



ESTATÍSTICA MULTIVARIADA E ESTATÍSTICA ESPACIAL PARA CURSOS DE ENGENHARIA: INFLUÊNCIA DAS FRONTEIRAS SECAS ENTRE NAÇÕES NOS SEUS PARÂMETROS SOCIOECONÔMICOS

DOI: 10.37702/2175-957X.COBENGE.2023.4543

Fábio Gerab - prifgerab@fei.edu.br
Centro Universitário FEI

Pedro Augusto Moraes de Oliveira - pedroaugustomoraes.oliveira@gmail.com
BASF

Resumo: Este trabalho apresenta a proposta de dois componentes curriculares abordando Ciência da Dados em cursos de engenharia utilizando técnicas estatísticas e computacionais avançadas, capazes desenvolver nos estudantes a competência de transformar dados em informações estruturadas e úteis, tanto para a compreensão de fenômenos e processos como para a tomada de decisões que possibilitem ganhos, sejam eles econômicos, sociais ou ambientais. Os componentes curriculares "Estatística Multivariada e Modelagem Estatística" e "Estatística Espacial" serão ofertados como disciplinas optativas. A metodologia de ensino a ser empregada baseia-se na modelagem estatística, construída pelos alunos a partir da proposta de diversos problemas reais e originais, utilizando dados reais não consolidados, que demandam uma análise complexa em distintos graus de profundidade. A proposta em questão foi exemplificada pelo desenvolvimento de um problema que envolve tanto Estatística Multivariada como Estatística Espacial. O problema apresentado utiliza mapas os digitais dos distintos países existentes no globo terrestre e os dados para estes países disponibilizados pela CIA - Central Intelligence Agency. Neste problema busca-se identificar a existência de dependência entre os padrões de desenvolvimento de um país com padrões encontrados em seus vizinhos fronteiriços. Outros problemas semelhantes estão sendo desenvolvidos.

Palavras-chave: Educação em Engenharia, Flexibilidade Curricular, Aprendizagem por Competências, Estatística Multivariada, Estatística Espacial

ESTATÍSTICA MULTIVARIADA E ESTATÍSTICA ESPACIAL PARA CURSOS DE ENGENHARIA: INFLUÊNCIA DAS FRONTEIRAS SECAS ENTRE NAÇÕES NOS SEUS PARÂMETROS SOCIOECONÔMICOS

1 INTRODUÇÃO

A Engenharia, em suas distintas modalidades, tem um caráter técnico, porém bastante dinâmico, de forma a acompanhar as rápidas mudanças tecnológicas em curso. Neste sentido, as mudanças decorrentes do uso em larga escala de ferramentas digitais e da possibilidade de armazenamento e de tratamento massivo de dados permitiu o surgimento de toda uma área do saber conhecida por Ciência de Dados.

A Ciência da Dados (Dahar, 2013) se utiliza de técnicas estatísticas e computacionais avançadas capazes de transformar dados em informações estruturadas e úteis, tanto para a compreensão de fenômenos e processos como para a tomada de decisões racionais que possibilitem ganhos, sejam eles econômicos, sociais ou ambientais. A Ciência de dados atua na intersecção entre três grandes áreas do conhecimento, a saber: Questão (ou negócio) em estudo; Matemática e Estatística; Computação. Este conjunto de conhecimentos permite a transformação de um conjunto de dados em um conjunto de informações relevantes para a compreensão do problema estudado.

Neste contexto, a agregação de informações disponíveis de forma dispersa, em novas bases de dados consolidadas, permite análises estatísticas multivariadas envolvendo os principais indicadores quantitativos de um processo. Os resultados destas análises, em geral, revelam as estruturas de dependência e de interdependência envolvidas no problema estudado.

Para um engenheiro inserido no ambiente contemporâneo de análise e tomada de decisão, torna-se cada dia mais interessante o domínio dos fundamentos da Ciência de Dados e dos métodos estatísticos multivariados. Para tanto, buscando garantir a flexibilidade curricular do curso, propôs-se um componente curricular a ser oferecido de forma optativa para os engenheiros, das distintas modalidades de cursos de engenharia oferecidos na instituição, denominada "Estatística Multivariada e Modelagem Estatística" que abordará os conteúdos programáticos de: Regressão Linear Múltipla; Análise de resíduos; Validação de modelos estatísticos; Regressão Logística; Árvores de classificação; Análise por agrupamentos (hierárquico e não hierárquico); Análise fatorial; Análise de Componentes Principais; Análise de Correspondência; Construção de modelos preditivos; Construção de modelos de Classificação. Tais conteúdos serão abordados tanto em aulas teóricas como em aulas práticas, na modelagem de problemas reais, utilizando-se de dados reais, em problemas inovadores, buscando-se agregar competências analíticas ao estudante. Para tanto, esta disciplina proposta necessita de um conjunto, sempre crescente, de problemas e possibilidades de estudo.

Entretanto, em Ciência de Dados, tem-se que, atualmente, para além das análises estatísticas convencionais é possível avançar para o que se entende por Estatística Espacial (Druck, 2004). Em uma abordagem envolvendo Estatística Espacial as informações existentes são agregadas a uma base espacial, em um mapa, seja sobre pontos com referenciamento geográfico, seja por associação destes dados às distintas seções deste mapa, representadas por polígonos, em mapas digitais (Druck, 2004). A construção de polígonos em um mapa, que podem representar, por exemplo: bairros em

um município; municípios em um estado, estados em um país; países em um continente etc., permite estabelecer métricas de vizinhança e de distância entre estes polígonos, na forma matemática de matrizes de distância ou de matrizes de vizinhança, com vizinhos de primeira ordem, se segunda ordem (vizinhos dos vizinhos), de terceira ordem, e assim por diante.

De posse de um banco de dados estruturado, devidamente associado aos seus respectivos polígonos, polígonos estes associados aos seus vizinhos, em distintas ordens de vizinhança, poder-se-á proceder o cálculo da Correlação Espacial existente para estes dados, por intermédio do cálculo da estatística denominada de I de Moran (Smith et al, 2007). Tal estatística, permite determinar se uma variável tem ou não correlação espacial com seus vizinhos, estatisticamente significativa, por intermédio de testes estatísticos apropriados.

A existência de correlação espacial para uma variável (ou para um grupo de variáveis) indica a dependência do valor desta variável em diferentes polígonos devido a sua proximidade espacial ou a dependência de fronteiras com os demais polígonos. Em outras palavras, a Correlação Espacial mensura, para cada variável, a influência de um ente espacial (bairro, cidade, estado etc.) sobre seus entes vizinhos, de maneira a medir o quanto um vizinho se assemelha a outro.

Assim, um segundo Componente curricular intitulado "Estatística Espacial", também será ofertado de forma optativa aos estudantes de engenharia das distintas modalidades. Este componente curricular abordará em laboratório, também utilizando dados reais em problemas contextualizados, os seguintes conteúdos: Visualização de dados espaciais. Análise de vizinhanças e Estatística Espacial; Correlação Espacial, Regressão espacial e modelos espaciais autorregressivos.

Durante o processo atual de elaboração destes componentes curriculares para os cursos de engenharia os proponentes estão desenvolvendo e testando problemas de interesse, impacto e originalidade envolvendo estatística multivariada e estatística espacial. Estes problemas estão em desenvolvimento em distintos trabalhos de Iniciação Científica, de estudantes de engenharia, em condições homólogas aos estudantes aptos a cursá-los.

Para exemplificar a metodologia proposta descreveremos neste trabalho o itinerário proposto para um típico problema que pode ser abordado nestes componentes curriculares. O problema ora discutido busca investigar a existência de correlação espacial para distintas variáveis socioeconômicas referentes aos países fronteiriços. Isto é, investigar para quais indicadores socioeconômicos a existência de fronteiras secas entre países torna-se relevante para o seu desenvolvimento.

2 REVISÃO BIBLIOGRÁFICA

2.1 Estatística multivariada

A Estatística Multivariada engloba uma série de técnicas de análise estatística, paramétricas ou não paramétricas, nas quais várias variáveis são analisadas simultaneamente, em um mesmo processo. Isto permite que as interações entre as variáveis que caracterizam um processo sejam corretamente consideradas, o que não ocorre nas análises uni ou bivariadas.

Em uma visão simplificada as Análises Multivariadas podem ser segmentadas em Análises de Dependência (ou supervisionadas) ou Análises de Interdependência (ou não supervisionadas) (Hair, 2005).

As Análises de Dependência são direcionadas principalmente pela construção de modelos preditivos. Nestes modelos, em geral, um conjunto de variáveis (variáveis independentes) é utilizado para prever o comportamento de uma outra variável (variável dependente). Estes modelos preditivos podem ser de natureza paramétrica ou não paramétrica. Dentre os modelos paramétricos destaca-se a Regressão Linear Múltipla. Dentre os Modelos não-paramétricos pode-se destacar a Regressão Logística, muitas vezes entendida como um modelo de classificação, e a Regressão de Cox, utilizada na construção de modelos de sobrevivência.

Já as Análises de interdependência permitem investigar as diversas relações de interdependência existentes em um conjunto de dados com distintas variáveis medidas em um mesmo ponto amostral. Tal abordagem busca identificar a estrutura dos dados avaliando padrões nas semelhanças dos comportamentos das variabilidades dos dados e das variáveis. Tais análises podem ser realizadas tomando por estrutura o conjunto de pontos amostrais ou o conjunto de variáveis. Tais análises permitem identificar subgrupos semelhantes em um banco de dados, mensurar o número de dimensões independentes existentes neste conjunto de dados e simplificar a interpretação da base de dados a partir da redução da sua dimensionalidade. Entre as Análises de Interdependência pode-se destacar a família das Análises por Agrupamento hierárquico e não-hierárquico (Clusters) e a família das Análises Fatoriais, com destaque para a Análise de Componentes Principais (Hair, 2005).

2.2 Estatística espacial e correlação espacial

Entende-se por estatística espacial um conjunto de técnicas estatísticas que expandem as análises estatísticas convencionais (uni, bi e multivariadas) por intermédio da utilização da informação de sua localização espacial associada a cada ponto amostral. Este procedimento pode ser pontual, usando o georreferenciamento de cada ponto amostral individualmente, ou de forma agregada, associando o pertencimento de cada ponto amostral a um determinado polígono, definido em um mapa digital.

Os modelos estatísticos espaciais podem ser, tal como os modelos tradicionais, modelos preditivos usando modelos de regressão espacial, ou classificatórios. Além disso, a Estatística Espacial permite estudar a existência de correlação espacial, através da estatística I de Moran (Smith et al, 2007) e da construção de Mapas LISA - Local Indicators of Spatial Association de Correlação Espacial (Druck, 2004) (figura 1). Tais abordagens, permitem, tanto de forma visual, como em uma abordagem inferencial, determinar, para um coletivo de regiões definidas por polígonos no espaço, o quanto uma determinada região é semelhante às suas regiões vizinhas. Esta análise pode ser realizada variável a variável ou utilizando-se de indicadores que representam o comportamento global de um dado subconjunto de variáveis.

Assim, as análises de correlação espacial são instrumentos eficientes para se avançar no entendimento das relações entre entes espaciais, sejam eles bairros, sejam eles municípios, ou sejam eles estados nacionais. Como será melhor apresentado adiante, neste trabalho o objeto de estudo será a investigação da existência de correlação espacial entre estados nacionais, em escala mundial.

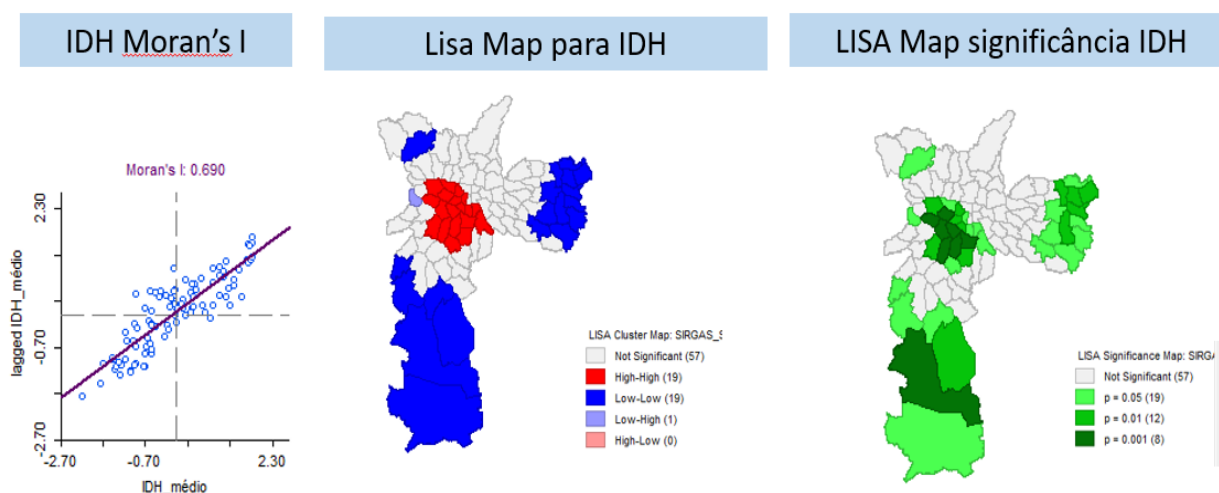
2.3 Bases de dados primárias e secundárias

Em qualquer estudo estatístico e matéria prima original são dados em quantidade e qualidade suficientes para permitir uma boa análise. Em uma análise espacial, os pontos amostrais podem ser os próprios polígonos que definirão as relações de vizinhança. Os

dados podem ser primários, obtidos diretamente pelo autor do estudo, ou secundários, extraídos de fontes de dados confiáveis. Este trabalho fará usos de dados secundários.

Figura 1: Exemplo de Análise de Correlação Espacial para a variável socioeconômica IDH nos distintos distritos do Município de São Paulo- SP

Análise de autocorrelação espacial IDH
(vizinho grau 1 - Queen)



Fonte: Autores

3 PROBLEMA DE ESTUDO PROPOSTO

O problema proposto investigará a existência de correlação espacial entre países, entendidos como estados nacionais autônomos, e seus vizinhos. Esta correlação será investigada por intermédio de um grande conjunto de variáveis geográficas, sociais e econômicas, associadas a cada país presente no estudo. Uma vez que estudos de estatística espacial exigem a alocação espacial destas variáveis, esta alocação fará uso de mapas digitais dos estados nacionais, representados por polígonos digitais sobre o globo terrestre. Estes polígonos digitais permitem a construção de matrizes de vizinhança entre estados nacionais, por intermédio de suas fronteiras secas. A proximidade entre nações, muitas vezes trazem consigo semelhanças e diferenciações. Processos culturais e políticos que envolveram o surgimento dos estados nacionais, tais como o processo de aglutinação de unidades políticas menores, os processos de colonização e de independência, os acidentes geográficos (rios e outros corpos de água, cordilheiras, desertos etc.) separando países, as guerras em momentos passados, a complementaridade econômica e as suas semelhanças linguísticas e culturais permitem entender, caso a caso, as razões para a existência de semelhanças ou diferenças nos padrões de conformação e de desenvolvimento das nações.

Entretanto, um olhar coletivo, baseado em análises estatísticas próprias, pode evidenciar quais são as variáveis socioeconômicas relevantes para o estabelecimento de semelhanças, ou de distinções entre nações vizinhas. Desta forma, este projeto buscará estabelecer, a partir de uma análise coletiva, quais são as variáveis impactantes para o

cenário de desenvolvimento de uma nação, a luz do cenário de desenvolvimento de seus vizinhos.

Para tanto, buscaremos uma análise global, envolvendo quase a totalidade dos estados nacionais hoje reconhecidos no mundo, utilizando de forma conjunta dados secundários das principais variáveis socioeconômicas quantitativas disponíveis e das representações espaciais e de fronteira destes respectivos estados nacionais.

4 OBJETIVOS

Este trabalho tem por objetivo:

- Desenvolver um problema de estudo que possa ser utilizado no componente curricular "Estatística Multivariada e Modelagem Estatística" e, posteriormente, revisitado no componente curricular "Estatística Espacial" utilizando o enriquecimento de dados na construção de bases de dados reais de forma a:

- Identificar constructos relativos aos padrões de desenvolvimento socioeconômicos de uma nação.

- Permitir estabelecer, a partir de uma análise estatística, o conjunto de variáveis com impacto relevante no cenário de desenvolvimento de uma nação, a luz do cenário de desenvolvimento de seus vizinhos.

- Desenvolver as habilidades analíticas do estudante de Iniciação Científica, tomado como representante dos estudantes candidatos ao curso, capacitando-o para a correta utilização de ferramentas estatísticas mais elaboradas, tanto em Estatística Multivariada como em Estatística Espacial.

5 METODOLOGIA

O bom andamento do projeto proposto perpassa uma série de condicionantes que envolvem a obtenção das bases de dados necessárias, a obtenção dos mapas digitalizados necessários para a análise, o acesso aos softwares de análise estatística compatíveis com a proposta e a correta concepção da estratégia de análise e sequenciamento das etapas envolvidas. Estes pontos serão abordados a seguir.

5.1 Bases de dados abertas e o The World Factbook

Neste trabalho necessitamos de dados confiáveis disponíveis para uma grande gama de estados nacionais reconhecidos como tal em nosso planeta. Para tanto, utilizaremos os dados secundários disponibilizados pela CIA - Central Intelligence Agency, do governo dos Estados Unidos da América. Estes dados são organizados por país pela CIA e disponibilizados em seu repositório The World Factbook (CIA, 2023) (<https://www.cia.gov/the-world-factbook/>). Como exemplo, podemos navegar pelos dados coletados para o Brasil em (<https://www.cia.gov/the-world-factbook/countries/brazil/>) ou para os dados referentes à Botswana em (<https://www.cia.gov/the-world-factbook/countries/botswana/>). Os dados em questão estão organizados em variáveis que perpassam doze dimensões referentes a cada país, a saber: Introduction; Geography; People and Society; Environment; Government; Economy; Energy; Communications; Transportation; Military and Security; Terrorism e Transnational Issues.

Estas informações podem ser quantitativas e qualitativas. Podem estar consolidadas para todos os países reconhecidos ou somente para alguns. O trabalho ora proposto fará uso de variáveis quantitativas consideradas relevantes, disponíveis para a grande maioria

dos países nas dimensões Geografia, População e Sociedade e Economia. Serão incorporadas à análise, em princípio, 32 variáveis quantitativas, listadas por país, conforme lista a seguir: Área territorial; População total; Idade média da população; Taxa de crescimento populacional; Taxa de natalidade; Taxa de mortalidade; Taxa líquida de migração; Razão por sexo médio populacional; Mortalidade materna; Mortalidade infantil; Expectativa de vida; Taxa de fertilidade; HIV/AIDS Taxa de prevalência em Adultos; Taxa de obesidade adulta; Consumo de álcool per capita; Tabagismo; Gasto % do GPD com educação; Produto interno bruto GPD; Taxa de crescimento do GPD; GPD per capita; Taxa de inflação; Taxa de crescimento da produção industrial; Força de trabalho; Taxa de desemprego; Índice Gini; Dívida pública % do GPD; Carga tributária; Balança comercial corrente; Exportações; Importações; Reservas em moeda estrangeira; Dívida externa.

Cada uma destas variáveis pode ser baixada em uma tabela de dados (.csv) distinta em processo de download direto do site da CIA. Assim, para o início do trabalho será necessário consolidar estes dados em uma única tabela multivariada.

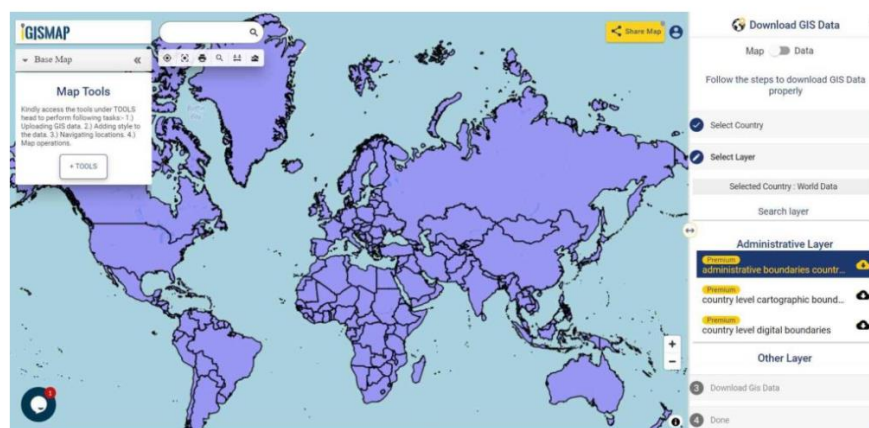
5.2 Bases de mapas digitais abertos e o IGISMAP

A fonte de mapas digitais referentes aos estados nacionais será a plataforma IGISMAP (IGISMAP, 2023) (<https://www.igismap.com/>). Entre diversas funcionalidades esta plataforma permite o download de diversos mapas digitais no formato shapefile (.shp), entre eles mapa mundi com os polígonos referentes aos estados nacionais, conforme Figura 2.

Figura 2: Mapa Mundi Digital

Download World Countries Boundaries Shapefile Data

This shapefile covers 246 countries in the world.



World Country Boundaries

[Download World Country Boundaries Shapefile](https://www.igismap.com/download-world-shapefile-free-country-borders-continent/)

Fonte: <https://www.igismap.com/download-world-shapefile-free-country-borders-continent/>
Acesso em 13/02/2023

5.3 Softwares de estatística multivariada e de estatística espacial

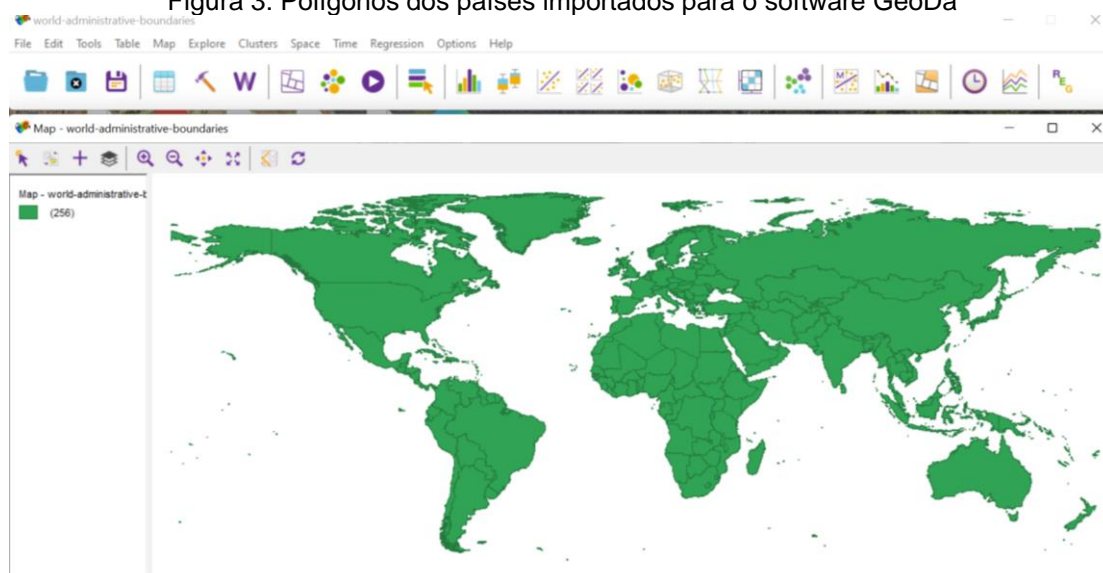
O software *R*: A language and environment for statistical computing (*R Core Team*, 2023) é um software livre desenvolvido para análises estatísticas, bastante completo, porém de utilização menos intuitiva. Já o software SPSS Statistics (IBM Corp., 2013) é um

software comercial dedicado à análise estatística bastante completo e de utilização mais amigável quando comparado ao R. Ambos os softwares podem ser usados na proposta deste trabalho.

O software GeoDa (Anselin, L. et al, 2023) (<https://geodacenter.github.io>) é um software livre, de fácil download e instalação, desenvolvido por pesquisadores do Centro de Ciência de Dados Espaciais da Universidade de Chicago (EUA), oferece ferramentas de análise estatística espacial para pesquisadores e usuários da área. O GeoDa é um software com excelente usabilidade adequado à elaboração de análises espaciais, concatenadas às principais técnicas de análise estatística multivariada, tais como: estatística descritiva com aglomeração espacial, confecção de gráficos adequados à espacialização, Análise de Componentes Principais, Análise por Agrupamento segundo distintos métodos de aglomeração, Correlação Espacial e construção de Modelos preditivos de Regressão Espacial. Para tanto, o GeoDa permite tanto a incorporação de mapas digitais como posicionamento espacial (geocodificação) além da incorporação de tabela de dados com distintos tipos de variáveis associadas aos dados espaciais, sejam eles polígonos, sejam eles pontos geocodificados.

A Figura 3 apresenta a máscara do mapa digital com todos os estados nacionais, extraído de (<https://www.igismap.com/download-world-shapefile-free-country-borders-continents/>) e já incorporados ao Software GeoDa.

Figura 3: Polígonos dos países importados para o software GeoDa



Fonte: Autores

Este projeto é um projeto de médio porte, que envolverá diversas etapas. Etapas estas que poderão ser reavaliadas em função dos resultados das análises estatísticas realizadas. A estratégia a ser seguida é, resumidamente:

- I. Identificar as fronteiras secas entre nações através de mapas digitais. – (<https://www.igismap.com/>)
- II. Identificar a fonte de dados socioeconômicos das nações – (<https://www.cia.gov/the-world-factbook/>)
- III. Definir variáveis disponíveis

- IV. Fazer a agregação destas variáveis em um único banco de dados
- V. Tratar adequadamente as variáveis para que sejam comparáveis em valores absolutos e/ou per capita.
- VI. Definir os critérios de inclusão de países na análise, considerando uma área territorial mínima e a existência de fronteiras secas.
- VII. Fazer uma Análise de Componentes Principais destas variáveis definindo o número de grupos característicos de nações
- VIII. Definir variáveis proxy para estes grupos
- IX. Realizar uma análise de Clusters, classificando as variáveis utilizadas na análise
- X. Importar as variáveis para o mapa global
- XI. Fazer correlação espacial para as variáveis proxy (I de Moran)
- XII. Ranquear as variáveis segundo o I de Moran
- XIII. Identificar as principais influências nos padrões de desenvolvimento e a respectiva influência da espacialidade, definida pelas vizinhanças com fronteiras secas.

6 RESULTADOS

O trabalho encontra-se em desenvolvimento, tendo várias das etapas descritas na metodologia já finalizadas. Abaixo estão apresentados os resultados parciais obtidos e, a seguir a descrição das etapas seguintes.

6.1 Resultados parciais

Foi realizada a construção e validação da base de dados agregando todas as 32 variáveis disponibilizadas pela CIA, segundo os estados nacionais correspondentes em uma única base de dados.

A partir desta base elaborou-se uma análise prévia das variáveis. As análises estatísticas foram realizadas utilizando-se o software SPSS v.22. A estatística descritiva apontou vários valores faltantes. Foram selecionados estados nacionais com ao menos um milhão de habitantes, excluindo-se assim unidades autônomas muito pequenas, consideradas não relevantes para esta análise. Após este filtro 160 estados nacionais se mantiveram na análise. O próximo passo da análise foi a seleção das variáveis com informações para a grande parte dos países. Selecionou-se 24 variáveis com medidas para simultâneas para 142 países. Este recorte da base de dados foi utilizado em uma análise multivariada preliminar. Utilizou-se uma análise de Componentes principais, retendo os fatores com autovalores superiores a 1 e aplicando-se posteriormente uma Rotação VARIMAX (Hair, 2004).

As variáveis incluídas no modelo, após a aplicação dos filtros, foram: Área territorial; População total; Idade média da população; Taxa de crescimento populacional; Taxa de natalidade; Taxa de mortalidade; Taxa líquida de migração; Mortalidade materna; Mortalidade infantil; Expectativa de vida; Taxa de fertilidade; HIV/AIDS Taxa de prevalência em Adultos; Taxa de obesidade adulta; Consumo de álcool per capita; Tabagismo; Gasto % do GPD com educação; Taxa de crescimento do GPD; GPD per capita; Taxa de inflação; Taxa de crescimento da produção industrial; Taxa de desemprego; Índice Gini; Dívida pública % GPD; Importações.

Como mostra a Tabela 1 abaixo, um conjunto de sete Componentes principais retidas explicou 75% da variabilidade total dos dados. Cada uma destas componentes está associada a um comportamento independente de variabilidade dos dados e relaciona-se a um distinto constructo.

Tabela 1: Componentes retidas na Análise de Componentes Principais

Componente	Variância total explicada								
	Valores próprios iniciais			Somadas de extração de carregamentos ao quadrado			Somadas rotativas de carregamentos ao quadrado		
	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	7,836	32,652	32,652	7,836	32,652	32,652	6,785	28,270	28,270
2	2,382	9,926	42,578	2,382	9,926	42,578	2,379	9,914	38,185
3	1,995	8,311	50,889	1,995	8,311	50,889	2,101	8,753	46,938
4	1,725	7,187	58,077	1,725	7,187	58,077	2,059	8,580	55,518
5	1,536	6,401	64,477	1,536	6,401	64,477	1,926	8,027	63,545
6	1,443	6,014	70,492	1,443	6,014	70,492	1,515	6,314	69,858
7	1,184	4,932	75,423	1,184	4,932	75,423	1,335	5,565	75,423
8	880	3,704	79,127						
22	,019	,079	99,982						
23	,004	,018	100,000						
24	2,989E-6	1,245E-5	100,000						

Método de Extração: Análise de Componente Principal.

Fonte: Autores

A Tabela 2 apresenta as variáveis associadas às componentes retidas por meio da sua maior carga fatorial. Estes agrupamentos permitem a identificação dos constructos associados a cada uma das componentes retidas e a atribuição *ad hoc* de um nome representativo do constructo. Os constructos identificados, em ordem de importância para a explicação da variabilidade dos dados, foram: Evolução Demográfica, Educação e Saúde, Tamanho da Nação, Mortalidade, Desemprego, Pujança econômica, Estabilidade econômica.

Tabela 2: Constructos e Proxys das Componentes Principais Retidas

Componente	Constructo associado	Correlação	Variável proxy
1	Evolução Demográfica	<ul style="list-style-type: none"> - Idade média da população (positiva) - Expectativa de vida (positiva) - Consumo de álcool per capita (positiva) - Taxa de obesidade adulta (positiva) - Tabagismo (positiva) - Taxa de fertilidade (negativa) - Taxa de natalidade (negativa) - Mortalidade infantil (negativa) - Taxa de crescimento populacional (negativa) - Mortalidade materna (negativa) 	- Idade média da população
2	Educação e Saúde	<ul style="list-style-type: none"> - HIV/AIDS Taxa de prevalência em Adultos - Taxa de crescimento da produção industrial - Taxa de crescimento do GPD - Índice Gini - Gasto % do GPD com educação 	- HIV/AIDS Taxa de prevalência em Adultos
3	Tamanho da nação	<ul style="list-style-type: none"> - Importações - População total - Área territorial 	- População total
4	Mortalidade	<ul style="list-style-type: none"> - Taxa de mortalidade 	- Taxa de mortalidade
5	Desemprego	<ul style="list-style-type: none"> - Taxa de desemprego 	- Taxa de desemprego
6	Pujança econômica	<ul style="list-style-type: none"> - GPD per capita - Taxa de líquida de migração 	- Taxa de líquida de migração
7	Estabilidade econômica	<ul style="list-style-type: none"> - Dívida pública 5 do GPD - Taxa de inflação 	- Dívida pública

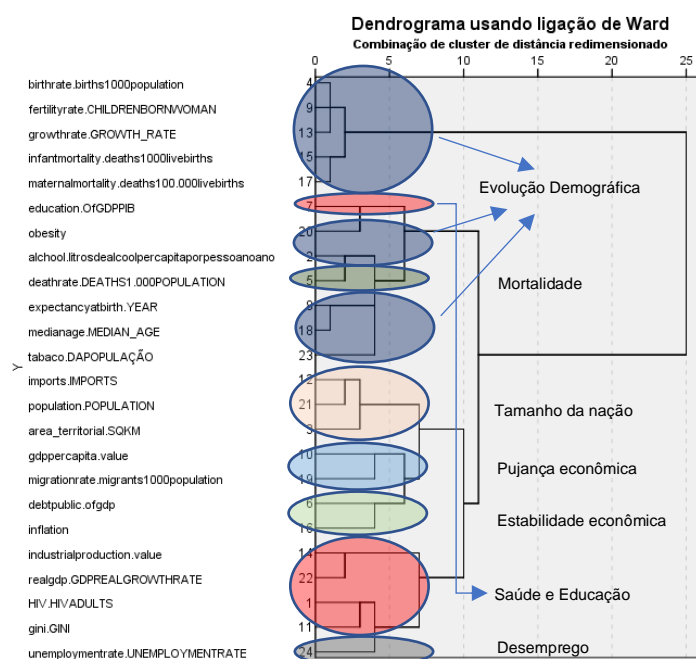
Fonte: Autores

Para cada um destes constructos selecionou-se uma variável proxy para representá-lo. A seleção foi realizada tanto pela proximidade semântica com o constructo como pela carga fatorial (maior carga fatorial) da variável. Assim: a variável Idade média da população será a representante do constructo Evolução Demográfica; a variável HIV/AIDS Taxa de prevalência em Adultos será a representante do constructo Educação e Saúde; a variável

População total será a representante do constructo Tamanho da nação; a variável Taxa de mortalidade será a representante do constructo Mortalidade; a variável Taxa de desemprego será a representante do constructo Desemprego; a variável Taxa líquida de migração será a representante do constructo Pujança econômica e a variável Dívida pública será a representante do constructo Estabilidade Econômica.

De maneira complementar à Análise de Componentes Principais procedeu-se, para o mesmo conjunto de 24 variáveis, a aglomeração por Cluster Hierárquico (Hair, 2005). As variáveis foram aglomeradas, após a sua padronização, utilizando as Distâncias Euclidianas Quadráticas e o método de Ward de aglomeração. A Figura 4 mostra o dendrograma relativo a esta aglomeração. Embora a estrutura de aglomeração siga pressupostos matemáticos distintos da análise de Componentes Principais, percebe-se que a estrutura de aglutinação das variáveis obtida por aglomeração é bastante semelhante àquela obtida por Análise de Componentes Principais. A maior diferença entre os métodos refere-se à componente Evolução Demográfica, pois, dada à existência de correlações positivas e negativas entre o componente principal e as suas variáveis (terceira coluna da Tabela 2), a análise de cluster desmembrou esta componente, pois não trata de forma automática a inversão de polos das escalas. Tal fato aumenta a distância entre as variáveis com correlação positiva das variáveis com correlação negativa.

Figura 4: Dendrograma resultante do Agrupamento Hierárquico das variáveis selecionadas



Fonte: Autores

6.2 Próximas Etapas

Uma vez já definidas as variáveis proxy, representativas dos constructos identificados, descritas acima, esta camada de variáveis proxy será adicionada ao Mapa Mundi Digital constante no software GeoDa. Este processo possibilitará a mensuração da correlação espacial entre nações fronteiriças para cada uma das variáveis proxy selecionadas e, portanto para cada um dos constructos identificados. A construção de

mapas LISA, como exemplificado na figura 1, permitira a visualização espacial destas correlações, respondendo ao problema de estudo proposto no item 3.

7 CONSIDERAÇÕES FINAIS

A construção de itinerários formativos flexíveis em cursos de engenharia, abordando temas atuais, de grande potencial de utilização nas atividades profissionais, utilizando um ensino baseado em Competências e Habilidades é a tônica da proposta em questão.

Para tanto, o foco deste trabalho foi o desenvolvimento de uma proposta de ensino que permita a obtenção de competências analíticas envolvendo Estatística e Ciência de Dados em dois componentes curriculares complementares, "Estatística Multivariada e Modelagem Estatística" e "Estatística Espacial", ofertados como optativas nestes cursos. Ambos estes componentes baseiam-se em atividades desenvolvidas em laboratório, versando sobre problemas reais, utilizando de dados reais, muitas vezes não consolidados. As atividades devem permitir o desenvolvimento do assunto em estudo em sua real complexidade, de maneira a escalar saberes adquiridos em distintas etapas do processo formativo em questão.

Para exemplificar o conceito destes componentes curriculares e das atividades por eles propostas, apresentou-se um problema de análise, em desenvolvimento. O problema em questão analisa dados de acesso livre, porém não consolidados e envolvem abordagens analíticas tanto pertinentes à Estatística Multivariada como à Estatística Espacial. O Problema está sendo elaborado por um estudante de Graduação em Engenharia, em um projeto de Iniciação Científica, de forma que a visão de um estudante do curso possa estar presente no desenvolvimento da atividade. Outros problemas, versando sobre conteúdos estatísticos distintos do aqui apresentado, seguindo a mesma proposta de ensino, estão em elaboração por outros estudantes de Iniciação Científica.

Os problemas desenvolvidos e propostos ao longo destes componentes curriculares servirão de alicerce para o desenvolvimento de soluções próprias dos estudantes destas disciplinas, em atividades desenvolvidas em grupo, abordando diferentes recortes dos problemas iniciais.

REFERÊNCIAS

ANSELIN, L. et al. **GeoDa - An Introduction to Spatial Data Science**. 2023) (disponível em: <https://geodacenter.github.io>) Acesso em: 20 mar, 2023.

CIA – **Central Intelligence Agency – USA**. (disponível em: <https://www.cia.gov/the-world-factbook/>). Acesso em: 27 fev. 2023.

DHAR, Vasant. Data science and prediction. **Communications of the ACM**, v. 56, n. 12, p. 64-73, 2013.

DRUCK, S. et al. **Análise Espacial de Dados Geográficos**". Brasília, EMBRAPA, 2004 (ISBN: 85-7383-260-6). (disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/index.html>). Acesso em: 10 Mar, 2023.

FÁVERO, P.F. et al. , **Análise de dados: modelagem multivariada para tomada de decisão**. Elsevier, Rio de Janeiro, 2009.

Hair, J.F. et al. **Análise multivariada de dados**, Bookman, 2005.

IBM Corp. Released 2013. **IBM SPSS Statistics for Windows**, Version 22.0. Armonk, NY: IBM Corp.

IGISMAP. Disponível em: <https://www.igismap.com/>. Acesso em: 27 fev., 2023.

MARQUES, J.; CASTRO, E.; BHATTACHARJEE, A. **A localização urbana na valorização residencial: Modelos de autocorrelação espacial**. Actas do XV Encontro da APDR, 2009.

R Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna. Disponível em: <https://www.R-project.org>. (Acesso em março 10, 2023).

SMITH, Michael J.; GOODCHILD, Michael F.; LONGLEY, Paul. **Geospatial analysis : a comprehensive guide to principles, techniques and software tools**. Leicester, UK: Matador, 2007. (disponível em: <https://www.spatialanalysisonline.com/HTML/index.html>)

Abstract: *This work proposes two curricular components addressing Data Science in engineering courses using advanced statistical and computational techniques, capable of developing in students the competence to transform data into structured and useful information, both for understanding phenomena and processes and for making decisions that enable gains, whether economic, social, or environmental. The curricular components "Multivariate Statistics and Statistical Modeling" and "Spatial Statistics" will be offered as optional subjects. The teaching methodology to be used is based on statistical modeling, built by the students from the proposal of several real and original problems, using unconsolidated real data, which demand a complex analysis in different degrees of depth. The proposal in question was exemplified by the development of a problem involving both Multivariate Statistics and Spatial Statistics. The presented problem uses digital maps of the different countries existing in the terrestrial globe and the consolidated data for these countries by the CIA - Central Intelligence Agency. In this problem, we seek to identify the existence of dependence between the development patterns of a country with patterns found in its border neighbors. Other similar study problems are being developed.*

Keywords: *Engineering Education, Curriculum Flexibility, Competency-Based Learning, Multivariate Statistics, Spatial Statistics*