Relatório de Dados Socioeconômicos & Obstétricos

Wesley Pereira da Silva [1]

24 de janeiro de 2023

[1] Lattes

Conteúdo

Apresentação

1. *Contexto:* O presente relatório objetiva apresentar informações detalhadas sobre como as características socioeconômicas estão relacionadas a riscos gestacionais. Investigamos um conjunto de variáveis demográficas e indicadores obstétricos do Brasil de dois anos distintos: 2000 e 2010. Por meio de análise de agrupamento, identificamos grupos de Estados com condições similares de desenvolvimento e variáveis de associação estatisticamente significativa. O conjunto de dados do experimento relatado neste relatório, foi reunido e organizado pelo Observatório Obstétrico Brasileiro (OOBr)[1].

2. *Métodos:* Esse foi um estudo ecológico baseado em população com análise de agrupamento de indicadores obstétricos associados a indicadores demográficos. O conjunto de dados utilizado foi construído a partir de dados oriundos do Censo Nacional - edições de 2000 e 2010, coletadas no sítio do Atlas Brasil[2]; e de dados do Sistema de Informação sobre Nascidos Vivos (SINASC)[3]. Para verificar a correlação de características socioeconômicas e associado a risco gestacional, buscamos agrupar as variáveis pela técnica k-Means. Os dados dos estados brasileiros e do Distrito Federal, foram tratados como vetores matemáticos e submetidos ao algorítimo citado. Foi utilizada uma técnica *naive* de otimização, denominada método do arco, em que buscamos avaliar a quantidade ideal de grupos para o conjunto de dados, a partir da maximização das distâncias vetoriais intra-extra grupos. As diferentes magnitudes de medidas do conjunto de dados foram padronizadas para valores de uma distribuição normal, assim, $X \to X_N$, sendo $X_N \sim N(\mu = 0, \sigma = 1)$. Evidenciamos a associação de variáveis por meio do calculo de correlação de Spearman e filtrando os valores de correlação superiores 0,6

[1] linktr.ee/observatorioobstetricobr
[2] www.atlasbrasil.org.br
[3] sinasc.saude.gov.br

e inferiores a -0,6.

3. *Fomento:* OOBr

1 Conjunto de Dados

A fim de buscar uma abordagem compatível com o estudo de [1], que aponta grupos socioeconômicos nos quais há há riscos gestacionais associados a contaminação por COVID-19 no Brasil, selecionamos um conjunto de variáveis socioeconômicas e obstétricas. A seguir, seus nomes, descrições e estruturas.

1.1 Descrição das Variáveis

Conjunto dados obstétricos:

- Nascimentos - Quantidade de Nascimentos
- Porc_prematuros - Percentual de Prematuros
- Porc_cesareas - Percentual de Cesáreas
- Porc_grav_multipla - Percentual de Gravidez Multiplas
- Porc_anomalias - Percentual de Anomalias
- Porc_nenhuma_consulta - Percentual de nenhuma consulta
- Porc_consulta7mais - Percentual de mais de 7 consultas
- Porc_feminino - Percentual do sexo feminino
- Porc_raca_mae_branca - Percentual de mãe com raça branca
- Porc_raca_mae_negra - Percentual de mãe com raça negra

- Porc_peso_menor_2500 - Percentual de peso menor de 2.5Kg
- Porc_apgar1_menor_7 - Percentual de APGAR1 menor 7
- Porc_apgar5_menor_7 - Percentual de APGAR5 menor 7

Conjunto de dados socioeconômicos:

- GINI - Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Seu valor varia de 0, quando não há desigualdade (a renda domiciliar per capita de todos os indivíduos tem o mesmo valor), a 1, quando a desigualdade é máxima (apenas um indivíduo detém toda a renda). O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.
- T_DES - Percentual da população economicamente ativa (PEA) nessa faixa etária que estava desocupada, ou seja, que não estava ocupada na semana anterior à data do Censo mas havia procurado trabalho ao longo do mês anterior à data dessa pesquisa.
- T_ANALF15M - Taxa de analfabetismo - 15 anos ou mais PAREDE - Percentual de pessoas em domicílios com paredes inadequadas
- T_AGUA - Percentual da população em domicílios com água encanada
- T_BANAGUA - Percentual da população em domicílios com banheiro e água encanada
- AGUA_ESGOTO - Razão entre as pessoas que vivem em domicílios cujo abastecimento de água não provem de rede geral e cujo esgotamento sanitário não é realizado por rede coletora de esgoto ou fossa séptica e a população total residente em domicílios particulares permanentes multiplicado por 100. São considerados apenas os domicílios particulares permanentes.
- T_LIXO - Percentual da população em domicílios com coleta de lixo

1.2 Estrutura dos Dados

A estrutura dos dados pode ser vista no Quadro 1.2. O conjunto possui 54 registros, relativos a 21 variáveis, descritas na seção anterior.

Quadro 1.2- Indicadores Socioecnômicos & Obstétricos

```
Data columns (total 21 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Nascimentos            54 non-null     int64
 1   Porc_prematuros        54 non-null     float64
 2   Porc_cesareas          54 non-null     float64
 3   Porc_grav_multipla     54 non-null     float64
 4   Porc_anomalias         27 non-null     float64
 5   Porc_nenhuma_consulta  54 non-null     float64
 6   Porc_consulta7mais     54 non-null     float64
 7   Porc_feminino          54 non-null     float64
 8   Porc_raca_mae_branca   27 non-null     float64
 9   Porc_raca_mae_negra    27 non-null     float64
 10  Porc_peso_menor_2500   54 non-null     float64
 11  Porc_apgar1_menor_7    54 non-null     float64
 12  Porc_apgar5_menor_7    54 non-null     float64
 13  GINI                   54 non-null     float64
 14  T_DES                  54 non-null     float64
 15  T_ANALF15M             54 non-null     float64
 16  PAREDE                 27 non-null     float64
 17  T_AGUA                 54 non-null     float64
 18  T_BANAGUA              54 non-null     float64
 19  AGUA_ESGOTO            54 non-null     float64
 20  T_LIXO                 54 non-null     float64
dtypes: float64(20), int64(1)
memory usage: 9.0 KB
```

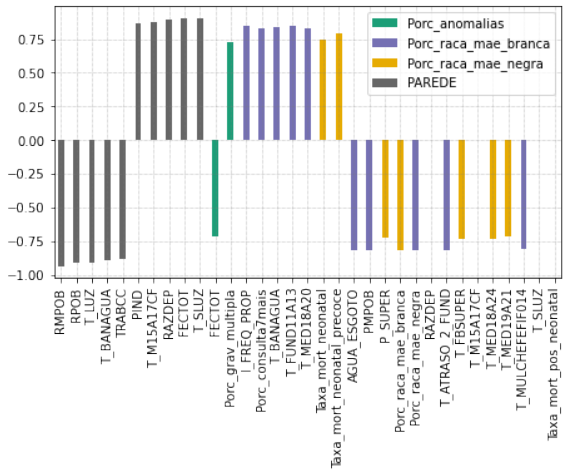
Do conjunto do Quadro 1.2, 17 variáveis estão completas e 4 possuem apenas dados relativos ao ano de 2010. A fim de manter a semântica dos dados e usar um conjunto de variáveis com as do trabalho de [1], tratamos essa situação como um prolema de dados faltantes e fizemos uso de técnicas regressão, como está descrito na próxima subseção.

1.3 Inserção de Dados Faltantes por Regressão

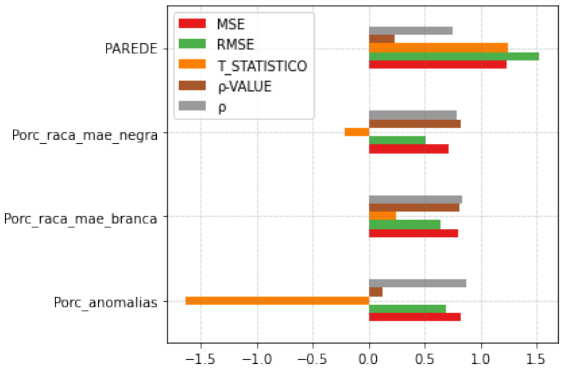
Para utilizar as variáveis Porc_anomalias; Porc_raca_mae_branca; Porc_raca_mae_negra; e PAREDE, que por sua vez não dispunham de dados relativo ao ano de 2000, utilizamos técnicas de regressão, utilizando variáveis disponíveis de 2010 e gerando um modelo que preenche os valores de 2000, com

Figura 1.1: Variáveis Elegíveis para Modelagem e Análise da Qualidade da Inserção

(a) Por Estados



(b) Por Variáveis



um grau de erro passível de mensuração.

1.3.1 Seleção de Variáveis para Regressão

Em busca das as melhores variáveis para compor os modelos preditivos, verificamos a existência de variáveis de forte correlação com as variáveis de predição. Então estabelecemos o filtro para a Correlações de Spearman, considerando apenas valores superiores 0,7 e inferiores a -0,7. O resultado pode ser conferido na Figura 1.1a. Ainda nesta figura, o eixo X lista as barras concernentes às variáveis de melhor correlação. A legenda destaca quais variáveis pertencem qual modelo. O eixo Y lista os valores de correlação de Spearman.

Os valore alcançados com inserção de dados para as variáveis de dados ausentes são apresentados na Figura 1.1b. A variável PAREDE foi de pior ajuste, por ter o maior valor de RMSE. A variável Porc_anomalias foi a de melhor ajuste à modelagem, considerando os valores de ρ-value e T_statistico.

As seguir, apresentamos a lista com a descrição das variáveis que foram utilizadas na regressão, dispostas na Figura 1.1a.

1. AGUA_ESGOTO - Razão entre as pessoas que vivem em domicílios cujo abastecimento de água não provem de rede geral e cujo esgotamento sanitário não é realizado por rede coletora de esgoto ou fossa séptica e a população total residente em domicílios particulares permanentes multiplicado por 100. São considerados apenas os domicílios particulares permanentes.

2. FECTOT - Número médio de filhos que uma mulher deverá ter ao terminar o período reprodutivo (15 a 49 anos de idade).

3. P_SUPER - Razão entre o número de pessoas de 18 anos ou mais de idade ocupadas e que já concluíram a graduação do ensino superior e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.

4. PIND - Proporção dos indivíduos com renda domiciliar per capita igual ou inferior a R$ 70,00 mensais, em reais de agosto de 2010. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.

5. Porc_consulta7mais - Porc. válida de 7 ou mais consultas de pré-natal

6. Porc_grav_multipla - Porc. válida de gestações múltiplas

7. RAZDEP - Razão de dependência é medida pela razão entre o número de pessoas com 14 anos ou menos e de 65 anos ou mais de idade (população dependente) e o número de pessoas com idade de 15 a 64 anos (população potencialmente ativa) multiplicado por 100.

8. RMPOB - Média da renda domiciliar per capita das pessoas com renda domiciliar per capita igual ou inferior a R$ 140,00 mensais, a preços de agosto de 2010. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.

9. RPOB - Média da renda domiciliar per capita das pessoas com renda domiciliar per capita igual ou inferior a R$ 255,00 mensais, a preços de agosto de 2010. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.

10. T_ATRASO_2_FUND - Razão entre o número de pessoas de 6 a 14 anos frequentando o ensino fundamental regular seriado com atraso idade-série de 2 anos ou mais e o número total de pessoas nessa faixa etária frequentando esse nível de ensino multiplicado por 100. O atraso idade-série é calculado pela fórmula: [(idade - 5) - número da série frequentada]. As pessoas de 6 a 14 anos frequentando a pré-escola foram consideradas como se estivessem no 1º ano do ensino fundamental.

11. T_BANAGUA - Razão entre a população que vive em domicílios particulares permanentes com água encanada em pelo menos um de seus cômodos e com banheiro exclusivo e a população total residente em domicílios particulares permanentes multiplicado por 100. A água pode ser proveniente de rede geral, de poço, de nascente ou de reservatório abastecido por água das chuvas ou carro-pipa. Banheiro exclusivo é definido como cômodo que dispõe de chuveiro ou banheira e aparelho sanitário.

12. T_FBSUPER - Razão entre o número total de pessoas de qualquer idade frequentando o ensino superior (graduação, especialização, mestrado ou doutorado) e a população na faixa etária de 18 a 24 anos multiplicado por 100.

13. T_FUND11A13 - Razão entre a população de 11 a 13 anos de idade que frequenta os quatro anos finais do fundamental (do 6º ao 9º ano desse nível de ensino) ou que já concluiu o fundamental e a população total nesta faixa etária multiplicado por 100.

14. T_LUZ - Razão entre a população que vive em domicílios particulares permanentes com iluminação elétrica e a população total residente em domicílios particulares permanentes multiplicado por 100. Considera-se iluminação proveniente ou não de uma rede geral, com ou sem medidor.

15. T_M15A17CF - Razão entre as mulheres de 15 a 17 anos de idade que tiveram filhos e o total de mulheres nesta faixa etária multiplicado por 100.

16. T_MED18A20 - Razão entre a população de 18 a 20 anos de idade que já concluiu o ensino médio em quaisquer de suas modalidades (regular seriado, não seriado, EJA ou supletivo) e o total de pessoas nesta faixa etária multiplicado por 100. As pessoas de 18 a 20 anos frequentando a 4ª série do ensino médio foram consideradas como já tendo concluído esse nível de ensino.

17. T_MED18A24 - Razão entre a população de 18 a 20 anos de idade que já concluiu o ensino médio em quaisquer de suas modalidades (regular seriado, não seriado, EJA ou supletivo) e o total de pessoas nesta faixa etária multiplicado por 100. As pessoas de 18 a 20 anos frequentando a 4ª série do ensino médio foram consideradas como já tendo concluído esse nível de ensino.

18. T_MED19A21 - Razão entre a população de 19 a 21 anos de idade que já concluiu o ensino médio em quaisquer de suas modalidades (regular seriado, não seriado, EJA ou supletivo) e o total de pessoas nesta faixa etária multiplicado por 100. As pessoas de 19 a 21 anos frequentando a 4ª série do ensino médio foram consideradas como já tendo concluído esse nível de ensino.

19. T_MULCHEFEFIF014 - Razão entre o número de mulheres que são responsáveis pelo domicílio, não têm o ensino fundamental completo e têm pelo menos 1 filho de idade inferior a 15 anos morando no domicílio e o número total de mulheres chefes de família multiplicado por 100. São considerados apenas os domicílios particulares permanentes.

20. T_SLUZ - Razão entre as pessoas que vivem em domicílios sem energia elétrica e população total residente em domicílios particulares permanentes multiplicado por 100.

21. Taxa_mort_neonatal - Taxa de mortalidade neonatal (até 27 dias) por 1000 nascidos vivos
22. Taxa_mort_neonatal_precoce - Taxa de mortalidade neonatal precoce (de 0 a 6 dias) por 1000 nascidos vivos
23. Taxa_mort_pos_neonatal - Taxa de mortalidade pós-neonatal (de 28 a 364 dias) por 1000 nascidos vivos
24. TRABCC - Razão entre o número de empregados de 18 anos ou mais de idade com carteira de trabalho assinada e o número total de pessoas ocupadas nessa faixa etária multiplicado por 100.
25. I_FREQ_PROP - Subíndice selecionado para compor o IDHMEducação, representando a frequência de crianças e jovens à escola em séries adequadas à sua idade. É obtido através da média aritmética simples de 4 indicadores:

1.3.2 Avaliação de Regressão

Os dados do ano de 2010, foram separados dois conjunto: uma porção de 33% dos utilizada para teste dos modelos e o restante dos dados, foi utilizado para treino. A qualidade do processo foi avaliado pelas métricas de:

1. Mean Squared Error - MSE: o erro quadrático médio [e]derivado do quadrado da distância euclidiana , é sempre um valor positivo que diminui à medida que o erro se aproxima de zero.
2. Root-Mean-Square Error - RMSE: A raiz quadrada do erro-médio Erro Residual Médio Quadrático: Raiz Quadrada das diferenças dos valores estimados e dos valores reais, elevados ao quadrado.
3. t-statístico: é a razão entre o desvio do valor estimado de um parâmetro de seu valor hipotético e seu erro padrão.
4. ρ-value: é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula.
5. ρ: coeficiente de correlação de Spearman é uma medida estatística da força de uma relação monotônica entre dados emparelhados.

Vemos na Tabela 1.1 o resultado da comparação da regressão. A variável Perc.Anomalias teve melhores valores de MSE e RMSE para Regresseão Linear - RegLin, visto em fonte escura na tabela citada. A Regressão Lasso teve melhor valor para T_STATISTICO e menor valor de ρ-value, apesar de não ser abaixo de 0,05, que asseguraria um nível de significância de 5% e um intervalo

de confiância de 95%. Por fim, os valores de correlação de Spearman (ρ) ficaram todos bastante próximos, não ultrapassando 0,11 de diferença entre o menor(0,76) da Regressão Linear e o maior valor (0,87), atingido pela Regressão Lasso, Lasso Cross Validation, entre as outras de mesmo resultado. A variável Perc. Mãe Negra teve melhores valores de MSE e RMSE para Regresseão Bayesian Ridge, visto em fonte escura na tabela citada. A Regressão Automatic Relevance Determination - ARDReg teve melhor valor para T_STATISTICO. a Regressão Linear teve e menor valor de ρ-value. Por fim, os valores de correlação de Spearman (ρ) tiveram amplitude de 0,66 a 0,78, também ficando todos bastante próximos.

Tabela 1.1: Avaliação Geral da Inserção de Dados Faltantes

Método	**Variável**	**MSE**	**RMSE**	**T_STATISTICO**	**ρ-VALUE**	ρ
RegLin	Perc.Anomalias	**0.75**	**0.56**	-1.30	0.21	0.76
Lasso	Perc.Anomalias	0.87	0.76	**-1.77**	**0.10**	0.87
LassoCV	Perc.Anomalias	0.84	0.70	-1.67	0.11	0.87
LassoLars	Perc.Anomalias	0.80	0.63	-1.52	0.15	0.85
ElasticNet	Perc.Anomalias	1.20	1.45	-2.61	0.02	0.87
ElasticNetCV	Perc.Anomalias	0.85	0.72	-1.70	0.11	0.87
BayesianRidge	Perc.Anomalias	0.80	0.64	-1.49	0.16	0.85
ARDReg	Perc.Anomalias	0.83	0.69	-1.64	0.12	0.87
RegLin	Perc. Mãe Negra	0.89	0.80	0.37	**0.71**	**0.66**
Lasso	Perc. Mãe Negra	0.72	0.51	-0.13	0.90	0.78
LassoCV	Perc. Mãe Negra	0.82	0.67	0.09	0.93	0.78
LassoLars	Perc. Mãe Negra	0.71	0.51	0.04	0.97	**0.65**
ElasticNet	Perc. Mãe Negra	0.98	0.96	0.34	0.74	-
ElasticNetCV	Perc. Mãe Negra	0.75	0.56	0.02	0.98	0.77
BayesianRidge	Perc. Mãe Negra	**0.67**	**0.45**	-0.08	0.93	0.77
ARDReg	Perc. Mãe Negra	0.71	0.51	**-0.22**	0.83	0.78
RegLin	Perc. Mãe Branca	1.20	1.44	0.57	0.57	0.80
Lasso	Perc. Mãe Branca	**0.67**	**0.45**	-0.10	0.92	0.77
LassoCV	Perc. Mãe Branca	0.73	0.53	-0.03	0.98	0.78
LassoLars	Perc. Mãe Branca	0.79	0.63	0.21	0.84	0.78
ElasticNet	Perc. Mãe Branca	0.76	0.58	**-0.74**	**0.47**	0.77
ElasticNetCV	Perc. Mãe Branca	0.73	0.53	-0.01	0.99	0.78
BayesianRidge	Perc. Mãe Branca	0.76	0.57	0.24	0.82	0.77
ARDReg	Perc. Mãe Branca	0.80	0.64	0.25	0.81	0.83
ElasticNet	PAREDE	1.71	2.93	2.20	**0.04**	-
RegLin	PAREDE	**1.06**	**1.12**	**0.98**	0.33	0.83

Lasso	PAREDE	1.44	2.07	1.71	0.11	0.83
LassoCV	PAREDE	1.12	1.25	1.08	0.30	0.80
LassoLars	PAREDE	1.27	1.61	1.36	0.19	0.77
ElasticNetCV	PAREDE	1.12	1.26	1.09	0.29	0.80
BayesianRidge	PAREDE	1.28	1.65	1.35	0.20	0.77
ARDReg	PAREDE	1.24	1.53	1.25	0.23	0.75

Fonte: elaborado pelo autor

A variável Perc. Mãe Branca teve melhores valores de MSE e RMSE para a Regressão Lasso, visto em fonte escura. A Regressão ElasticNet teve melhor valor para T_STATISTICO e menor valor de ρ-value. Por fim, os valores de correlação de Spearman (ρ) tiveram amplitude de 0,77 a 0,80. Finalmente, a variável PAREDE teve melhores valores de MSE e RMSE para a Regressão Linear, visto em fonte escura. A Regressão Linear tave ainda o menor valor para T_STATISTICO. A regressão ElasticNet obteve o melhor valor de ρ-value, este, abaixo de 0,05, ficando com intervalo de confiança de 95%. Por fim, os valores de correlação de Spearman (ρ) tiveram amplitude de 0,75 a 0,83.

Inserimos os dados faltantes utilizando Regressão Linear, todavia, temos clareza das vantagens e desvantagens em relação aos outros métodos.

1.4 Medidas de Posição e Dispersão Por Variáveis e Estados

Na Figura 1.2a, apresentamos a distribuição das variáveis. Dentre as variáveis obstétricas a de maior amplitude Porc_raca_mae_branca, como pode ser visto no gráfico, na nona posição, colorido em cinza. Ainda na figura citada, na décima e décima primeira posição, as variáveis Porc_apgar1_menor_7 e Porc_apgar5_menor_7, apresentam intervalo interquantil e medianas bastante próximos.

As medianas de Porc_prematuros, Porc_cesareas e Porc_grav_multipla também aparem coincidentes, nas posições segunda, terceira e quarta da Figura 1.2a.

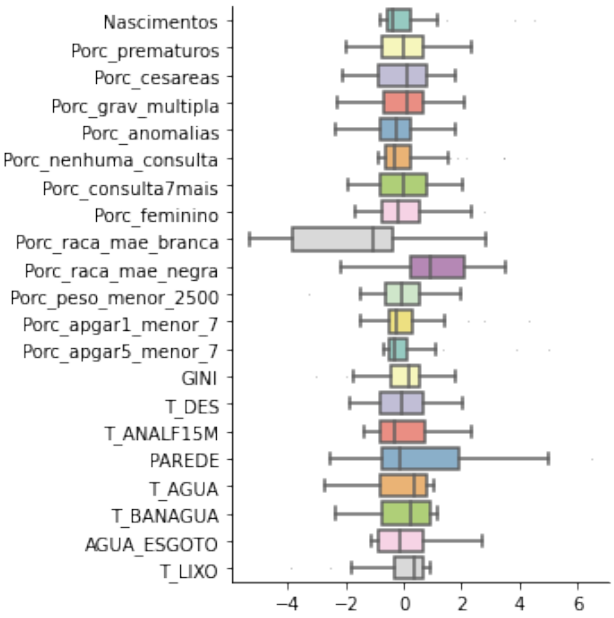
Das variáveis socioeconômicas, PAREDE é a de maior amplitude, seguida de T_ANALF15M, AGUA_ESGOTO e T_DES, vistas na parte inferior das Figura 1.2a, nas cores azul, vermelha, rosa e lilás, respectivamente.

Na Figura 1.2b, os valores dos indicadores por ano dos estados foram tratados como uma mesma distribuição, para efeito de comparação das medidas. As maiores amplitudes podem ser observadas nos registros dos anos 2000, para os estados do Acre e Maranhão. Ainda na figura citada, nota-se que estado de São Paulo aumentou o limite superior e reduziu o limite inferior da sua distribuição do ano 2000 para o ano 2010. Uma situação contrária é vista para o Estado de Tocantins, abaixo de São Paulo no fim do gráfico.

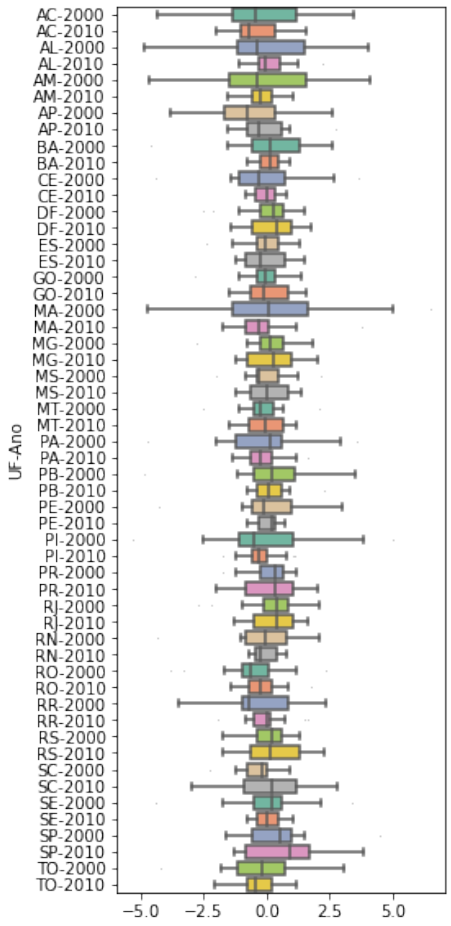
A seguir, aprofundamos a observação por estado, correlacionando seus indicadores. Apresentamos também, as correlações entre as variáveis. De modo a ter um panorama de quais estados e variáveis tendem a ser similares.

Figura 1.2: Medidas por Variável e Por Estado

(a)



(b)



1.5 Correlações Identificadas nos Dados Completos

Após a inserção de dados faltantes, foi realizada o calculo da correlação de Spearman, a fim de evidenciar quais variáveis dispunham de forte correlação. Neste ponto da investigação, em função de haver dois registros de indicadores por Estado, foi calculada a média simples de cada variável. Assim a correlação foi tomada com base nas médias, de modo a agregar os dados temporais. Na Figura 1.3a, os valores de correlação positiva estão postos na coloração do gradiente azul, sendo mais escuro quanto maior o valor de correção. E as correlações negativas, estão dispostas na coloração do gradiente vermelho. De mesmo modo, tem-se que a cor mais sendo ocorre conforme a magnitude da métrica.

Na Figura 1.3a, vê-se a forte correlação entre os estados de SP, MG, RS, PR, RJ, SC, DF, ES, variando na amplitude positiva de 0,67 a 0,87. No espectro das correlações negativas, alguns estados da Região Nordeste e Norte forte associação inversa aos estados de SC, RS, e SP, por exemplo. A amplitude neste caso, foi de -0,67 a -0,81.

Na Figura 1.3b, estão postas, com a mesma lógica de coloração, as correlações das variáveis, também calculada pelas médias nos períodos apurados. Nela, dentre as demais correlações, destacamos as seguintes:

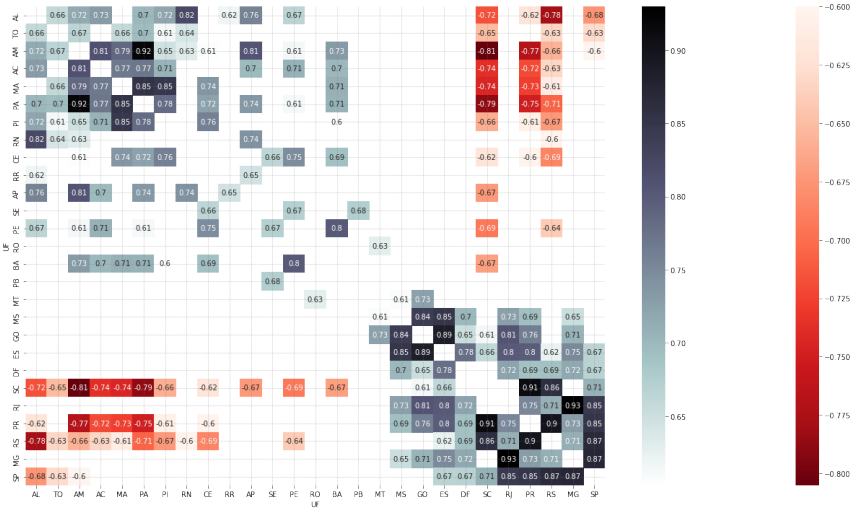
1. Porc_raca_mae_branca(+): T_AGUA, T_BANAGUA E T_LIXO.
2. Porc_raca_mae_branca(-): T_ANALFABETISMO15M, GINI, PAREDE e AGUA_ESGOTO.
3. Porc_raca_mae_negra(+): Porc_nenhuma_consulta, GINI, T_ANALFABETISMO15M e AGUA_ESGOTO.
4. Porc_raca_mae_negra(-): T_AGUA e T_BANAGUA.
5. Porc_prematuros(+): Porc_anomalia e Porc_grav_multipla
6. Porc_cesareas(+): Porc_consulta7mais, Porc_mae_branca, T_AGUA, T_BANAGUA E T_LIXO
7. Porc_cesareas(-): Porc_nenhuma_consulta, GINI, T_ANALFABETISMO15M, PAREDE e AGUA_ESGOTO.

Para compreender a estrutura dos dados, seguindo as práticas atuais da literatura [2], foi feito uso da técnica de análise de agrupamento, variando os critérios de seleção de características, de modo a reduzir a dimensionalidade do conjunto.

Na sequência, foram discutidas as peculiaridades de cada arranjo identificado, relatando as variáveis de preponderância discriminativa.

Figura 1.3: Correlações de Spearman - Conjunto de Dados Completo

(a) Por Estados



(b) Por Variáveis

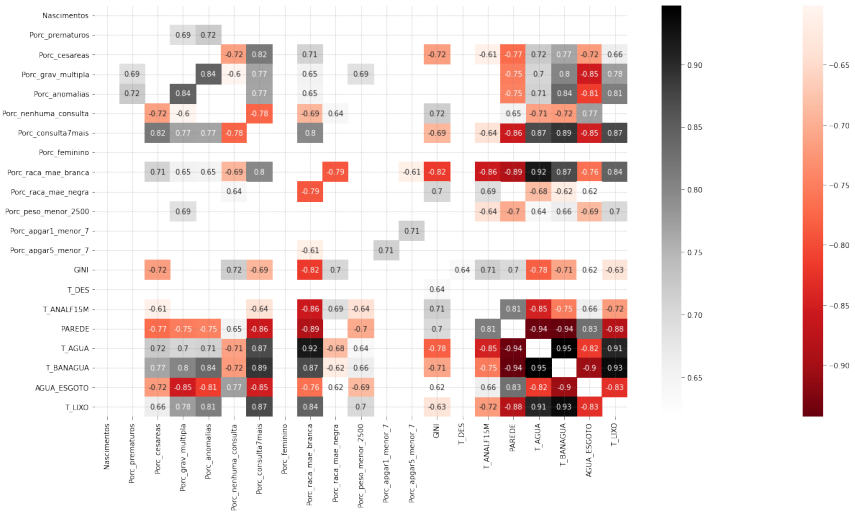
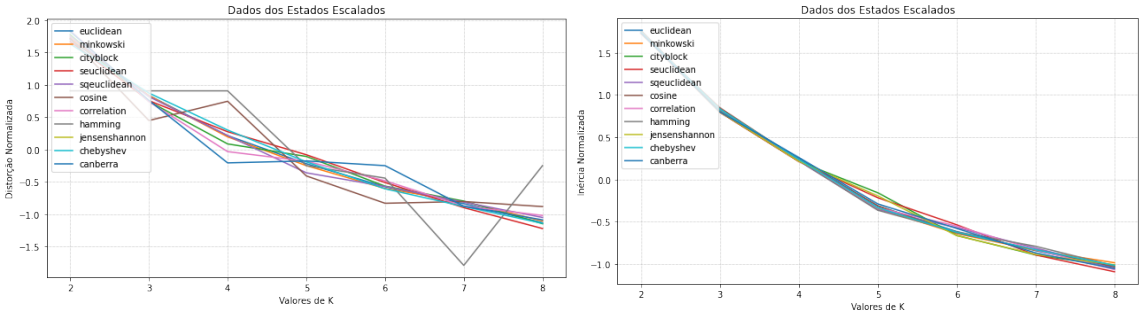


Figura 2.1: Avaliação de Clusters Sobre os Dados Reais Padronizados

(a) (b)



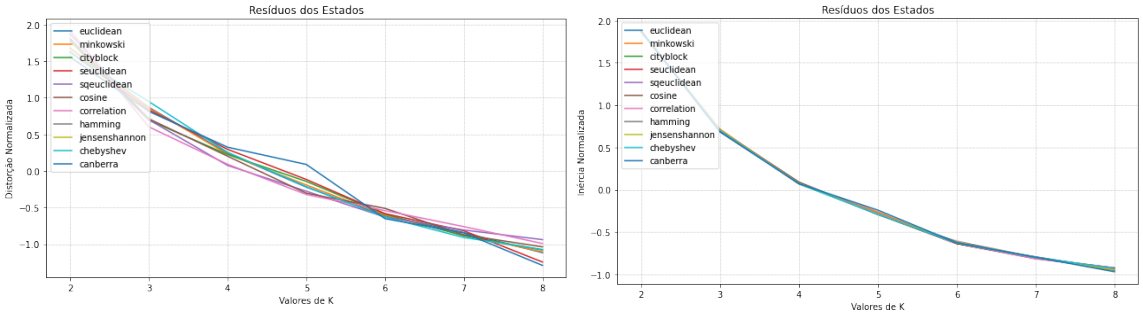
2 Análise de Agrupamento

A análise de agrupamento foi feita com absolutos escalados e com os dados residuais, oriundos da subtração dos valores dos indicadores de 2000 dos valores de 2010. A Figura 2.1a mostra para o conjunto de distâncias, o valor da pontuação de silhueta calculado para cada quantidade de grupos avaliados. Naquela, o eixo X representa a quantidade de grupos e o eixo Y determina o valor de silhueta, os quais foram escalados para uma distribuição normal padronizada. A Figura 2.1b mostra a Pontuação de Inércia também por quantidade de clusters.

As Figuras 2.2a e 2.2b apresentam a mesma avaliação citada, mas processando os dados residuais. A quantidade ótima de *clusters* para os dados reais foi de 6 grupos e para os dados residuais, foi de 7 grupos. A seguir, mostramos características dos agrupamentos, bem como variáveis fortemente relacionados por grupo.

Figura 2.2: Avaliação de Clusters Sobre os Dados Residuais

(a) (b)



2.1 Agrupamentos Direto e Indireto

O agrupamento executado diretamente nos valores padronizados pode ser visto na Figura 2.3a. Nela, os clusters estão identificadas pelo eixo X. No eixo Y, a faixa de valores das distribuições dos clusters. O cluster 1 é o de maior amplitude e o cluster 2 é o de menor. As medianas dos clusters 3 e 5 são bem próximas, contudo, o intervalo interquantil do cluster 3 é maior que o dobro do intervalo interquantil do cluster 5. Igualmente, o limite superior do cluster 3 é maior que o limite superior do cluster 5.

O agrupamento executado sobre as métricas da diferença dos indicadores por estado, o qual denominamos, agrupamento indireto, está posto na Figura 2.3b. Uma vez que trata de dados residuais, os valores estão predominantemente distantes de 0, pelo fato dos indicadores do estados terem variado entre os anos observados. As medianas dos clusters 1, 2, 4 e 6 são bem próximas. O cluster 4 é o de menor amplitude, quase imperceptível na figura. Os clusters 4 e 6 possuem amplitude ajustada ao intervalo interquantil.

A seguir, buscamos apresentar uma arrazoado sobre quais variáveis são determinantes para que o estado seja classificado nos clusters. Para isso, fizemos uso da técnica de árvore de decisão, utilizando os valores das variáveis e os clusters identificados no agrupamento.

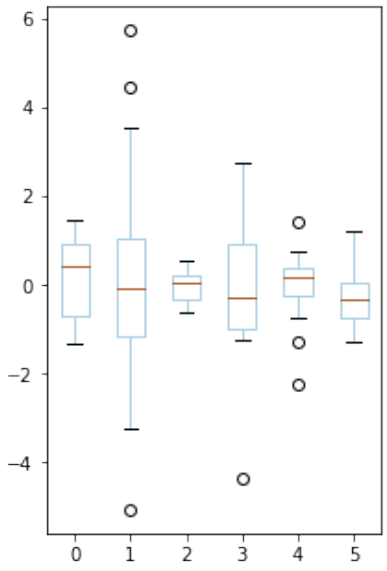
2.2 Importâncias das Variáveis para do Agrupamento

Com vistas a compreender quais variáveis foram determinantes para a formação dos grupos, utilizamos a técnica de árvore de decisão. Na Figura 2.4a, estão em 5 níveis, para os 6 grupos classificados. A variável *Porc_conlusta_7mais* é a de maior importância para a separação dos grupos. Seguida de PAREDE. Em paralelo, tem-se na sequencia, Porc_apgar_menor_7 e Porc_raca_mae_branca. A seguir, novamente Porc_raca_mae_branca e por fim, em paralelo, PAREDE E Porc_consulta_7mais.

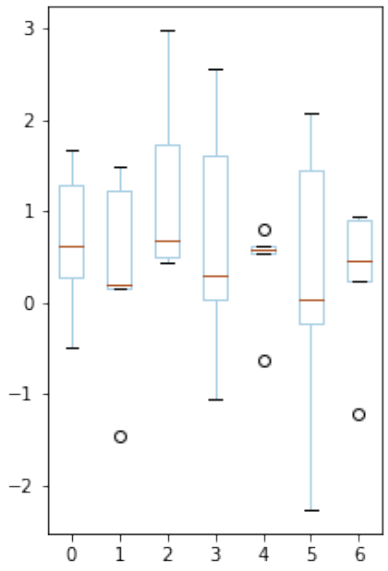
No agrupamento pelas métricas, visto na Figura 2.4b o MSE se destava no primeiro nível. No segundo, reaparece junto do T_STATISTICO. No teceiro nível, tem-se a ρ e T_STATISTICO. No quarto nível, apenas o T_STATISTICO.

Figura 2.3: Medidas de Posição e Dispersão pelas Médias dos Clusters

(a) Clusters Por Variáveis



(b) Clusters Por Métricas de Resíduos



2.3 Descrição dos Agrupamentos

Na Figura 2.5a, apresentamos por estado, os valores das métricas residuais. O estado do Maranhão apresenta pico de MSE, visto na linha azul tracejada. Os estados Maranhão e Acre integram o cluster 1. Corroborando com a magnitude do erro encontrado neste cluster, a Figura 2.3a evidenciava variação atípica, pela grande amplitude vista na distribuição estatística.

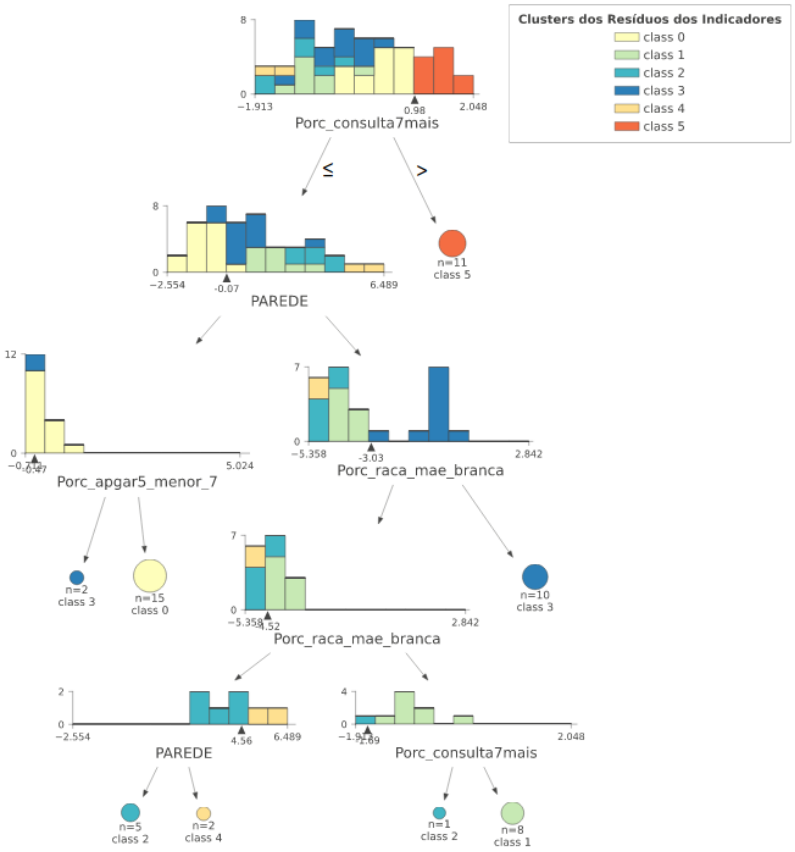
Na Figura 2.5a, os valores de correlação negativo, demonstram a variação grande entre os anos observados, como ocorreu nos estados de Pernambuco, Rio Grande do Norte e Ceará. Para compreender essa variação, a Figura 2.5b lista a transição dos estados por cluster, nos anos observados. Em azul escuro a posição do estado em 2010, e em laranja, a posição do estado no ano 2000.

Os estados do Maranhão e Piaui, nos anos 2000, ocupavam sozinhos o cluster 4. No ano de 2010, progrediram para o cluster 1, junto de Acre, Amazonas, Amapá e Pará.

O Cluster 3 apresenta exclusivamente valores relativos ao ano 2000, sendo composto por estados da região norte e nordeste. De mesmo modo, o Cluster 0 apresenta predominantemente dados dos anos 2000, excetuando-se o estado da Bahia que em 2010 compos o cluster sozinho. Ainda sobre o cluster 0, os estados que o compunham em 2000 progrediram para o cluster 2, majoritariamente, e cluster 5, que por sua vez reúne 13 dos 27 estados. Interno ao Cluster 5, a Figura 2.7a lista a disposição dos valores de suas variáveis. Comparado ao Cluster 0, na Figura 2.7b, nota-se uma taxa de Porc_prematuros, de Porc_anomalias, por exemplo. Todavia, as variáveis Porc_apgar1_menor_7 e Porc_apgar5_menor_7 na mesma comparação, aparecem com intervalos interquantis maiores, bem como suas amplitudes.

Figura 2.4: Árvores de Decisão dos Clusters

(a) Perspectiva sobre os Dados Diretos



(b) Perspectiva sobre os Resíduos dos Dados

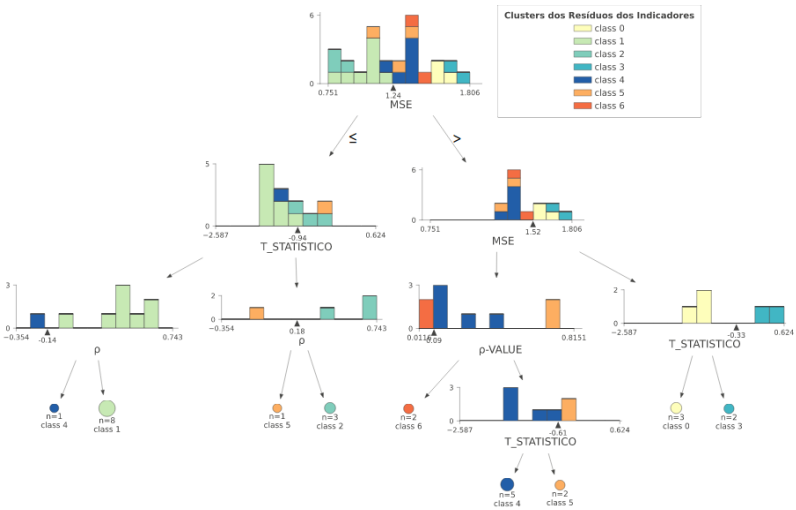


Figura 2.5: Clusterização por Resíduos e por Valores

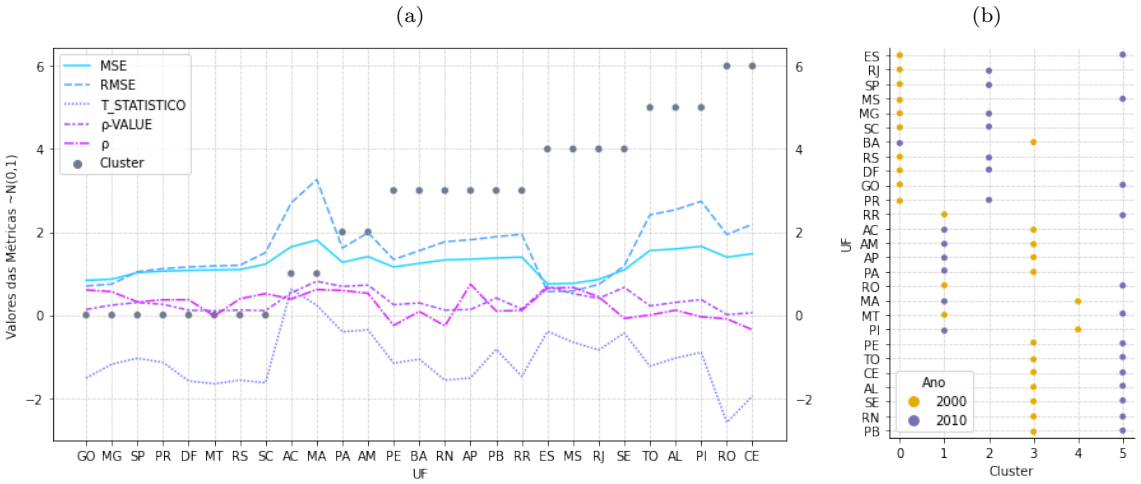


Figura 2.6: Agrupamento Indireto - Correlação Métricas de Avaliação dos Dados Residuais

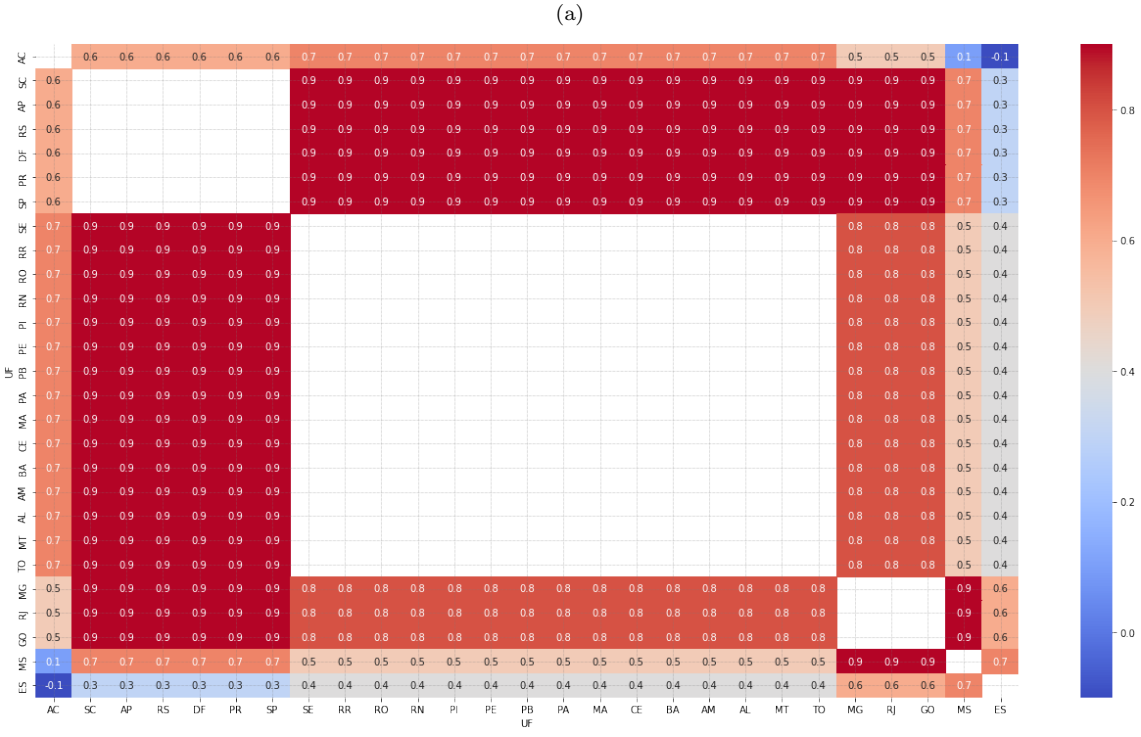
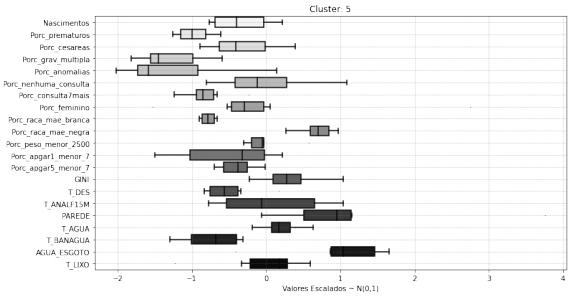


Figura 2.7: Medidas de Posição e Dispersão - Contrastes dos Clusters

(a) Cluster Moderno



(b) Cluster Antigo

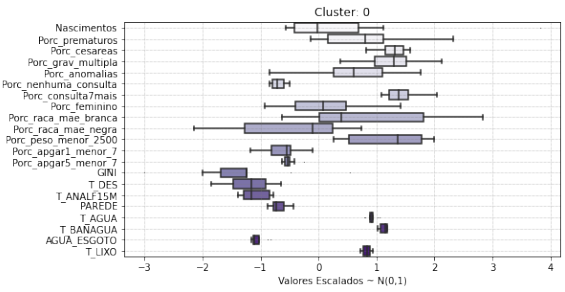
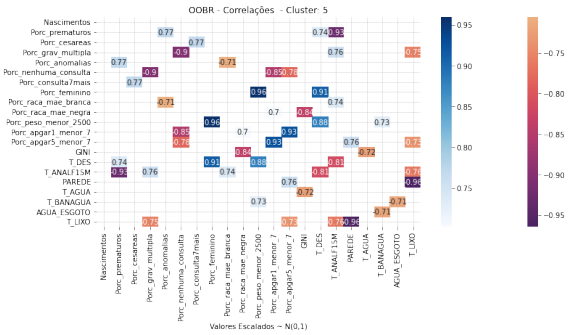
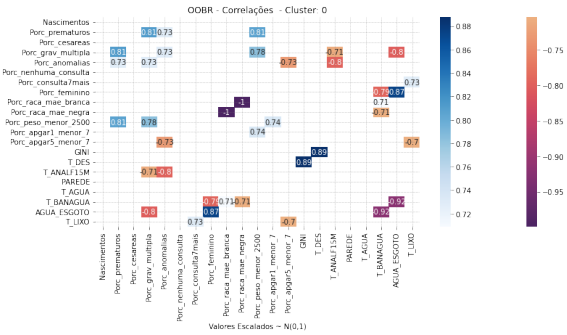


Figura 2.8: Medidas de Posição e Dispersão - Constrastes dos Clusters

(a) Cluster Moderno



(b) Cluster Antigo



Observando as variáveis fortemente relacionadas nos clusters, a Figura 2.8a mostra o Cluster 5 evidencia correlação de 0,96 entre Porc_feminino e Porc_peso_menor_2500. Das variáveis sócioeconomicas, a T_DES aparece correlacionada com Porc_feminino em 0,91. Neste cluster, verifica-se também a correlação entre Porc_apgar1_menor_7 e Porc_apgar5_menor_7 da ordem de 0,93. A variável T_ANALF15M está correlacionada com Porc_grav_multipla em 0,76. A variável Porc_peso_menor_2500 está correlacionada a T_DES no valor de 0,88. No Cluster 5, inversamente correlacionada estão Porc_prematuros e T_ANALF15M em -0,93. Igualmente Porc_grav_multipla está associada a T_LIXO em -0,79. A variável Porc_raca_mae_negra possui correlação com a variável GINI no valor de -0,84.

Na Figura 2.8b, a maior correlação positiva entre variável socioeconômica e obstétrica, ocorre entre e Porc_feminino e AGUA_ESGOTO, no valor de 0,87. A segunda segunda maior correlação, ocorre entre Porc_consulta7mais e T_LIXO. Na mesma Figura, inversamente correlacionadas, tem-se Porc_grav_multipla e AGUA_ESGOTO, no valor de -0,8. As variáveis Porc_feminino e T_BANAGUA possuem correlação de -0,79.

Verificadas as correlações das variáveis por agrupamento e o posicionamento dos estados nos grupos, passamos ao rankeamento dos estados. A intenção é verificar níveis de desenvolvimento similar e ocorrência de quadros regionais similares.

2.4 Agrupamentos GeoSES e Indicadores Obstétricos

Figura 2.9: Correlações GeoSES e Indicadores Obstétricos

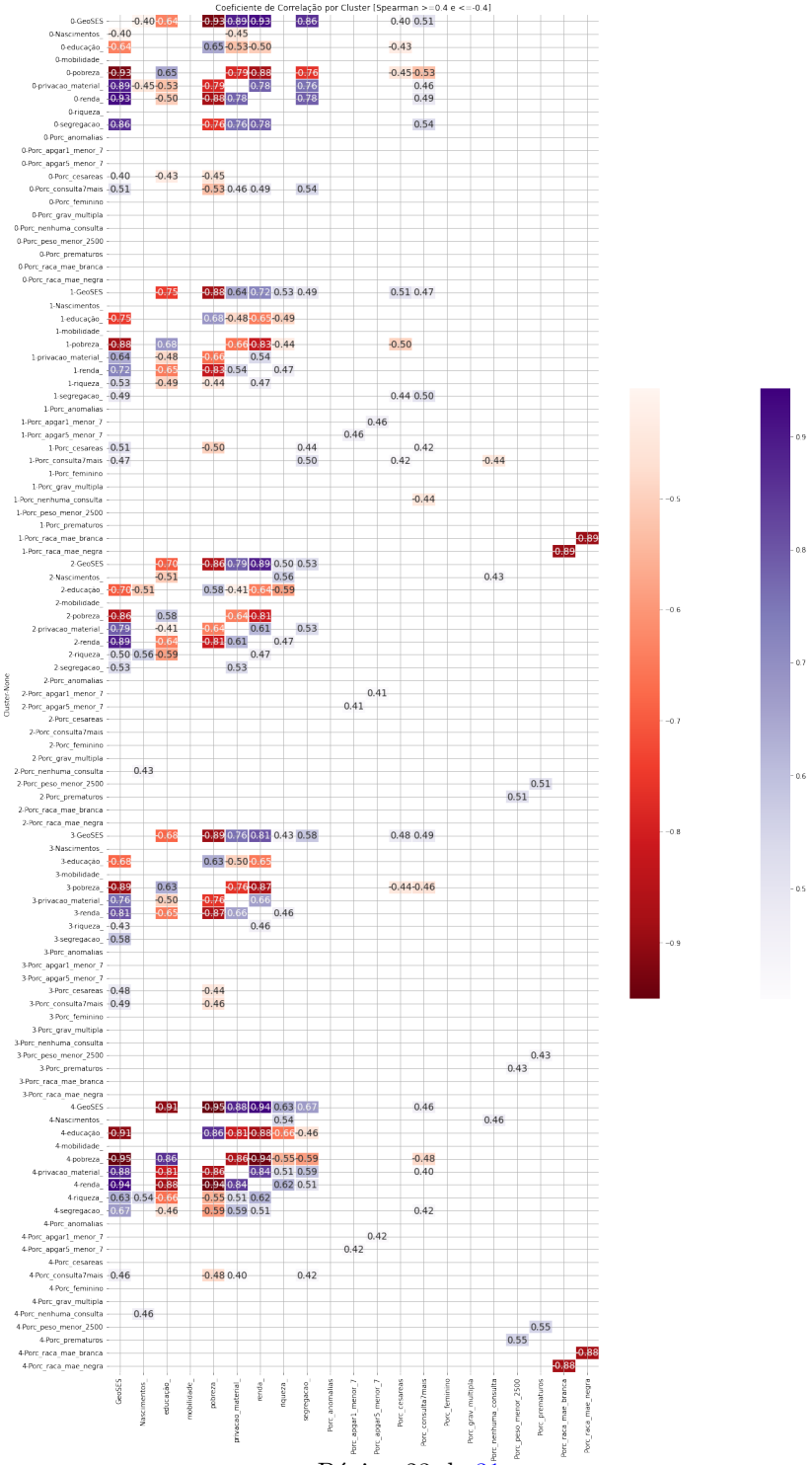
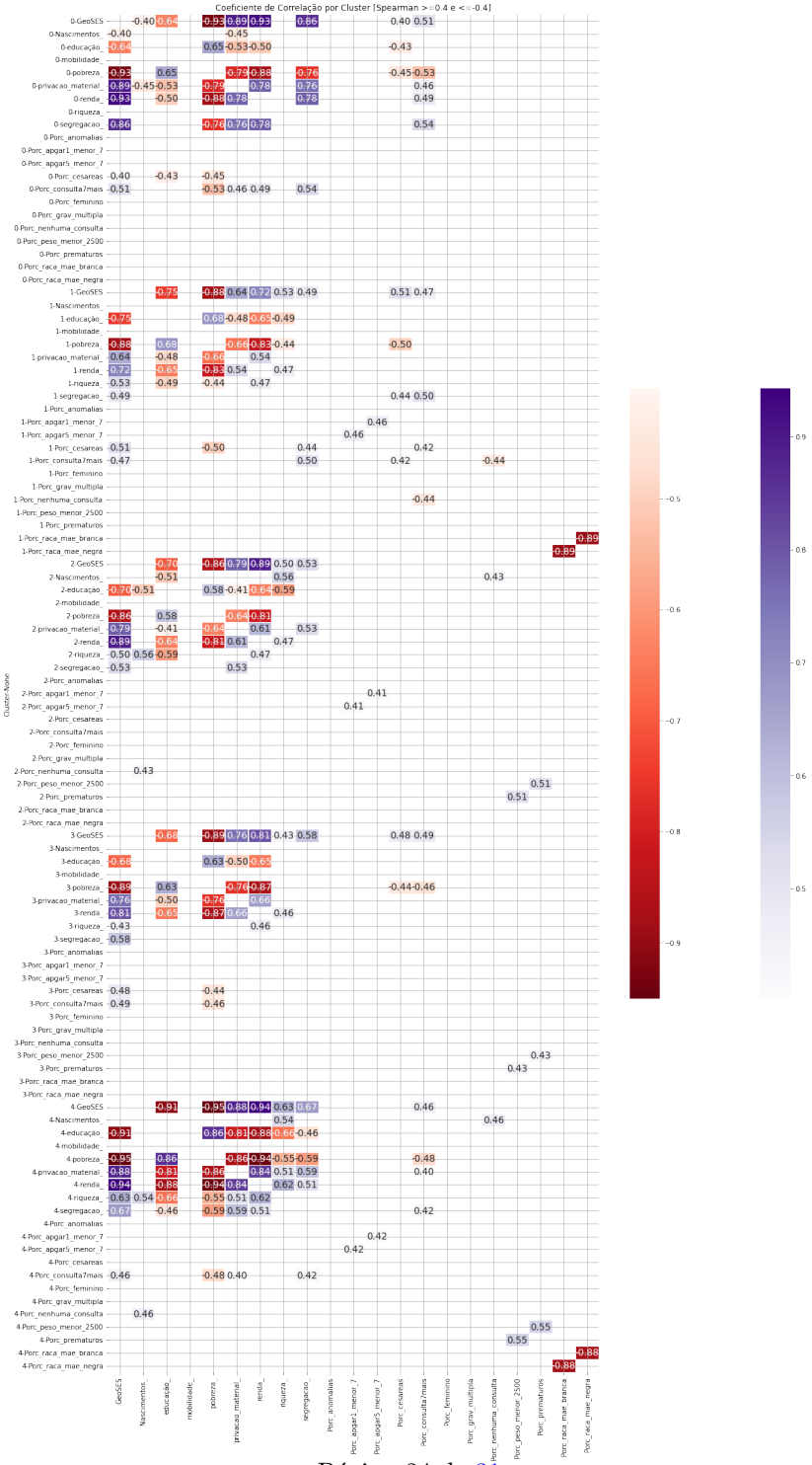


Figura 2.10: Correlações GeoSES e Indicadores Obstétricos



3 Ranking dos Estados e Mapas de Indicadores

3.1 Ranking Geral por Resíduos

A Figura 3.1a ilustra nossa proposta de ranking dos estados por análise de resíduo. A Figura, cujos dados estão ordenados pelo MSE, mostra em sua primeira posição, com menor valor residual o estado do Espírito Santo; enquanto em última posição figura o estado do Maranhão. Na mesma figura apresentamos com maior clareza a correlação negativa do estado do Ceará, Pernambuco, Rio Grande do Norte, Rondônia, Sergipe e Mato Grosso. O que implica ter havido grande variação em seus indicadores. Os estados de correlação positiva não passaram pro grandes variações em seus indicadores, mas a diferenças absolutas de suas métricas, ficam evidenciados pelo RMSE. Portanto, a classificação de estados pode ter mais de uma ordenação, de acordo com o destaque desejado.

De modo a compreender as métricas geradas pela análise de resíduo, a Figura 3.1b mostra as distribuições, por meio de suas curvas de densidade. Neste gráfico, destacamos a tendencia da métrica ρ em formar uma bimodal. O que era esperado em virtude daqueles estados em que a correlação de seus indicadores foi negativa.

3.2 Ranking de Contrastes

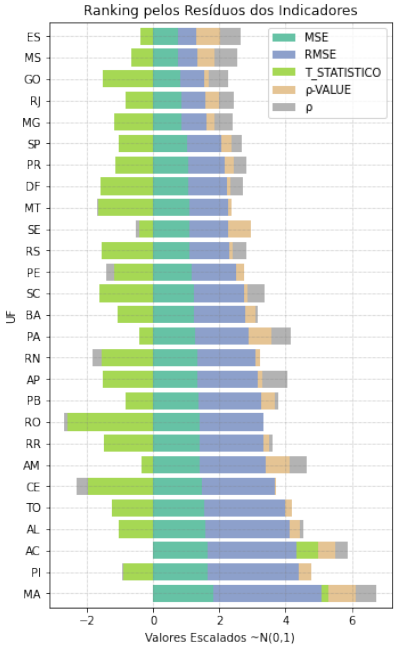
A fim de verificar os valores padronizados, para os estados contrastantes no Ranking Geral, na Figura 3.2 apresentamos as variáveis por estado, relativos aos anos apurados. Nele, vemos que São Paulo tende a ter as variáveis predominantemente acima da média, visto nas cores verde claro e lilás. O estado do Maranhão tende a ter indicadores inferiores à média. O Espírito Santo oscila intermediariamente em torno da média.

3.3 Mapas

Por fim, ilustrando no mapa brasileiro os valores encontrados por grupos e anos, as variáveis de maior ocorrência nas correlações, temos as Figuras 3.3a, 3.3b, 3.4a, 3.4b, 3.5a, 3.5b, 3.6a, 3.6b, 3.7a, e 3.7b.

Figura 3.1: Ranking e Distribuições das Métricas de Avaliação

(a)



(b)

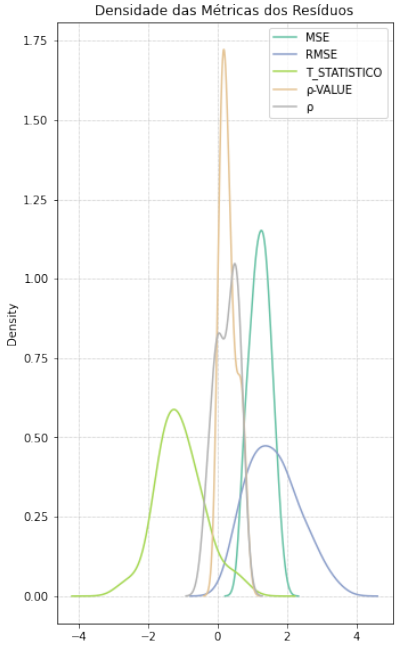


Figura 3.2: Contraste das Variáveis por Estados por Ano

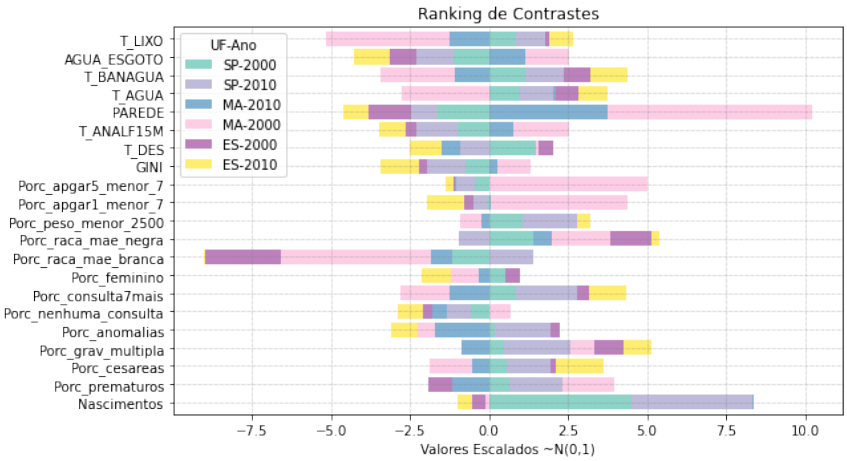


Figura 3.3: Mapas Brasileiros a Variável Padronizada

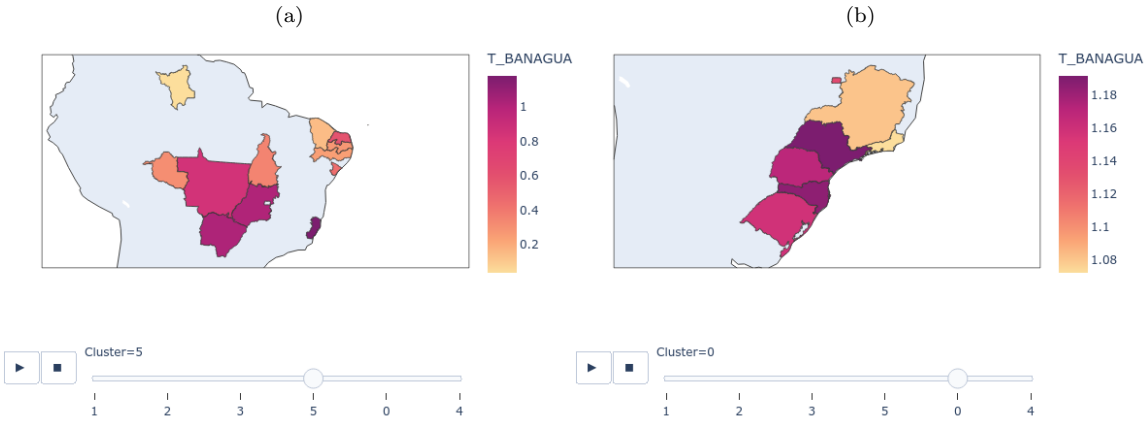


Figura 3.4: Mapas Brasileiros a Variável Padronizada

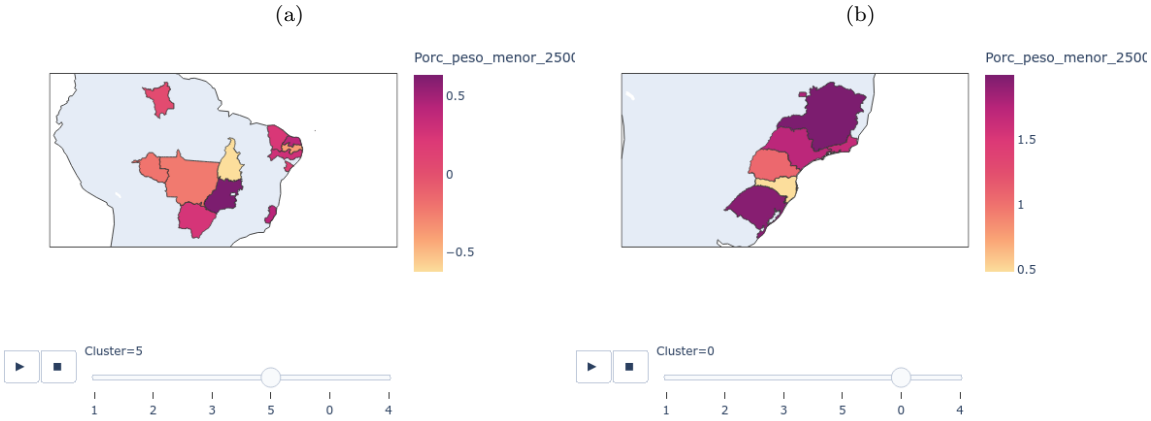


Figura 3.5: Mapas Brasileiros - Constrastes Porc_apgar5_menor_7

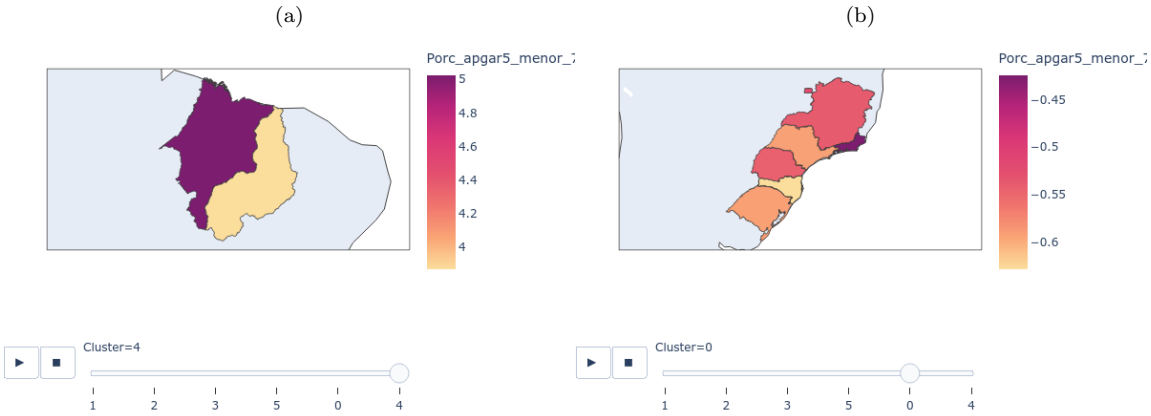


Figura 3.6

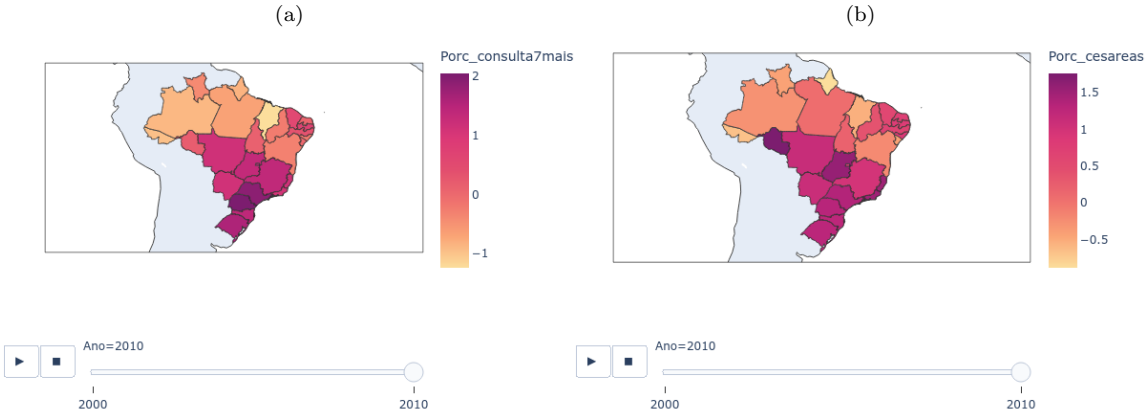
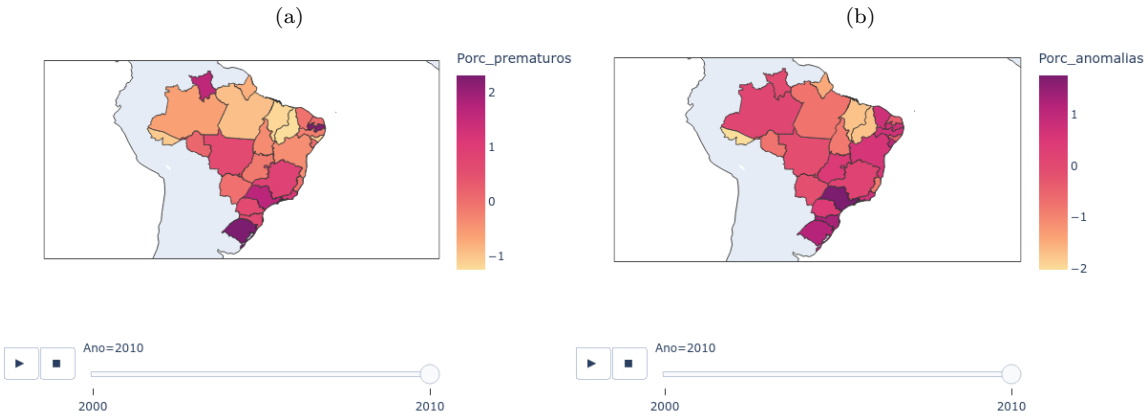


Figura 3.7: Mapas Brasileiros a Variável Padronizada



Bibliografia

[1] Thayane Santos Siqueira, José Rodrigo Santos Silva, Mariana do Rosário Souza, Débora Cristina Fontes Leite, Thomas Edwards, Paulo Ricardo Martins-Filho, Ricardo Queiroz Gurgel, and Victor Santana Santos. Spatial clusters, social determinants of health and risk of maternal mortality by COVID-19 in Brazil: a national population-based ecological study. *The Lancet Regional Health - Americas*, 3:100076, 2021.

[2] Marcos A. Spalenza, Juliana P. C. Pirovani, and Elias de Oliveira. Structures discovering for optimizing external clustering validation metrics. In Ajith Abraham, Patrick Siarry, Kun Ma, and Arturas Kaklauskas, editors, *Intelligent Systems Design and Applications - 19th International Conference on Intelligent Systems Design and Applications (ISDA 2019), Auburn, WA, USA, December 3-5, 2019*, volume 1181 of *Advances in Intelligent Systems and Computing*, pages 150–161. Springer, 2019.

Siglas

OOBr Observatório Obstétrico Brasileiro 2, 3

SINASC Sistema de Informação sobre Nascidos Vivos 2