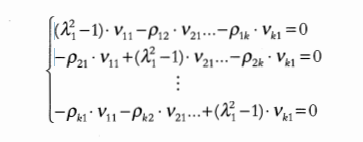


Figura 11. Expressão para extração do um (1) autovetor, partindo de um (1) autovalor.

Fonte: (Fávero; Belfiore, 2017)

Teremos uma matriz de autovetores que de forma sequente, nos dará os scores fatoriais através das figuras abaixo.

Figura (12)

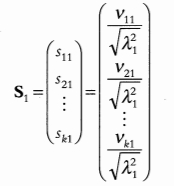


Figura 12. Expressão para extração do um (1) autovetor, partindo de um (1) autovalor.

Fonte: (Fávero; Belfiore, 2017).

O cálculo deve ser repetido para cada uma das variáveis, tendo a finalidade obter-se o score fatorial de cada fator. Onde por fim, avançaremos para a equação, que padronizados através do procedimento de Zscores, nos dará os fatores, propriamente dito com correlações de Pearson iguais a 0, tornando-os ortogonais entre si.

Figura (13)

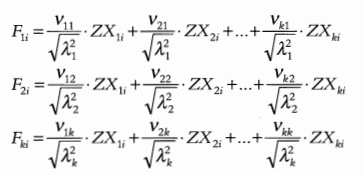


Figura 13. Extração dos fatores.

Fonte: (Fávero; Belfiore, 2017).

**Resultados e Discussão**

A base de dados atribuída ao estudo, foi extraída no portal do IBGE, onde passou por uma seleção e pré-processamento das variáveis essenciais à pesquisa, evitando a saturação de informações implicando em problemáticas durante a exploração, tais como: grandes quantidades de fatores irrelevantes, processamento computacional desnecessário e falta de lógica para dados dentro do contexto de desenvolvimento humano ou dados que introduzem uma abstração não condizendo com a realidade (Nunes, 2012). Considerando as particularidades e necessidades para o diagnóstico dos fatores, foi dada a exclusão da variável IDH, correspondente ao último censo de 2010, onde interfere de forma direta durante o estudo, sendo irrelevante na elaboração do trabalho, cujo a finalidade principal é descobrir e encontrar grupos e especificidades que apresentem um determinado grau de correlação positiva ou negativa, ao qual a variável traz de forma não assertiva dentro da pesquisa.

Com o objetivo de que os gestores ou decisores tenham uma visibilidade coerente sobre o comportamento conjunto dos dados, será dada uma matriz de correlações ilustrada, utilizando todas as variáveis presentes, a priori de forma visual, por meio de 2 visualizações distintas formadas por pares de variáveis. A primeira, é uma projeção de pontos, detalhando através de uma linha, a correlação linear positiva, nula ou negativa; A segunda é um dado numérico quantitativo sobre o grau de correlação, que por questões presentes no pacote atribuído ao software R, é mostrado um asterisco ou mais, nos informando sobre a significância estatística do dado e o tamanho da fonte é aumentado ou reduzido de acordo com o critério. Na figura 1 e 2, será trazida a parcela para uma dupla de variáveis com nível de correlação alta e baixa. A figura 3, mostrará a matriz completa e podermos visualizar o montante de todos os dados.

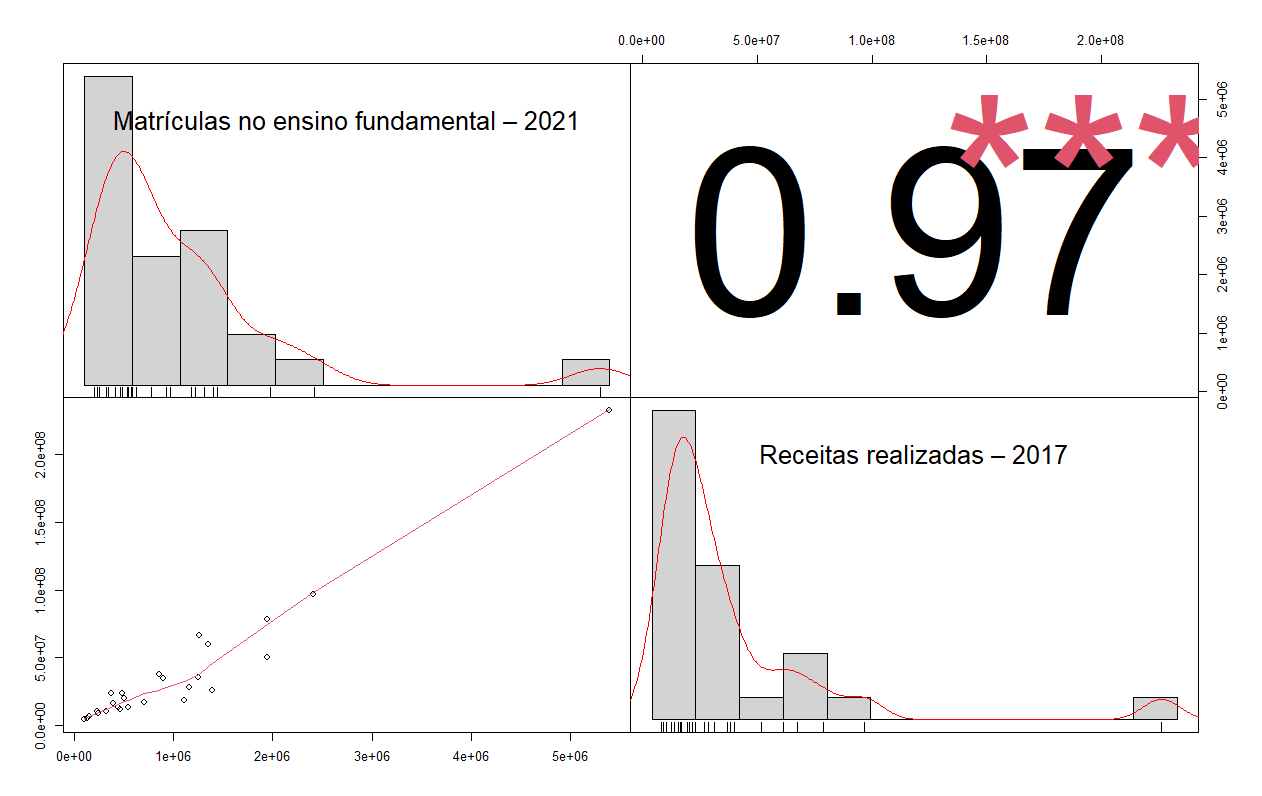


Figura 1. Correlação linear de Pearson alta.

Fonte: Dados originais da pesquisa.

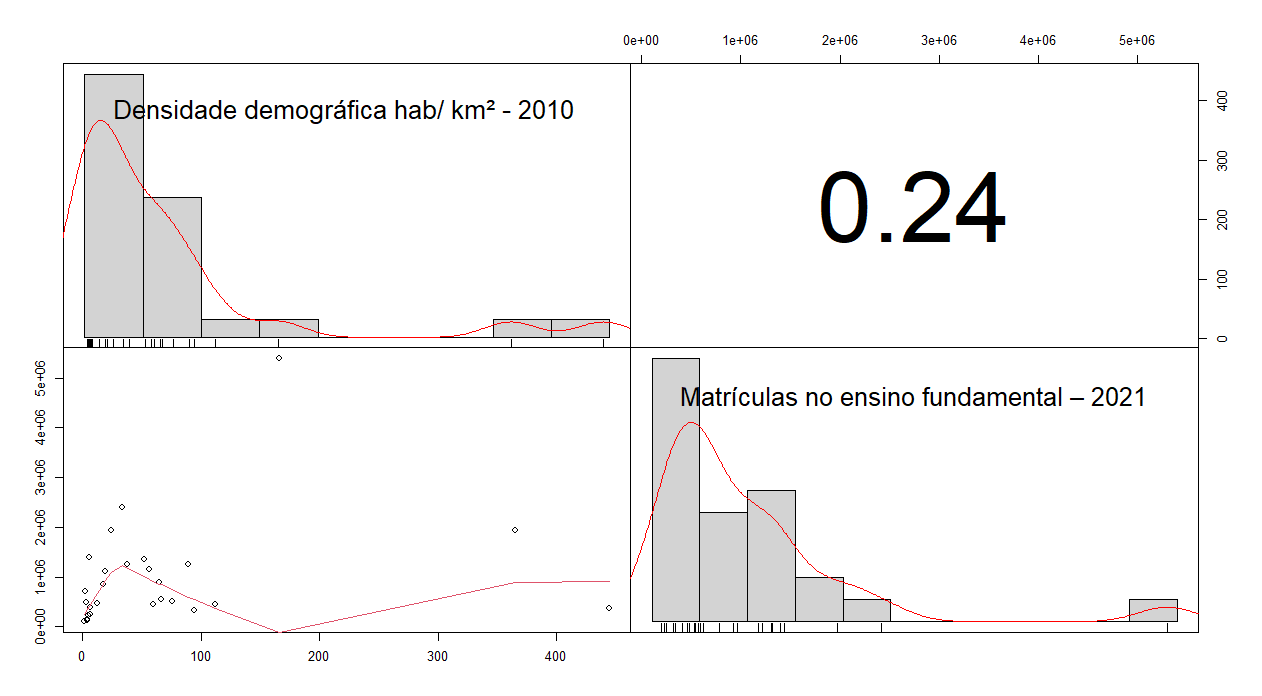


Figura 2. Correlação linear de Pearson baixa.

Fonte: Dados originais da pesquisa.

Observou-se na figura 3, que grande parcela das variáveis, apresentam um grau de significância estatística em relação a outros dados. Todavia, há uma face que mostra dados com correlações extremamente baixas ou até negativas. Embasando-se no conceito que podem variar de -1 a 1, nota-se uma amostragem, onde algumas variáveis tem a capacidade de inferir diretamente em semelhantes, e ao mesmo tempo podem ser irrelevantes para determinadas áreas.

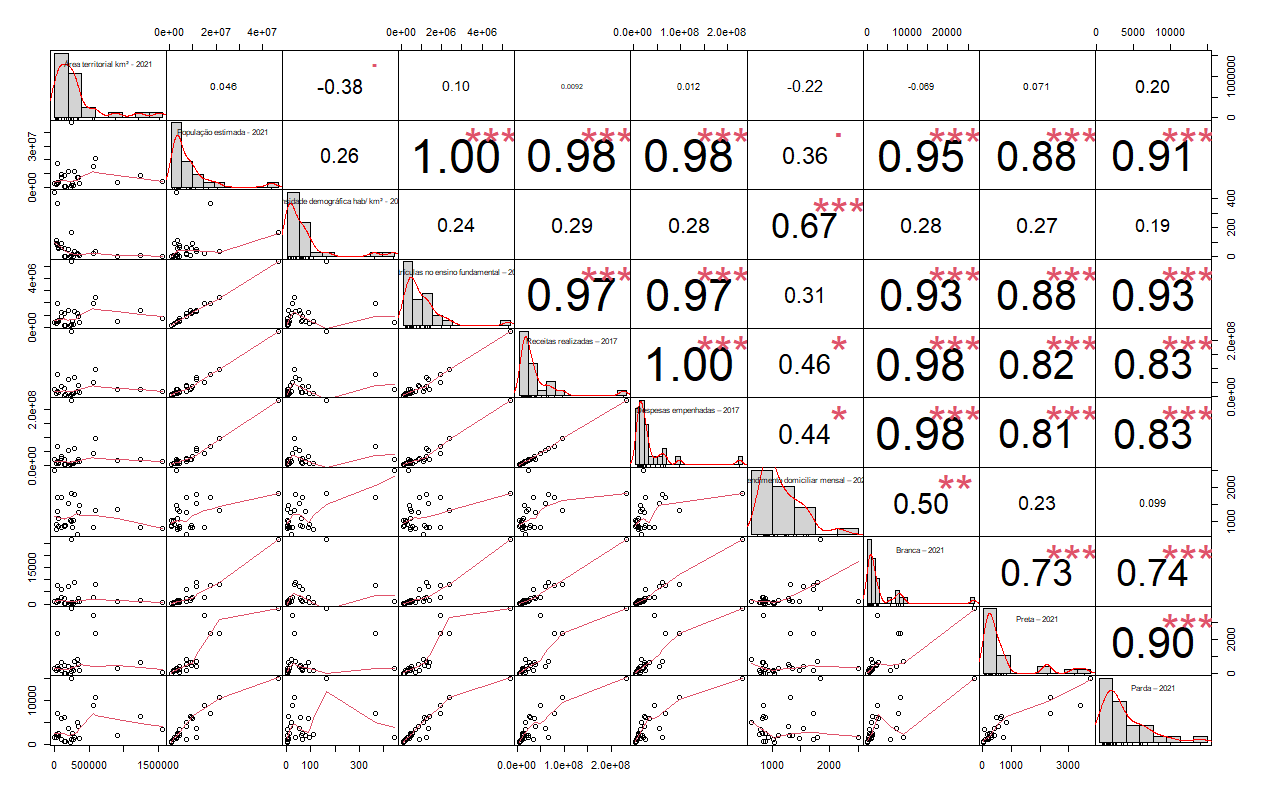


Figura 3. Matriz de correlações com toda a base de dados.

Fonte: Dados originais da pesquisa.

A seguir, temos a matriz ρ que será aplicada durante o estudo e destaca-se que as diagonais, são sempre denominadas com valores iguais a 1, o que diz respeito à correlação de uma variável com ela mesma.

(Continua)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | População estimada - 2021 | Densidade demográfica hab/ km² - 2010 | Matrículas no ensino fundamental – 2021 | Receitas realizadas – 2017 |
| População estimada - 2021 | 1 | 0.264210785142825 | 0.995654419342973 | 0.983734628673267 |
| Densidade demográfica hab/ km² - 2010 | 0.264210785142825 | 1 | 0.235557405576142 | 0.291421631812046 |
| Matrículas no ensino fundamental – 2021 | 0.995654419342973 | 0.235557405576142 | 1 | 0.967978700887015 |
| Receitas realizadas – 2017 | 0.983734628673267 | 0.291421631812046 | 0.967978700887015 | 1 |
| Despesas empenhadas – 2017 | 0.98159994920821 | 0.278559313764603 | 0.966703293286743 | 0.997450334948588 |
| Rendimento domiciliar mensal – 2021 | 0.356005471876482 | 0.668457425168091 | 0.306241127690848 | 0.463393196091971 |
| Branca – 2021 | 0.949168313040063 | 0.276927927806884 | 0.92839562186767 | 0.980645535585754 |
| Preta – 2021 | 0.875540439990127 | 0.272959654496579 | 0.879066408662091 | 0.817993159241045 |
| Parda – 2021 | 0.910439015626992 | 0.192940462369643 | 0.929433614889684 | 0.833717592886805 |

(Conclusão)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Despesas empenhadas – 2017 | Rendimento domiciliar mensal – 2021 | Branca – 2021 | Preta – 2021 | Parda – 2021 | Despesas empenhadas – 2017 |
| 0.98159994920821 | 0.356005471876482 | 0.949168313040063 | 0.875540439990127 | 0.910439015626992 | 0.98159994920821 |
| 0.278559313764603 | 0.668457425168091 | 0.276927927806884 | 0.272959654496579 | 0.192940462369643 | 0.278559313764603 |
| 0.966703293286743 | 0.306241127690848 | 0.92839562186767 | 0.879066408662091 | 0.929433614889684 | 0.966703293286743 |
| 0.997450334948588 | 0.463393196091971 | 0.980645535585754 | 0.817993159241045 | 0.833717592886805 | 0.997450334948588 |
| 1 | 0.44350154381482 | 0.977547193599471 | 0.81325558356004 | 0.833906037870181 | 1 |
| 0.44350154381482 | 1 | 0.499025817414566 | 0.229407867804128 | 0.0988486910372701 | 0.44350154381482 |
| 0.977547193599471 | 0.499025817414566 | 1 | 0.727910334686704 | 0.737521545953239 | 0.977547193599471 |
| 0.81325558356004 | 0.229407867804128 | 0.727910334686704 | 1 | 0.896069100873064 | 0.81325558356004 |
| 0.833906037870181 | 0.0988486910372701 | 0.737521545953239 | 0.896069100873064 | 1 | 0.833906037870181 |

Fonte: Dados originais da pesquisa.

Tabela 1 Matriz de correlações de Pearson

Com a matriz previamente definida, pontua–se o principal recurso para execução da PCA, porém o algoritmo prcomp(), do pacote psych, presente no software R, exige que os dados estejam padronizados pelo procedimento zcores, o que não implica em regra neste momento para o estudo e sim em limitações impostas pelo software. Averiguou-se que, representam uma quantidade de desvios-padrão que diferem uma variável da média. Em outras palavras, essa transformação fará com que todas as variáveis padronizadas, disponham de variância igual a 1 e a média igual a 0, tornando possível o cálculo de distância entre observações com ordens de grandeza ou métricas diferentes (Fávero; Belfiore, 2017). Em seguida, será necessária a rejeição de uma hipótese nula, pressupondo que os dados não se associam de forma aleatória, através do teste de esfericidade de Bartlett para adequabilidade à PCA. Utiliza-se durante este estudo, o nível de confiança de 95% que é análogo ao nível de significância de 5%. Verifica-se que o qui-quadrado da matriz ρ é equivalente a 2918.536, nos dando um p.value maior ou igual a 0,05. Salienta-se que a matriz de correlações é diferente da matriz de identidade.

Conforme descrito no tópico anterior, será realizada a extração dos dados autovalores com base na matriz de correlações de Person, portanto será gerada uma nova matriz de polinômio característico, e fazendo a extração matricial das raízes dessas informações, teremos os nossos eigenvalues, cuja a finalidade é descobrir a variabilidade máxima dos dados, ao que em outra perspectiva, pode ser vista como parcela da variação total dos dados, evidenciando um caminho inicial para a definição dos fatores propriamente ditos pelo estudo. Como resultado, temos os dados apresentados abaixo na tabela 3.

Eigenvalues

|  |  |
| --- | --- |
| PCA 1 | 6.69314175452 |
| PCA 2 | 1.48007033402 |
| PCA 3 | 0.55127053217 |
| PCA 4 | 0.17827058438 |
| PCA 5 | 0.07761751561 |
| PCA 6 | 0.01484948067 |
| PCA 7 | 0.00337809757 |
| PCA 8 | 0.00137849200 |
| PCA 9 | 0.00002320906 |

Fonte: Dados originais da pesquisa.

Tabela 2 Resultados obtidos na extração dos autovalores

Para continuidade, será evidenciada a variância compartilhada dos autovalores, onde a priori, nota-se que a soma total dos mesmos, trará como resultado o número 9, correspondente a quantidade de variáveis presentes no estudo, ressaltando que este, diz respeito a parcela da variação dos dados. A resolução obtida, presente na tabela 4, possibilita uma visão percentual, ajudando a entender a característica e importância para a construção dos fatores. Auxiliando no entendimento, o primeiro componente principal descrito como (PCA1), explica 74% da variância, o segundo componente principal descrito como (PCA2), explica 16% da variância, sendo assim de forma subsequente.

Variância compartilhada

|  |  |
| --- | --- |
| PCA 1 | 0.7436824 |
| PCA 2 | 0.1644523 |
| PCA 3 | 0.0612523 |
| PCA 4 | 0.0198078 |
| PCA 5 | 0.0086242 |
| PCA 6 | 0.0016499 |
| PCA 7 | 0.0003753 |
| PCA 8 | 0.0001532 |
| PCA 9 | 0.0000026 |

Fonte: Dados originais da pesquisa.

Tabela 3 Resultados da variância compartilhada

A tabela 5 mostrará a variância acumulativa, sendo uma forma diferente de visualizá-lo, cujo o objetivo é somar a variância até obter-se o resultado de 100% e descrever o peso de cada autovalor. Vale ressaltar que apenas 5 dos 9 autovalores, expõe 99% das relações de interdependência dos dados. O que predispõe um norte, sobre a quantidade de fatores necessários para aplicabilidade da pesquisa, entendendo que os fatores são novas variáveis que devem compartilhar a carga correlacional das variáveis anteriores e devem ser ortogonais entre si (Enap, 2019).

Variância acumulativa

|  |  |
| --- | --- |
| PCA 1 | 0.7436824 |
| PCA 2 | 0.9081347 |
| PCA 3 | 0.9693870 |
| PCA 4 | 0.9891948 |
| PCA 5 | 0.9978190 |
| PCA 6 | 0.9994689 |
| PCA 7 | 0.9998443 |
| PCA 8 | 0.9999974 |
| PCA 9 | 1.0000000 |

Fonte: Dados originais da pesquisa.

Tabela 4 Resultados da variância acumulativa

A figura 4 exibirá um relatório visual do compartilhamento das variâncias, distribuídas entre os componentes principais, conforme descrito anteriormente.

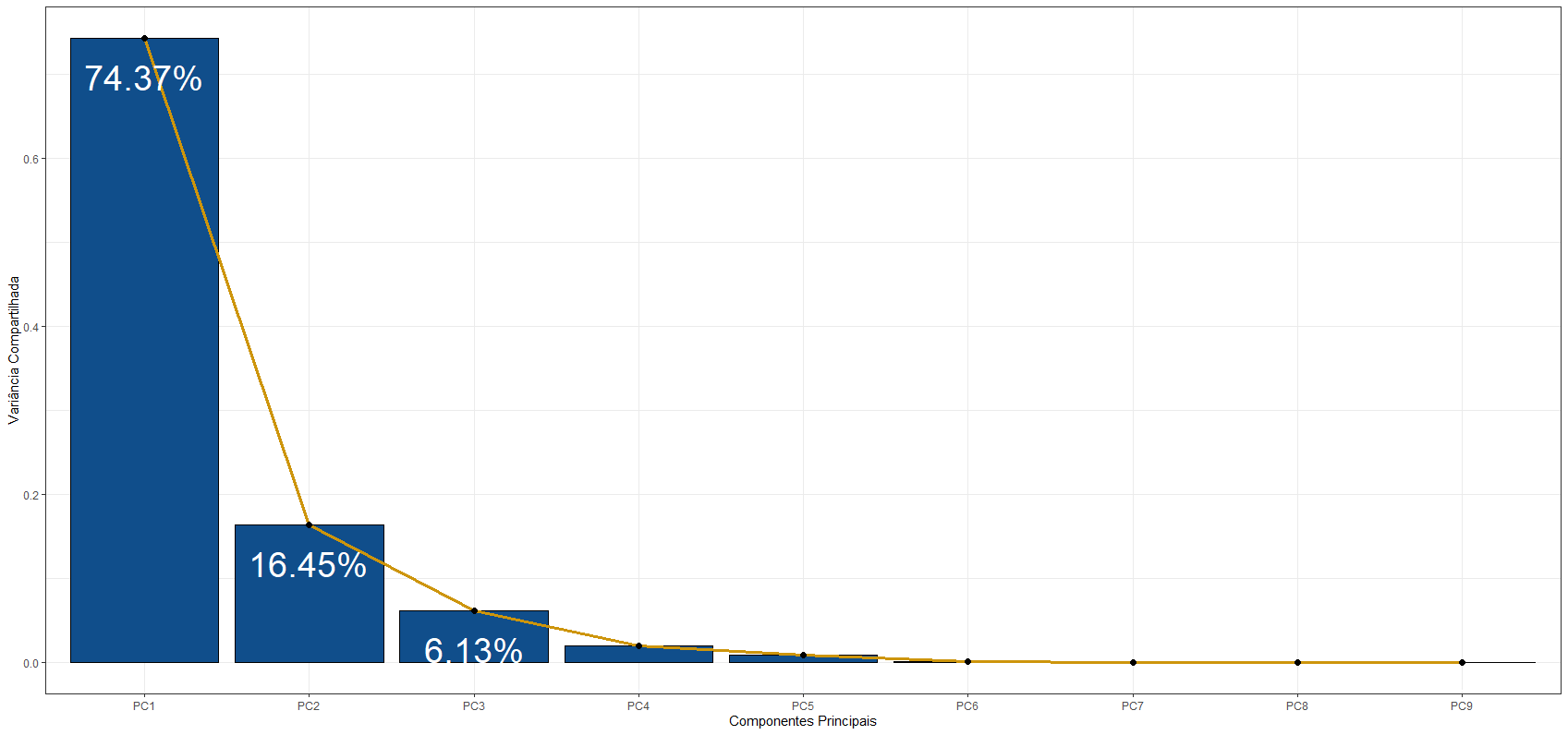


Figura 4. Exploração visual do compartilhamento das variâncias compartilhadas entre os componentes principais.

Fonte: Dados originais da pesquisa.

Posto isso, através de um sistema de equações na matriz p de autovalores, cujo descrito no tópico anterior, totaliza-se para cada autovalor, um autovetor representando o fator em que a variância entre os quais, é igual a 0. Para o estudo, será assumido o critério da raiz latente, conhecido como critério estatístico de Kaiser-Meyer-Olkin [KMO], onde serão considerados eingenvalues iguais ou maiores que 1, pois explicam ao menos uma variável inteira, o que auxiliará na definição dos fatores para análise de forma lógica e consistente não sendo um dogma para estudos desta área (Pascal, 2021). A figura 5, mostrará os 3 componentes principais da PCA por meio de uma ilustração, evidenciando o peso das variáveis para o estudo, sendo possível notar que, os dois primeiros componentes, explicam uma grande parcela dos dados, dentro do contexto que todas as variáveis encontram o grupo comum entre si e ortogonais entre os outros.

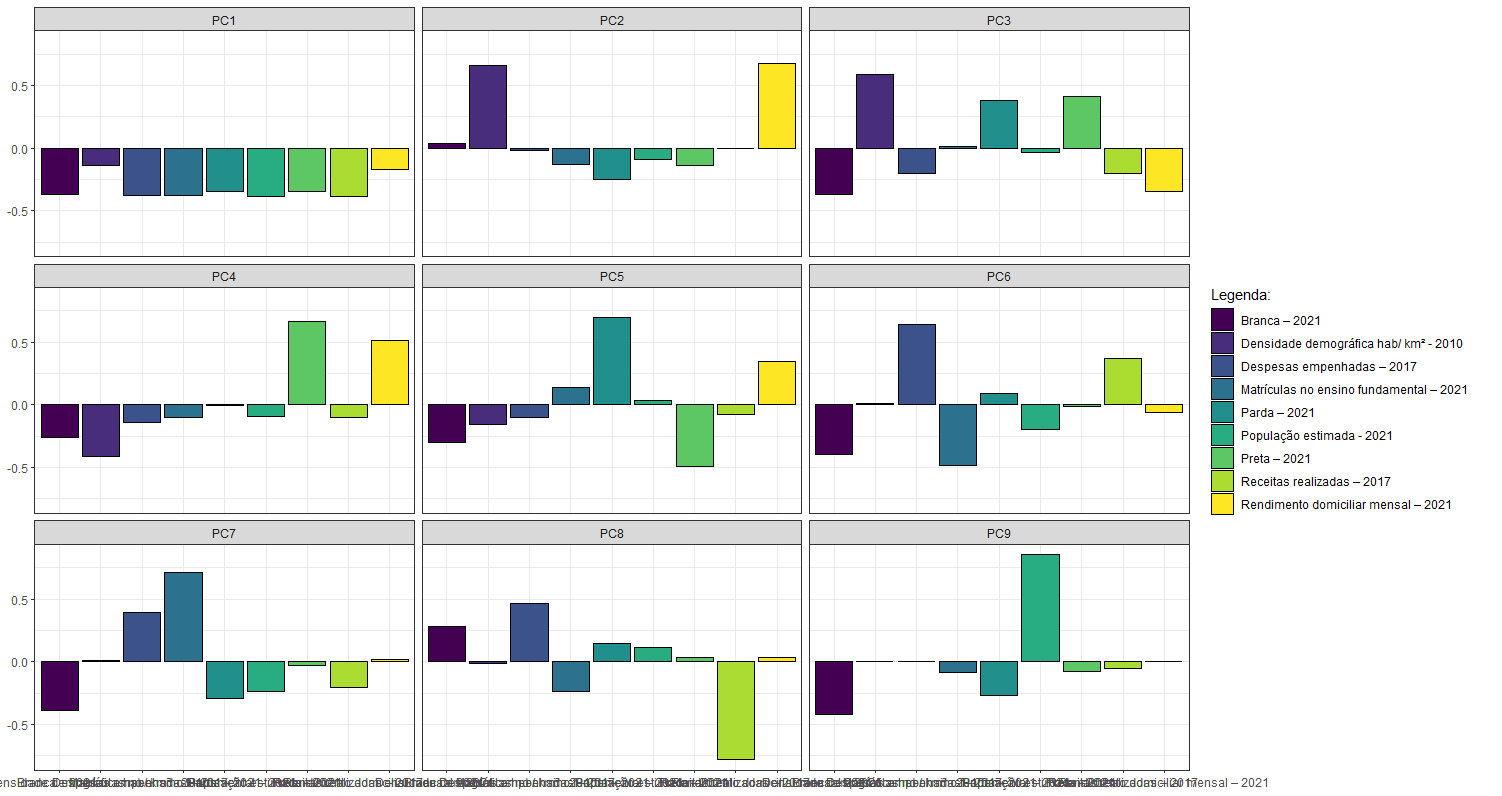
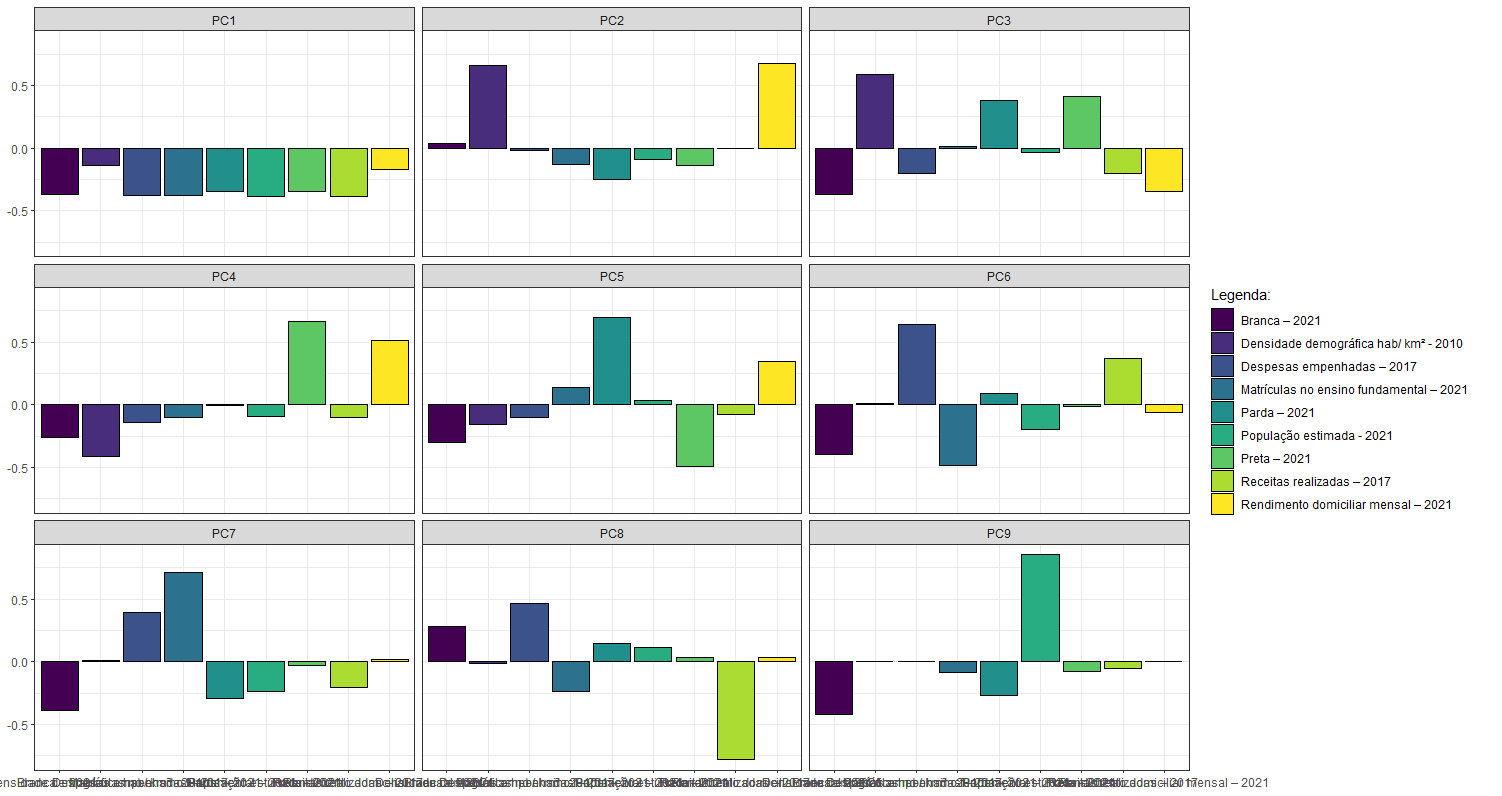


Figura 5. Peso e influência das variáveis dentro dos componentes principais.

Fonte: Dados originais da pesquisa.

Por meio da solução das raízes utilizando no cálculo, os autovalores junto aos autovetores, será dado os scores fatoriais que multiplicados pelas variáveis padronizadas pelo procedimento zscores, trará os scores fatoriais, descrito com mais informações no tópico anterior. A tabela 6, exibe 4 dos 9 fatores representados por cada estado, sendo o resultado da inferência imposta pelas variáveis da pesquisa, com correlação igual a 0.

Fatores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | **...** |
| Acre | 0.76 | -0.32 | -0.35 | -0.26 | **...** |
| Alagoas | 0.51 | -0.06 | 0.76 | -1.61 | **...** |
| Amapá | 0.78 | -0.36 | -0.30 | -0.33 | **...** |
| **...** | **...** | **...** | **...** | **...** | **...** |
| Sergipe | 0.58 | 0.08 | 0.44 | -0.86 | **...** |
| Tocantins | 0.66 | -0.19 | -0.41 | 0.24 | **...** |

Tabela 5. Fatores obtidos pela [PCA].

Fonte: Dados originais da pesquisa.

Para continuidade, serão extraídas as cargas fatoriais, que dizem respeito aos coeficientes de correlação de Pearson para cada fator com variáveis originais para elaboração do estudo, possibilitando a validação de constructos. Também será importante, a evidência das comunalidades, referindo-se à variância total compartilhada de cada variável com todos os fatores presentes na amostra (Fávero; Belfiore, 2017). Por critérios previamente estabelecidos, serão utilizados apenas dois fatores à aplicabilidade do estudo, pois ambos explicam o comportamento de ao menos, uma variável inteira. A tabela 7, apresenta um breve relatório dos resultados pontuados com as cargas fatoriais e as comunalidades.

Cargas fatoriais e comunalidades

|  |  |  |  |
| --- | --- | --- | --- |
| Variáveis | Fator 1 | Fator 2 | Comunalidades |
| População estimada - 2021. | 0.99 | -0.10 | 0.99 |
| Densidade demográfica hab km² - 2010. | 0.35 | 0.80 | 0.77 |
| Matrículas no ensino fundamental - 2021 | 0.98 | -0.15 | 0.99 |
| Receitas realizadas - 2017 | 0.98 | -0.00 | 0.97 |
| Despesas empenhadas - 2017. | 0.98 | -0.02 | 0.96 |
| Rendimento domiciliar mensal - 2021 | 0.44 | 0.82 | 0.87 |
| Branca - 2021 | 0.94 | 0.04 | 0.90 |
| Preta - 2021 | 0.88 | -0.16 | 0.80 |
| Parda - 2021 | 0.88 | -0.29 | 0.88 |

Tabela 6. Cargas fatoriais e comunalidades.

Fonte: Dados originais da pesquisa.

Para avaliação do comportamento dos dados mediante os resultados obtidos, será apresentada na figura 6, um mapa percpetual considerando as 10 variáveis específicas no âmbito de desenvolvimento humano brasileiro, que são atraídas para determinado fator com comportamento inerente, durante a composição da análise dos componentes principais. Sendo o plano com (eixo x), representando o fator 1 e o (eixo y), representando o fator 2.

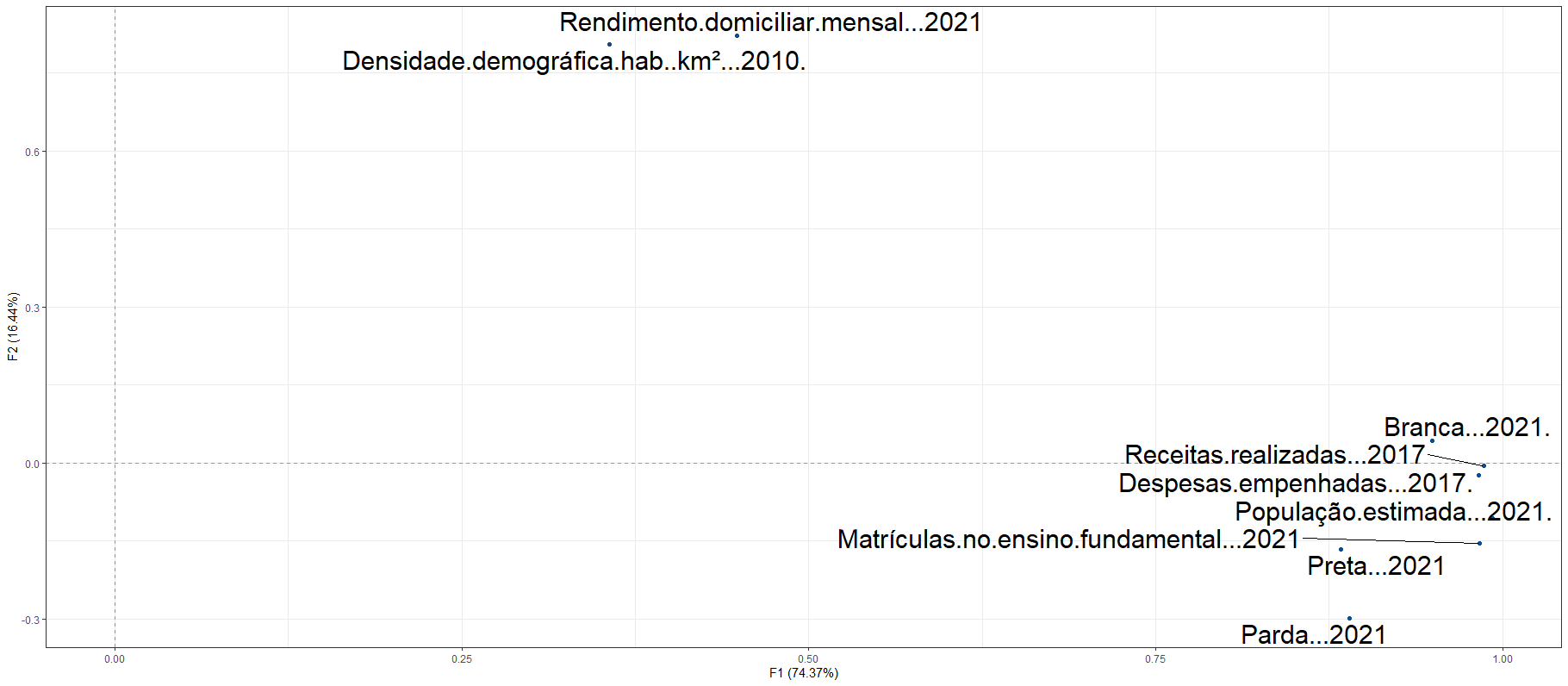


Figura 6. Mapa perceptual da distribuição dos dados.

Fonte: Dados originais da pesquisa.

Com os fatores definidos, pontua-se que temos a nossa matéria primeira para retirada de insights e validação de constructos, impostos pela pesquisa que implica em, entender o comportamento de diversas áreas com métricas e pesos diferentes no âmbito social e econômico com embasamento étnico-racial para atribuir uma lógica durante a aplicação de políticas públicas e projetos, como chave para geração de igualdade em todos os estados brasileiros. Em razão do comportamento conjunto da relação de interdependia das 10 variáveis originais, a variância foi capturada na ordem de 74,37% e foi agrupada na primeira variável, chamada de (Fator 1) e a ordem de 16,45% foi subjacente ao (Fator 2). Com a completude dessas informações, é vista a possibilidade para elaboração de um ranking, ajudando a entender as oportunidades para enfoque.

Durante a elaboração dessa métrica para nivelamento, será utilizado o critério de soma ponderada e ordenamento, pois tende à uma aceitação no meio acadêmico, onde determinada observação recebe a soma dos scores fatoriais de todas as variáveis, de acordo com as exigências impostas pelo KMO e são ponderados pelos subsequentes percentuais da variância compartilhada e ordenados pelas observações em conformidade com os resultados (Fávero; Belfiore, 2017).

Ranking de indicadores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estados | **...** | Fator 1 | Fator 2 | Pontuação | Ranking |
| São Paulo |  | 4.0787490 | 0.3483089 | 6.0665878 | 1º |
| Minas Gerais |  | 1.4278807 | 0.8245968 | 2.1237796 | 2º |
| Rio de Janeiro |  | 1.2282785 | -1.6691251 | 1.8268982 | 3º |
| **...** | **...** | **...** | **...** | **...** | **...** |
| Roraima |  | -0.7824546 | 0.1124417 | -1.1637955 | 26º |
| Amapá |  | -0.7852926 | 0.3618015 | -1.1680166 | 27º |

Figura 7. Ranking decrescente de indicadores atribuídos à estados brasileiros.

Fonte: Dados originais da pesquisa.

Observa-se durante a análise dos fatores que, estados com uma população estimada maior, são mais suscetíveis à obtenção de melhores resultados econômicos e sociais (Mariano, 2019). Entretanto, não é uma regra pois esses resultados positivos dependem de outras vaiáveis, que são fortemente afetadas pelo número populacional, tais como; despesas realizadas pelo estado e despesas empenhadas, onde apontam de forma breve e parcial, mas com seguridade, os fatores econômicos de determinada localidade. O fator educacional, previsto pela variável, dada como quantitativo de matriculas no ensino fundamental, se mostrou relevante pois os 2 primeiros estados no ranking possuem os maiores valores neste quesito. Todavia, não implica em regra pois esta sofre grande impacto dos fatores descritos anteriormente. Em questão do fator racial proposto pela pesquisa e embasado no censo do ano de 2021 publicado pelo IBGE, foi possível notar algumas similaridades, onde a população autodenominada como branca é muito maior em estados mais ricos e desenvolvidos, bem como a população parda e negra que tende a ser, diversas vezes maior em comparação com estados abaixo no ranking. Não obstante, a amostra traz parcialidade, onde pontua-se que a desigualdade racial existe, porém é possível visualizar caminhos para diminuição desta realidade, apresentando políticas e meios para implementação de renda nos estados e viabilização de educação a priori para crianças e jovens.

**Considerações Finais**

Ao final deste trabalho, é notório que os objetivos propostos inicialmente foram alcançados, cujo o princípio da análise exploratória a priori, seria a obtenção de insights e validação de constructos previamente evidenciados através de dados divulgados de forma pública para esta avaliação. Entende-se que hoje, o fator étnico-racial tem grande relevância e traz impactos diretos na qualidade de vida e desenvolvimento humano, mas o estudo propôs um aprofundamento neste conceito para avaliar de forma concisa e embasada o comportamento desta, em conjunto com outros fatores determinantes para tal desdobramento.

Conforme esperado, a análise dos componentes principais, mostrou-se eficaz à proposta estabelecida, pois apresentou resultados expressivos de fácil compreensão e visualização para pessoas de diversos nichos, sendo decisores ou não, por meio de dados estruturados e plotagem ilustrativa para os mesmos. O comportamento conjunto das variáveis selecionadas através de estados brasileiros para o estudo, serviu como ponto de partida para o seu entendimento de forma conclusiva e sincronizada com o tema, que por sua vez, dirigiu-se à uma análise para investigação das variâncias. Os 2 primeiros fatores com suas devidas cargas correlacionais, explicaram 90% do comportamento das variáveis, onde apresentaram as suas correlações positivas e negativas em conjunto com as outras. A PCA com sua especificidade, visa o agrupamento de observações com comportamentos semelhantes, pontua-se a relevância de variáveis de caráter econômico, educacional bem como, as quantitativas populacionais que potencializam o peso dos fatores, tais como, despesas realizadas, matriculas no ensino fundamental e despesas empenhadas.

A limitação presente na pesquisa, foi a coleta dos dados, pois o estudo aponta, indicadores em nível nacional, dificultando o acesso à algumas informações de forma linear para todos os estados. Houve um questionamento inicial sobre as observações utilizadas, visando o impacto positivo e negativo que elas teriam para o estudo, evitando uma extrapolação de informações que afetaria de forma direta à pesquisa, onde a decisão foi dada por meio de levantamentos teóricos e informações atuais neste ramo. É interessante notar que existem diversas ponderações a serem consideradas para o entendimento de um fator que pode ser afetado por questões culturais e regionais e as diferentes formas para tratativa e aplicabilidade em diversas áreas, servindo como um próximo passo da pesquisa às demais formas para interpretação e instrução dos fatores extraídos.

Por fim, entende-se que as funções previamente estabelecidas, são inerentes para cognição dos dados através da explicação de seus comportamentos e redução estrutural para visualização de forma didática, o que provê impactos para geração de insights, tratando-se de gestão pública e servindo como ferramenta para criação de transparência.

**Referências**

Fávero; Belfiore. 2017. Manual de Análise de Dados Estatística e Modelagem Multivariada com Excel ®, SPSS ® e Stata®. 1ed. GEN LTC. São Paulo, SP, Brasil.

Barberio. 2019. Progresso e desenvolvimento humano. 1ed. Alta Books. Rio de Janeiro, RJ, Brasil.

Nunes. 2012. Introdução a Abstração de Dados. 1ed. Bookman. Porto Alegre, RS, Brasil.

Dalberto; Ervilha; Bohn; Gomes. Índice de Desenvolvimento Humano Eficiente: Uma Mensuração Alternativa do Bem-estar da Nações. 1: 1- 5.

Pascal. 2021. Indústrias e Desenvolvimento Sustentável na África Sub-Sahariana: Desafios e Perspectivas da Industrialização Africana. Edições Nosso Conhecimento. Marselha. França.

Torres; Ferreira; Dini. 2003. Indicadores sociais: por que construir novos indicadores como o IPRS. 1: 81 – 84.

Stanton. 2017. The Human Development Index: A History. 1. UMASS. Amherst, MA, Estados Unidos.

Abud. 2019. Escola Nacional de Administração Pública [ENAP]. Análise Fatorial. 1: 1 – 18

Instituto de Pesquisa Econômica Aplicada [IPEA]. Belchior; Charão. 2011. O Longo Combate às desigualdades raciais. Disponível em: <https://www.ipea.gov.br/igualdaderacial/index.php?option=com_content&view=article&id=711>. Acesso em: 08. Jun. 2022.

**Anexos**

Link de acesso a base de dados aplicada ao estudo:

<https://docs.google.com/spreadsheets/d/1_kdO_iX-iFCHJpFAA47J3Oce6pBz0adL/edit?usp=sharing&ouid=108696731067059320539&rtpof=true&sd=true>