Aprendizagem Não-Supervisionada Aplicada à Estratificação de Risco de Nascimentos Prematuros no Brasil com Recorte em Dados Socioeconômicos

Márcio L. B. Lopes Jr.[*], Raquel de M. Barbosa[*†] e Marcelo A. C. Fernandes[*‡]
[*]Laboratório de Aprendizagem de Máquina e Instrumentação Inteligente, nPITI/IMD, UFRN, Natal, RN, Brasil.
[†]Laboratório de Desenvolvimento de Medicamentos, DFAR, UFRN, Natal, RN, Brasil.
[‡]Departamento de Engenharia da Computação e Automação, UFRN, Natal, RN, Brasil.
Email: [*]marcio.lopes.099@ufrn.edu.br,[†]m.g.barbosafernandes@gmail.com,[‡]mfernandes@dca.ufrn.br

***Resumo*—O nascimento prematuro (*Preterm birth* - PTB) é um fenômeno que traz diversos riscos e desafios à sobrevivência dos recém-nascidos. Apesar de muitos avanços, ainda não foram esclarecidas todas as causas desse fenômeno. Entende-se que o risco ao PTB é multi-fatorial e também pode estar associado a fatores socioeconômicos. Assim, este artigo tem como meta a utilização de técnicas de aprendizagem não-supervisionada para obter um estrato de risco de PTB no Brasil a partir de dados socioeconômicos. Através de utilização de bancos de dados públicos disponibilizados pelo Governo Federal do Brasil, foi gerado um *dataset* a nível municipal com informações socioeconômicas e uma taxa de ocorrência de prematuridade. Este *dataset* foi clusterizado utilizando várias abordagens e técnicas de aprendizagem de máquina não-supervisionada como k-means, *principal component analysis* (PCA) e *density-based spatial clustering of applications with noise* (DBSCAN) . Após a validação, foram descobertos 4 *clusters* com ocorrências de PTB muito acima da média nacional, e 3 muito abaixo. Os *clusters* de alta prematuridade apontaram para municípios de menor nível educacional, de pior qualidade de serviços públicos, como saneamento e coleta de lixo, e de população menos branca. Foi observado ainda o aspecto regional, com os municípios em *clusters* de maior risco ao PTB estando localizados principalmente nas regiões Norte e Nordeste. Os resultados apontam para uma influência positiva da qualidade de vida e da oferta de serviços públicos na diminuição da ocorrência de nascimentos prematuros.**

***Palavras-chave*—Parto prematuro, Clusterização, Aprendizagem Não-Supervisionada, SINASC, Cadastro Único**

I. Introdução

O nascimento prematuro (*Preterm birth* - PTB), definido como aquele que ocorre antes das 37 semanas de gestação, é a maior causa de mortalidade no mundo para crianças com até 5 anos de vida [1]–[3]. Adicionalmente, foi demonstrado como sendo um fator crítico da sobrevivência de recém-nascidos [2]. Os nascidos prematuros apresentam um maior desafio à assistência médica, que precisa suprir o ainda incompleto desenvolvimento de alguns órgãos vitais [4]. Tentar entender e prever as causas do PTB tem sido cada vez mais frequente em pesquisas científicas, especialmente com o surgimento de bancos de dados governamentais progressivamente mais confiáveis e complexos. Aproximar-se desse objetivo significaria encontrar formas de prevenir casos de PTB, ou de antecipar a assistência às mães quando a prevenção não for possível, reduzindo assim o número de vidas perdidas.

O trabalho apresentado em [5] mostra que a etiologia do PTB é multi-fatorial e o fator de risco pode estar associado a situação socioeconômica de uma dada região (*neighbourhood socioeconomic status* (*neighbourhood* SES). A região de SES (*neighbourhood* SES) é uma medida em nível de área que agrega fatores de SES (como renda, educação e situação de emprego) em um determinado nível geográfico [6]. Trabalhos na literatura mostram que a taxa de PTB em áreas de baixo SES é maior que a taxa de PTB em áreas de alto SES [7].

Várias técnicas de aprendizagem de máquina foram previamente aplicadas ao problema para predição ou estratificação do risco PTB, incluindo SVMs [8], redes neurais [9]–[11] e árvores de decisão [12], [13]. No entanto, as aplicações mais comuns são as técnicas de regressão logística e de regressão linear, empregadas na análise e na predição de PTB para diferentes fatores: a pobreza [14], as condições de trabalho da gestante [15], [16], fatores sociais em geral [17], [18] e, principalmente, fatores clínicos ou hereditários [19]–[22]. Há ainda uma vasta literatura associando diferentes fatores com a prematuridade utilizando métodos estatísticos tradicionais [23]–[25], entre os quais fatores socioeconômicos [26], [27].

Uma das formas de compreender a associação entre os diversos fatores de SES relacionados ao risco de PTB são as técnicas de agrupamento de dados (ou clusterização). A clusterização é uma linha da aprendizagem de máquina não-supervisionada que busca associar elementos a grupos sem que haja uma compreensão inicial dos dados. Para tanto, utiliza-se de algoritmos de distanciamento para julgar o quão próximos dois pontos estão e se estes devem pertencer ou não a um mesmo grupo (ou *cluster*). As técnicas de clusterização são utilizadas há décadas para análise científica em diversas áreas do conhecimento, como por exemplo na psicologia [28], na genética [29] e na geofísica [30].

A clusterização como meio de encontrar grupos mais vulneráveis ao risco de PTB é menos comum que os métodos estatísticos tradicionais. Todavia, sua aplicação já é observada em diversos estudos recentes. Em [31], o uso da clusterização espacial mostra uma possível relação entre morar próximo a

lixões e de PTB. Os trabalhos apresentados em [32], [33] mostram a geração de agrupamentos associando fatores hereditários e comportamentais ao risco de PTB. Já em [34] é apresentado um estudo que investiga a distribuição geográfica do risco de PTB em nível de pequenas áreas chamadas de "*census blocks*" da cidade de Paris, França.

Assim, o objetivo deste trabalho é estratificar o risco de PTB no Brasil a partir de fatores de SES. O processo de estratificação é realizado a partir de uma análise por clusterização com base em técnicas de aprendizagem de máquina não-supervisionada. A análise foi realizada a partir da combinação de três bases de dados (*datasets*) coletadas pelo Governo Federal do Brasil: o Sistema de Informações sobre Nascidos Vivos (SINASC) [35], que contém informações sobre gestação, parto, recém-nascidos e mães; o Cadastro Único (CADU) [36], contendo uma ampla variedade de dados socioeconômicos de brasileiras a nível de família e pessoal; e a estimativa populacional do IBGE [37].

A partir das bases de dados foi criado um novo *dataset* e uma nova métrica chamada de Taxa Municipal de Prematuridade (TMP), com os quais se faz uma análise, a nível municipal, buscando observar as relações entre os fatores de SES e o risco de PTB.

Os resultados mostram um estrato formado por 7 grupos (*clusters*) no qual 3 apresentam uma TMP abaixo da média nacional e 4 apresentam uma TMP acima da média nacional. Com base nos *clusters* encontrados foi possível mapear geograficamente municípios no Brasil em relação ao risco de PTB. Além disto, este artigo mostra a análise de alguns fatores de SES associado aos *clusters* encontrados. Assim, os resultados apresentados neste trabalho poderão contribuir para a elaboração de políticas mais eficientes e especializadas para o Sistema Único de Saúde (SUS).

II. Materiais e métodos

A. Base de dados

Como descrito na introdução, foram utilizados três *datasets* para a geração do conjunto de treinamento, o SINASC, CADU e o IBGE. A análise realizada neste trabalho foi para os dados de 2018 em todos os três *datasets*.

O *dataset* SINASC, caracterizado aqui pela variável $\mathbf{T}_{\mathrm{SN}}$, é uma base formada por 61 atributos e quase 3 milhões de amostras, ela armazena dados associados aos nascimentos ocorridos em território brasileiro e pode ser encontrada no site DATASUS [35]. Para a proposta deste trabalho, foram utilizadas duas colunas de $\mathbf{T}_{\mathrm{SN}}$, relativas à duração da gestação e à residência da mãe, conforme descrito na Tabela I.

Tabela I: Variáveis do *dataset* SINASC, $\mathbf{T}_{\mathrm{SN}}$.

SINASC ($\mathbf{T_{SN}}$)	
Indexador	Código do Município de Residência da Mãe
Selecionadas	Semanas de Gestação
Descartadas	59 outras

O *dataset* CADU foi dividida em dois *datasets* distintos, chamados aqui de CADU Pessoa, expresso pela variável $\mathbf{T}_{\mathrm{P}}$, e CADU Família, expresso pela variável $\mathbf{T}_{\mathrm{F}}$.

O *dataset* $\mathbf{T}_{\mathrm{P}}$, possui mais de 12 milhões de amostras de cidadãos brasileiros com 26 atributos cada, contendo desde informações pessoais como sexo, idade, e raça, até informações mais específicas sobre educação e empregabilidade, como descrito na Tabela II.

Tabela II: Variáveis do *dataset* CADU Pessoa, $\mathbf{T}_{\mathrm{P}}$.

CADU Pessoa ($\mathbf{T_P}$)	
Indexador	ID Pessoa, ID Família
Selecionadas	Sexo, Idade, Raça, Código do Município de Residência, Local de Nascimento, Deficiência, Alfabetização, Tipo de Escola, Nível Escolar, Situação de Emprego, Tipo de Emprego, Remuneração, Valor de Benefícios Sociais
Descartadas	Grau de parentesco com Responsável da Família, Informações Regionais

Para o *dataset* $\mathbf{T}_{\mathrm{F}}$, existem cerca de 4 milhões de amostras de famílias com 23 atributos cada, contendo informações sobre condições de moradia, renda familiar, ou tipo de família, como detalhado na Tabela III.

Tabela III: Variáveis do *dataset*, $\mathbf{T}_{\mathrm{F}}$.

CADU Família ($\mathbf{T_F}$)	
Indexador	ID Família
Selecionadas	Propriedade do Domicílio, Quantidade de Cômodos, Material das Paredes, Material do Piso, Abastecimento de Água, Escoamento Sanitário, Coleta de Lixo, Iluminação, Calçamento, Classificação de Grupos Especiais, Renda Familiar Média
Descartadas	Data de Cadastramento, de Alteração, de Atualização, Código EAS/MS, Código CRAS/CREAS

A base de população usada é a estimativa da população brasileira do IBGE para o ano de 2018, que contém 5.570 amostras de municípios. Representada aqui pelo *dataset* $\mathbf{T}_{\mathrm{IBGE}}$, esta base possui 5 colunas, descritas na Tabela IV, das quais apenas as colunas referentes à população e ao código de município foram utilizadas.

Tabela IV: Variáveis do *dataset* IBGE, $\mathbf{T}_{\mathrm{IBGE}}$.

Estimativa de População do IBGE ($\mathbf{T_{IBGE}}$)	
Indexador	Código do Município
Selecionadas	População
Descartadas	Nome do Município, UF, Nome da UF

As fontes e dimensões de cada *dataset* para o ano de 2018 pode ser conferida na Tabela V.

Tabela V: Resumo dos *Datasets* utilizados.

Dataset	Base	Ano	Amostras	Variáveis
$\mathbf{T}_{\mathrm{SN}}$	SINASC	2018	2.944.932	61
$\mathbf{T}_{\mathrm{IBGE}}$	IBGE	2018	5.570	5
$\mathbf{T}_{\mathrm{P}}$	CADU (pessoa)	2018	12.852.599	26
$\mathbf{T}_{\mathrm{F}}$	CADU (família)	2018	4.807.996	23

B. Pré-processamento

Com o objetivo de unir todas as informações desejadas em um único *dataset* que possa servir de entrada para um algoritmo de clusterização, foi realizada uma etapa de pré-processamento das quatro bases, subdividida nos processos P_1, em que são processados os *datasets* $\mathbf{T}_{\text{SN}}$ e $\mathbf{T}_{\text{IBGE}}$, e P_2, que trata dos *datasets* $\mathbf{T}_{\text{P}}$ e $\mathbf{T}_{\text{F}}$. Os processos tem como saídas os *datasets* intermediários $\mathbf{I}_1$ e $\mathbf{I}_2$, respectivamente. Os *datasets* intermediários $\mathbf{I}_1$ e $\mathbf{I}_2$, possuem 5.570 amostras cada, referentes aos municípios. Os dois *datasets* intermediárias são unidos utilizando como chave o Código do Município, gerando $\mathbf{A}_0$. A Figura 1 detalha o esquema geral do pré-processamento.

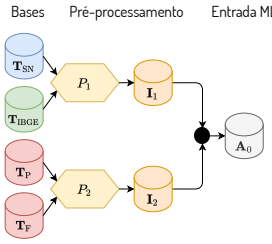


Figura 1: Esquema geral de pré-processamento para geração do *dataset* $\mathbf{A}_0$ (entrada do processo de aprendizagem de máquina), incluindo os pré-processamentos P_1 e P_2, suas saídas $\mathbf{I}_1$ e $\mathbf{I}_2$, e os *datasets* originais $\mathbf{T}_{\text{SN}}$, $\mathbf{T}_{\text{IBGE}}$, $\mathbf{T}_{\text{P}}$ e $\mathbf{T}_{\text{F}}$.

1) Pré-processamento P_1: Como descrito na Figura 1, o *dataset* SINASC, caracterizado pela variável $\mathbf{T}_{\text{SN}}$, foi pré-processado para obter os números de nascimentos prematuros a nível de município.

Primeiramente, o *dataset* $\mathbf{T}_{\text{SN}}$ foi filtrado pelas semanas de gestação, mantendo somente os casos de nascimentos com menos de 37 semanas. Em seguida, foram agrupados os valores pelo município de residência da mãe, contabilizando o número de amostras de nascimentos prematuros existentes para cada município.

Em seguida, através de união com a base de estimativa populacional do IBGE ($\mathbf{T}_{\text{IBGE}}$), utilizando o Código do Município, foi adicionada à $\mathbf{T}_{\text{SN}}$ a informação de população dos municípios, gerando o *dataset* intermediário $\mathbf{I}_1$.

No $\mathbf{I}_1$, foi então calculada a Taxa Municipal de Prematuridade, TMP. A TMP é a métrica proposta neste trabalho para mensurar a frequência dos nascimentos prematuros por município, expressa como:

$$\text{TMP} = \frac{N_{\text{NP}}}{N_{\text{P}}} \tag{1}$$

onde N_{NP} é o número de nascimentos prematuros ocorridos em um município e N_{P} é a população deste mesmo município.

Foi tomada a decisão de utilizar a população – em vez do número total de nascimentos, que pode ser obtido pelo *dataset* $\mathbf{T}_{\text{SN}}$ – devido à observação de percentuais municipais de prematuridade muito acima de valores realistas: um município estava com 70% de nascimentos prematuros, por exemplo. Mesmo esses valores extremos sendo removidos mais à frente no processamento P_3, foi utilizada a população total para tentar reduzir a chance desses valores desbalanceados afetaram os municípios de percentual de ocorrências de parto prematuro mediana, assumindo a possibilidade de dados faltantes em $\mathbf{T}_{\text{SN}}$.

A saída do pré-processamento P_1 é o *dataset* intermediário $\mathbf{I}_1$ contendo somente duas colunas: código do município e TMP.

2) Pré-processamento P_2: Como descrito na Figura 1, os *datasets* CADU Pessoa, expresso pela variável $\mathbf{T}_{\text{P}}$, e CADU família, caracterizada pela variável $\mathbf{T}_{\text{F}}$, foram pré-processados objetivando uma mudança principalmente nas colunas categóricas, de forma que essas sejam convertidas em variáveis numéricas, aptas a serem utilizadas pelos algoritmos de clusterização escolhidos.

Para o *dataset* $\mathbf{T}_{\text{P}}$, primeiro foi realizada uma filtragem, tendo sido removidas as pessoas do sexo masculino, assim como mulheres com idade inferior a 14 ou superior a 40 anos, como forma de excluir as faixas reconhecidamente consideradas de menor fertilidade [38]. Em seguida, foi aplicada a técnica de *one-hot encoding* em todas as colunas categóricas, gerando 29 colunas binárias. Finalmente, para o caso das variáveis educacionais foi feito um processamento adicional para unir diferentes colunas referentes a um mesmo nível de ensino. No caso do *dataset* $\mathbf{T}_{\text{F}}$, que representa os dados a nível familiar, todas as colunas categóricas foram transformadas pelo *one-hot encoding*, o que resultou em 48 novas colunas binárias. Os dois *datasets* foram unidos através da coluna ID Família, presente em ambos, adicionando a cada amostra de $\mathbf{T}_{\text{P}}$ os dados de sua família presentes em $\mathbf{T}_{\text{F}}$. Ao final do processo, os valores foram agrupados pelo município de residência, calculando o valor médio de cada coluna para cada município. Assim, a saída do pré-processamento P_2 gerou um *dataset* intermediário chamado de $\mathbf{I}_2$.

3) Saída $\mathbf{A}_0$: Para gerar a saída do pré-processamento geral, indicada pelo *dataset* $\mathbf{A}_0$, foram combinadas as saídas do processo P_1, o *dataset* $\mathbf{I}_1$ e a saída do processo P_2 o *dataset* $\mathbf{I}_2$, pelo Código do Município. O *dataset* $\mathbf{A}_0$ ficou com 5529 amostras e 104 variáveis, cada amostra representa um município brasileiro.

C. Metodologia

A metodologia de geração dos *clusters* finais seguiu a sequência de passos observada no diagrama ilustrado na Figura 2. No qual o *dataset* $\mathbf{A}_0$, foi transformado em *dataset* intermediário chamado de $\mathbf{A}_{\text{RN}}$ através do pré-processamento P_3. Em seguida, o *dataset* intermediário $\mathbf{A}_1$ é processado pelo bloco de processamento chamado de múltiplos k-means, identificado por MkM. O MkM gera uma matriz de centroides, caracterizado aqui pela variável $\mathbf{C}_0$. A matriz $\mathbf{C}_0$ passa pelo pré-processamento P_4, e gera uma matriz chamada de $\mathbf{C}_{\text{RN}}$. Finalmente, a matriz $\mathbf{C}_{\text{RN}}$ é processado pelo algoritmo *density-based spatial clustering of applications with noise* (DBSCAN) com os agrupamentos (*clusters*) finais encontrados. O módulo de processamento associado DBSCAN é chamado neste trabalho com DBS.

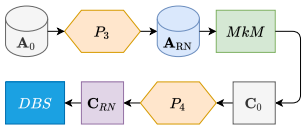


Figura 2: Diagrama de sequência do processo de clusterização, incluindo os dois algoritmos empregados MkM e DBS, os pré-processamentos P_3 e P_4, e os *datasets* intermediários gerados por cada passo do processo.

1) Pré-processamento P_3: Como apresentado na Figura 3, pré-processamento P_3 é caracterizado pela remoção de *outliers*, redução de dimensionalidade e normalização. Inicialmente os municípios considerados *outliers* com relação aos valores de TMP, isto é, aqueles cuja TMP é mais que três desvios-padrão acima ou abaixo da média geral, foram removidos do *dataset* $\mathbf{A}_0$, gerando um *dataset* $\mathbf{A}_1$. A coluna referente aos valores da TMP foi removida de $\mathbf{A}_1$ e suas informações armazenadas separadamente para uso em operações posteriores. O $\mathbf{A}_1$ possui 103 colunas e considerando a dificuldade natural da otimização dos *clusters* em alta dimensionalidade, foi aplicada uma redução de dimensionalidade utilizando o algoritmo *principal component analysis* (PCA), gerando um *dataset* reduzido chamado de $\mathbf{A}_R$ que mantém 95% da variância original de $\mathbf{A}_1$ em 58 colunas. Em seguida o *dataset* A_R foi normalizado através de 3 técnicas aplicadas em cascata: (1) Transformação de Yeo-Johnson, aproximando a distribuição das dimensões de uma distribuição normal; (2) normalização amostral L2, as mostras foram rebalanceadas individualmente de forma a capturar pontos de maior e menor destaque em cada; (3) e por fim, normalização de 0 a 1 por coluna. As normalizações geram o *dataset* $\mathbf{A}_{RN}$ de entrada do k-means (MkM).

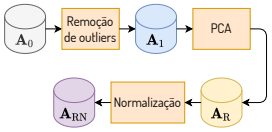


Figura 3: Diagrama detalhado do Pré-processamento P_3.

2) Múltiplos k-means (MkM): Com objetivo de encontrar centros representativos associados ao problema, foi proposta uma estratégia de processamento chamada aqui de MkM. O MkM é caracterizado por grupo de N modelos de k-means que são executados para o mesmo conjunto de entradas, o *dataset* $\mathbf{A}_{RN}$. Cada i-ésimo modelo de k-means, chamado daqui de kM_i, é executado com um específico número de centros, Nc_i, inicializados de forma aleatória. Ao final, é gerada uma matriz expressa como

$$\mathbf{C}_0 = \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_N^T \end{bmatrix} \tag{2}$$

onde cada i-ésimo G_i, é caracterizado por um conjunto de Nc_i centros associado a cada i-ésimo modelo kM_i e é expresso como

$$\mathbf{G}_i = [\mathbf{c}_{i,1}, \ldots, \mathbf{c}_{i,Nc_i}] \tag{3}$$

onde $\mathbf{c}_{i,j}$ é o j-ésimo centro do i-ésimo modelo kM_i e é expresso como

$$\mathbf{c}_{i,j} = [c_{i,j,1}, \ldots, c_{i,j,H}] \tag{4}$$

onde H representa o número total de colunas do *dataset* de entrada, no qual neste trabalho $H = 58$, como foi detalhado na subseção anterior. Assim, a matriz $\mathbf{C}_0$ pode ser re-escrita como

$$\mathbf{C}_0 = \begin{bmatrix} \mathbf{c}_{1,1} \\ \vdots \\ \mathbf{c}_{1,Nc_1} \\ \vdots \\ \mathbf{c}_{N,1} \\ \vdots \\ \mathbf{c}_{N,Nc_N} \end{bmatrix} = \begin{bmatrix} c_{1,1,1} & \cdots & c_{1,1,H} \\ \vdots & \ddots & \vdots \\ c_{1,Nc_1,1} & \cdots & c_{1,Nc_1,H} \\ \vdots & \ddots & \vdots \\ c_{N,1,1} & \cdots & c_{N,1,H} \\ \vdots & \ddots & \vdots \\ c_{N,Nc_N,1} & \cdots & c_{N,Nc_N,H} \end{bmatrix}. \tag{5}$$

O número de linhas da matriz $\mathbf{C}_0$ pode ser caracterizado como

$$L = \sum_{i=1}^{N} Nc_i. \tag{6}$$

Para este trabalho, foi utilizado $N = 290$, ou seja, foram executados 290 modelos de k-means para o *dataset* $\mathbf{A}_{RN}$. A quantidade de centros variou de 2 a 30, ou seja, $Nc_i \in \{2, \ldots, 30\}$ (29 quantidades diferentes de centros) onde para cada quantidade foram executadas 10 realizações, totalizando $N = 290$ modelos. Cada i-ésimo modelo, kM_i, foi otimizado com a técnica de maximização de expectativa, com tolerância para convergência de 10^{-7} e número máximo de 10.000 iterações. O resultado dos $N = 290$ modelos de k-means se agrupam em uma matriz $\mathbf{C}_0$ com $L = 4640$ linhas, representando todos centros encontrados. É importante observar que cada centro da $\mathbf{C}_0$ representa um possível *cluster* associado a um n-ésimo modelo gerado pelo MkM.

Além da matriz $\mathbf{C}_0$ é criado também um *dataset* intermediário chamado de $\mathbf{A}_C$ formado por 5529 amostras (linhas) e $N+1$ variáveis (colunas). Cada amostra (ou linha) representa um município e as variáveis (as colunas) são o valor de TMP de cada município e os *clusters* que cada amostra (ou município) está associadao em cada n-ésimo modelo de k-means encontrado pelo MkM. Assim, pode-se afirmar que cada município, em cada i-ésima linha, pertence a um k-esimo *cluster* em cada n-ésima coluna de $\mathbf{A}_C$. Cada n-ésimo modelo de k-means gerado pelo MkM possuie N_c *clusters* que agrupa um conjunto de B municípios.

3) Pré-processamento P_4: No pré-processamento P_4, são filtrados da matriz $\mathbf{C}_0$ os centros que representam *clusters* considerados "*clusters* de interesse", isto é, aqueles que apresentaram uma TMP média superior ou inferior à TMP média nacional em mais de 10%. Os dados associados aos *clusters*

são obtidos do *dataset* $\mathbf{A}_{\mathrm{C}}$ e a TMP média pode ser calculada como

$$TMP_{\text{media}} = \frac{1}{B}\sum_{i=1}^{B} TMP_i \tag{7}$$

onde TMP_i é a TMP referente ao i-ésimo município e B é o número total de municípios de um dado *cluster*. Para o caso do cálculo da TMP média nacional, o *cluster* de municípios são todos os municípios do Brasil presentes nos *datasets* já trabalhados, e a variável B na equação 7 equivale ao número total deles, sendo neste caso $B = 5529$. Após a filtragem em $\mathbf{C}_0$, uma nova matriz é gerada apenas com os centros que representam os "*clusters* de interesse", matriz esta chamada aqui de $\mathbf{C}_{ci}$.

Objetivando agrupar os *clusters* de interesse, $\mathbf{C}_{ci}$, a partir das várias execuções do k-means realizadas pelo MkM, foi calculada matriz correlação entre as amostras de $\mathbf{C}_0$. A ideia é trabalhar com a similaridade entre os centros, que representam os *clusters* de interesse, a fim de facilitar o agrupamento nas próxima etapas. A matriz de correlação está caracterizada pela variável $\mathbf{C}_1$.

Objetivando a redução de dimensionalidade da matriz de correlação $\mathbf{C}_1$, o algoritmo do PCA foi aplicado em $\mathbf{C}_1$, gerando uma matriz reduzida chamada aqui de $\mathbf{C}_R$. A matriz $\mathbf{C}_R$ é composta por 4 colunas que mantém 99% da variância de $\mathbf{C}_1$. Em seguida foi feita uma normalização dos valores de cada coluna entre 0 e 1, gerando então a matriz $\mathbf{C}_{RN}$. A Figura 4 detalha o Pré-processamento P_4.

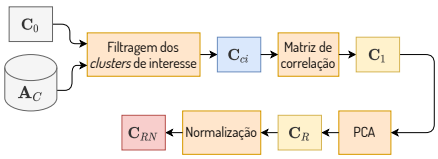


Figura 4: Diagrama detalhado do Pré-processamento P_4.

4) DBSCAN (DBS): O DBSCAN recebe como entrada a matriz $\mathbf{C}_{RN}$ que contém a similaridade amostral. Diferentemente do k-means, o DBSCAN não precisa de um número alvo fixo de centros (*clusters*), e sua funcionalidade é definida unicamente pelos ajustes de dois parâmetros: distância euclidiana mínima entre os pontos (ϵ) e número mínimo de pontos por *cluster*. Como saída, DBS vai gerar a classificação dos *clusters* encontrados a partir dos centros gerados pelo MkM, agrupando aqueles que podem ser tratados como um só *cluster*, e descartando os *clusters* menos frequentes (casuais).

Para o DBS, o número mínimo de ocorrências definido foi de 50, e $\epsilon = 0{,}06$. Como não há uma forma clara e objetiva para validar a qualidade dos *clusters* descobertos, foram utilizadas técnicas de visualização de dados para verificar a consistência dos *clusters*. Foram geradas representações cartográficas dos municípios de cada *cluster* e também visualizações de $\mathbf{C}_{RN}$ pós-classificação em 2D (utilizando t-SNE). Buscou-se ajustar os parâmetros de forma a evitar muitas fragmentações (sobre-ajuste) ou regiões sem qualquer *cluster* validado (sub-ajuste).

III. Resultados

Após a realização dos pré-processamentos P_1, P_2 e P_3, a aplicação do MkM resultou num total de 1337 "*clusters* de interesse". A quantidade de *clusters* de interesse descobertos cresceu conforme o número de *clusters* total definido nos modelos de k-Means, Nc_i, como pode ser observado na Figura 5, em que os primeiros casos aparecem quando o número de centros (*clusters*) de entrada, Nc_i, é igual a 5, chegando a encontrar cerca de 90 *clusters* de interesse para as entradas, Nc_i, mais altas (27 a 30).

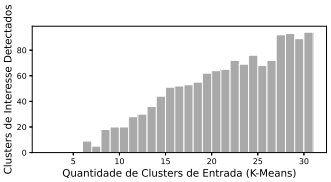


Figura 5: *Clusters* encontrados em MkM por quantidade de *clusters* de entrada.

A matriz de correlação, $\mathbf{C}_1$, e seu versão reordenada podem ser observadas, respectivamente, nos itens (a) e (b) da Figura 6. É possível observar possíveis padrões de *clusters*. Destaca-se no item (c) da Figura os *clusters* das respectivas amostras descobertos ao final do processo, permitindo uma comparação visual do resultado do DBS com um algoritmo de cálculo de distanciamento.

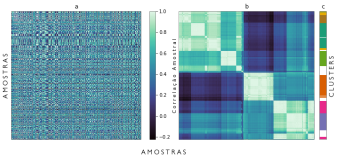


Figura 6: (a) Matriz de correlação (b) Matriz de correlação reordenada por distâncias entre amostras (c) Classificação das amostras reordenadas pós-DBSCAN

Após a aplicação do pré-processamento P_4, foi realizada a clusterização secundária através do DBS. O DBS encontrou então 7 *clusters* finais, divididos em 4 *clusters* de alta TMP (Taxa Municipal de Prematuridade) e 3 *clusters* de baixa TMP. Ainda na Figura 7, é possível observar, no item (a), uma estagnação, ou até diminuição na identificação de *clusters* válidos para os valores mais altos de quantidade de *clusters* de entrada quando comparado aos valores medianos. No item (b), observa-se os mesmos *clusters* identificados, mas agora não por tipo (alta e baixa TMP), e sim pelos *clusters* únicos encontrados.

Foi calculada a distribuição da TMP dos sub-clusters (clusters de interesse) de cada um dos *clusters* finais. Esta distribuição é observada na Figura 8, onde cada *cluster* está representado no eixo x, as distribuições no eixo y, e a média municipal nacional da TMP é indicada visualmente (aproximadamente $1{,}4 \times 10^{-4}$). Observa-se que quase todos os *clusters* validados tem sua TMP centroide variando numa

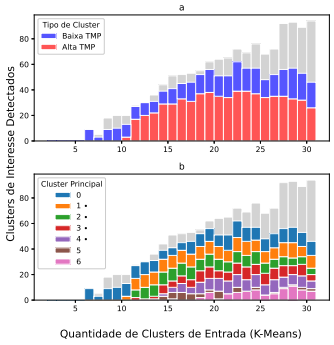


Figura 7: Clusters encontrados por época em MkM. (a) por tipo de *cluster* (b) por *cluster*. O símbolo (•) indica *cluster* com alta TMP

amplitude de 1×10^{-4} a $2{,}5 \times 10^{-4}$ em relação à média dos municípios de todo país, excetuando-se apenas o *cluster* **1**, com TMP centroide superior à média em quase 8×10^{-4} unidades.

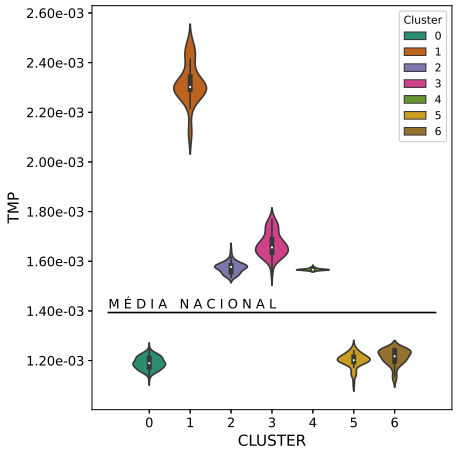


Figura 8: Distribuição de TMP dos *clusters* finais.

Também foi observada a presença geográfica desses *clusters*, isto é, quais municípios pertencem a quais *clusters*. Através das visualizações é possível contextualizar os *clusters* encontrados, assim como validar os mesmos. Como a entrada do problema foram dados sociais, era esperado que alguns ou todos os *clusters* possuíssem alguma concentração geográfica. Para permitir essa avaliação, foram geradas três visualizações do mapa do Brasil.

A primeira visualização - de valores binários - é observada na Figura 9, tendo sido gerada utilizando o tipo de *cluster* (alta ou baixa TMP). Foi contada a quantidade de vezes que cada município do banco de dados foi classificado num *cluster* de alta ou baixa TMP, e o município foi classificado no mapa de acordo com o tipo mais frequente.

A segunda visualização foi gerada a partir da subtração do total de vezes em que um município foi classificado em algum *cluster* com alta TMP pelo total de vezes em que ele foi classificado em *cluster* de baixa TMP, desta forma

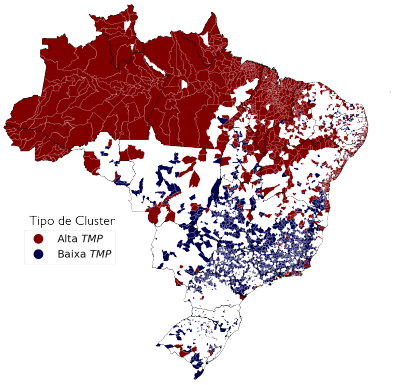


Figura 9: Municípios por tipo de *cluster* mais comum (baixa ou alta TMP).

obtendo também um grau de intensidade ou pertencimento do município em relação a um tipo de *cluster*, o que pode ser visto na Figura 10.

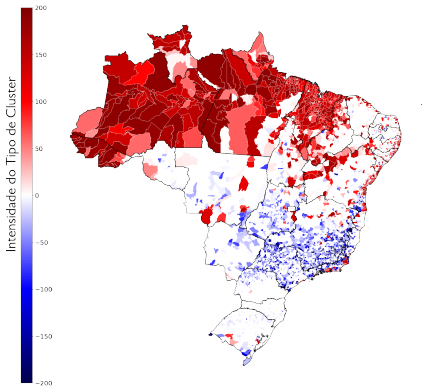


Figura 10: Municípios por diferença na quantidade de vezes classificado num tipo de *cluster*. Valores negativos (■ azuis) indicam que o município foi mais vezes classificado no tipo Baixa TMP, valores positivos (■ vermelhos) indicam mais vezes em Alta TMP.

A terceira visualização, observada na Figura 11, revela em qual dos 7 *clusters* finais cada município foi mais comumente classificado, permitindo a visualização dos núcleos regionais encontrados, bem como dos *clusters* com municípios mais dispersos.

Finalmente, foram extraídos a partir dos *clusters* validados, seus núcleos, contendo as informações da média e da variância observada para cada um dos 10 *clusters* encontrados. Com essa informação é possível observar as características de cada um dos *clusters*.

IV. Análise dos Resultados

Em relação ao funcionamento dos modelos de clusterização, observa-se primeiramente que o filtro secundário DBS funcionou conforme o esperado, selecionando as amostras similares

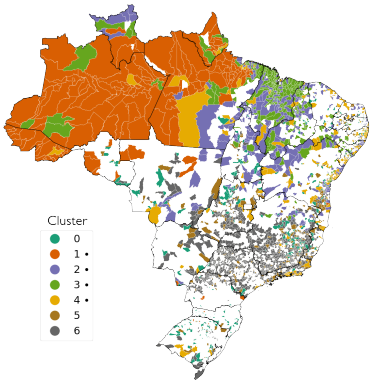


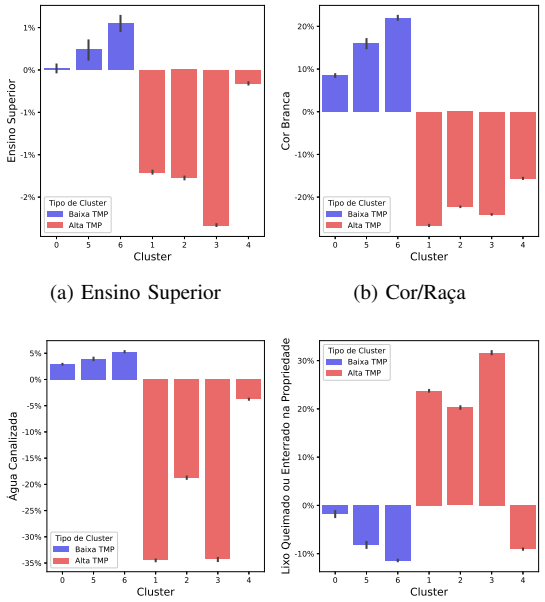
Figura 11: Municípios por *cluster* mais comum. O símbolo (•) indica *cluster* de alta TMP.

e eliminando o ruído derivado do "sobreajuste" advindo do alto número de *clusters* de entrada em algumas unidades do bloco MkM. No item (c) da Figura 6, pode-se verificar o filtro ao ver como algumas faixas/linhas não foram selecionadas a nenhum *cluster* final. Além disso, observando a Figura 11, é possível ver o caráter regional da maioria dos *clusters* encontrados, que ocupam uma zona específica do mapa, com alguns poucos municípios de exceção. É esperado que municípios próximos tenham características sociais mais parecidas, portanto a regionalidade observada no mapa é um forte indício de que a clusterização, apesar de tantas transformações ocorridas entre os conjuntos de dados iniciais até a aplicação do DBS, foi bem sucedida em traduzir e processar os dados do problema.

Através das Figuras 9 e 10, é possível observar um claro caráter regional nos *clusters* encontrados, com os *clusters* de alta TMP situando-se mais fortemente no Norte e Nordeste, e os de baixa TMP no Centro-Sul. No Nordeste, os *clusters* de alta TMP são mais presentes no estado do Maranhão e no Vale do Rio São Francisco. Os *clusters* de baixa TMP mais intensos são observados no estado de São Paulo e no sul de Minas Gerais. A região Norte está quase toda classificada em *clusters* de alta TMP, e como pode ser visto na Figura 11, o *cluster* predominante na região é o **1**, notadamente o de maior TMP.

Observando as características de cada *cluster* individualmente, é possível observar como é a relação entre os fatores sociais utilizados como entrada e a TMP. Na Figura 12, observa-se a diferença percentual do *cluster* em relação à média populacional para alguns desses fatores. Vê-se que, ao selecionar essas características (educação superior, etnia, abastecimento de água e destino do lixo), há uma clara distinção entre os valores para *clusters* de alta e baixa TMP.

O *cluster* **4** chama atenção por ser o que mais foge à regra. Observando a Figura 11, vê-se que este *cluster* é o mais disperso dentre os de alta TMP, e destaca-se pela quantidade de municípios litorâneos no Nordeste e no Rio de Janeiro. Na Figura 8, o mesmo *cluster* se revela como o de menor TMP entre os de taxa alta. Em contrapartida, os *clusters* **1**, **2** e **3**, de maior TMP, concentram-se no Norte e no interior do Nordeste.



(a) Ensino Superior (b) Cor/Raça

(c) Água Canalizada (d) Destino do Lixo

Figura 12: Características dos *clusters* finais.

V. Conclusões

Através da utilização conjunta de métodos de clusterização e redução de dimensionalidade baseados em aprendizagem não-supervisionada, aplicados a dados socioeconômicos a nível de município, foi possível extrair informações importantes sobre *clusters* de municípios com alta e baixa ocorrência de PTB. Os resultados obtidos revelaram uma distinção socioeconômica clara entre os *clusters* com alto e baixo risco de PTB, com os *clusters* com alto risco de PTB ocupando predominantemente áreas com piores índices sociais. O PTB é um fenômeno complexo muti-fatorial e a busca por sua redução demanda análises de diversos aspectos capazes de influenciar em sua ocorrência. Aqui, foi observado que a qualidade de vida e a oferta de serviços públicos possivelmente afetam, positivamente, na redução da prematuridade, podendo e devendo ser incluídos dentre esses aspectos.

Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

Referências

[1] E. França, S. Lansky, M. Rego, D. Carvalho Malta, J. Santiago França, R. Teixeira, D. Porto, M. Almeida, M. D. F. Marinho de Souza, C. Szwarcwald, M. Mooney, M. Naghavi, and A. Vasconcelos, "Principais causas da mortalidade na infância no brasil, em 1990 e 2015: estimativas do estudo de carga global de doença," *Revista Brasileira de Epidemiologia*, vol. 20, pp. 46–60, 05 2017.

[2] B. Modell, R. Berry, C. A. Boyle, A. Christianson, M. Darlison, H. Dolk, C. P. Howson, P. Mastroiacovo, P. Mossey, and J. Rankin, "Global regional and national causes of child mortality," *The Lancet*, vol. 380, no. 9853, p. 1556, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673612618789

[3] W. H. Organization, "Born too soon: the global action report on preterm birth," p. 112 p., 2012.

[4] I. of Medicine, *Preterm Birth: Causes, Consequences, and Prevention*, R. E. Behrman and A. S. Butler, Eds. Washington, DC: The National Academies Press, 2007. [Online]. Available: https://www.nap.edu/catalog/11622/preterm-birth-causes-consequences-and-prevention

[5] K. Adhikari, S. B. Patten, T. Williamson, A. B. Patel, S. Premji, S. Tough, N. Letourneau, G. Giesbrecht, and A. Metcalfe, "Does neighborhood socioeconomic status predict the risk of preterm birth? a community-based canadian cohort study," *BMJ open*, vol. 9, no. 2, p. e025341, 2019.

[6] I. Kawachi and L. F. Berkman, *Neighborhoods and health*. Oxford University Press, 2003.

[7] A. Metcalfe, P. Lail, W. A. Ghali, and R. S. Sauve, "The association between neighbourhoods and adverse birth outcomes: A systematic review and meta-analysis of multi-level studies," *Paediatric and perinatal epidemiology*, vol. 25, no. 3, pp. 236–245, 2011.

[8] N. Santoso and S. Wulandari, "Hybrid support vector machine to preterm birth prediction," *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 8, p. 191, 10 2018.

[9] T. Włodarczyk, S. Płotka, P. Rokita, N. Sochacki-Wójcicka, J. Wójcicki, M. Lipa, and T. Trzciński, "Spontaneous preterm birth prediction using convolutional neural networks," 2020.

[10] C. Catley, M. Frize, C. Walker, and D. Petriu, "Predicting high-risk preterm birth using artificial neural networks," *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 10, pp. 540–9, 08 2006.

[11] Y.-S. Kim, "Analysis of spontaneous preterm labor and birth and its major causes using artificial-neural-network," *Journal of Korean Medical Science*, vol. 34, 04 2019.

[12] J. Hill, M. Campbell, G. Zou, J. Challis, G. Reid, H. Chisaka, and A. Bocking, "Prediction of preterm birth in symptomatic women using decision tree modeling for biomarkers," *American journal of obstetrics and gynecology*, vol. 198, pp. 468.e1–7; discussion 468.e7, 05 2008.

[13] J. Lee, J. Cai, F. Li, and Z. Vesoulis, "Predicting mortality risk for preterm infants using random forest," *Scientific Reports*, vol. 11, p. 7308, 03 2021.

[14] E. DeFranco, M. Lian, L. Muglia, and M. Schootman, "Area-level poverty and preterm birth risk: A population-based multilevel analysis," *BMC public health*, vol. 8, p. 316, 10 2008.

[15] M. Buen, E. Amaral, R. Souza, R. Passini, G. Lajos, R. Tedesco, M. Nomura, T. Dias, P. Rehder, M. Sousa, and J. Cecatti, "Maternal work and spontaneous preterm birth: A multicenter observational study in brazil," *Scientific Reports*, vol. 10, 06 2020.

[16] M.-J. Saurel-Cubizolles, J. Zeitlin, N. Lelong, E. Papiernik, G. Renzo, and G. Bréart, "Employment, working conditions, and preterm birth: Results from the europop case-control survey," *Journal of epidemiology and community health*, vol. 58, pp. 395–401, 06 2004.

[17] J. Kaufman, F. Alonso, and P. Pino, "Multi-level modeling of social factors and preterm delivery in santiago de chile," *BMC Pregnancy and Childbirth*, vol. 8, pp. 46 – 46, 2008.

[18] K. Beeckman, S. Putte, K. Putman, and F. Louckx, "Predictive social factors in relation to preterm birth in a metropolitan region," *Acta obstetricia et gynecologica Scandinavica*, vol. 88, pp. 787–92, 06 2009.

[19] A. Grjibovski, L. Bygren, A. Yngve, and M. Sjostrom, "Large social disparities in spontaneous preterm birth rates in transitional russia," *Public health*, vol. 119, pp. 77–86, 03 2005.

[20] A. A. d. Oliveira, M. F. d. Almeida, Z. P. d. Silva, P. L. d. Assunção, A. M. R. Silva, H. G. d. Santos, and G. P. Alencar, "Fatores associados ao nascimento pré-termo: da regressão logística à modelagem com equações estruturais," *Cadernos de Saúde Pública*, vol. 35, 00 2019.

[21] M. Chen, N. Xie, Z. Liang, T. Qian, and D. Chen, "Early prediction model for preterm birth combining demographic characteristics and clinical characteristics," 11 2020.

[22] B. Alleman, A. Smith, H. Byers, B. Bedell, K. Ryckman, J. Murray, and K. Borowski, "A proposed method to predict preterm birth using clinical data, standard maternal serum screening, and cholesterol," *American journal of obstetrics and gynecology*, vol. 208, 03 2013.

[23] S. Sun, K. Weinberger, K. Spangler, M. Eliot, J. Braun, and G. Wellenius, "Ambient temperature and preterm birth: A retrospective study of 32 million us singleton births," *Environment International*, vol. 126, 02 2019.

[24] R. Granese, E. Gitto, G. D'Angelo, R. Falsaperla, G. Corsello, D. Amadore, G. Calagna, I. Fazzolari, R. Grasso, and O. Triolo, "Preterm birth: Seven-year retrospective study in a single centre population," *Italian Journal of Pediatrics*, vol. 45, 12 2019.

[25] J. Huang, Y. Qian, M. Gao, H. Ding, L. Zhang, and R. Jia, "Analysis of factors related to preterm birth: a retrospective study at nanjing maternity and child health care hospital in china," *Medicine*, vol. 99, p. e21172, 07 2020.

[26] K. Baker, W. Story, E. Walser-Kuntz, and M. B. Zimmerman, "Impact of social capital, harassment of women and girls, and water and sanitation access on premature birth and low infant birth weight in india," *PLoS ONE*, vol. 13, p. e0205345, 10 2018.

[27] M. Ruiz, P. Goldblatt, J. Morrison, L. Kukla, J. Švancara, M. Riitta-Järvelin, A. Taanila, M.-J. Saurel-Cubizolles, S. Lioret, C. Bakoula, A. Veltsista, D. Porta, F. Forastiere, M. Eijsden, T. Vrijkotte, M. Eggesbø, R. White, H. Barros, S. Correia, and H. Pikhart, "Mother's education and the risk of preterm and small for gestational age birth: A drivers meta-analysis of 12 european cohorts," *Journal of epidemiology and community health*, vol. 69, 04 2015.

[28] F. Borgen and D. Barnett, "Applying cluster analysis in counseling psychology research," *Journal of Counseling Psychology*, vol. 34, pp. 456–468, 10 1987.

[29] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of computational biology : a journal of computational molecular cell biology*, vol. 6, pp. 281–97, 08 1999.

[30] J. Sun and Y. Li, "Multidomain petrophysically constrained inversion and geology differentiation using guided fuzzy c-means clustering," *Geophysics*, vol. 80, pp. ID1–ID18, 07 2015.

[31] M. Istvan, F. Rouget, L. Michineau, C. Monfort, L. Multigner, and J.-F. Viel, "Landfills and preterm birth in the guadeloupe archipelago (french west indies): A spatial cluster analysis," *Tropical Medicine and Health*, vol. 47, 12 2019.

[32] R. Passini, Jr, J. G. Cecatti, G. J. Lajos, R. P. Tedesco, M. L. Nomura, T. Z. Dias, S. M. Haddad, P. M. Rehder, R. C. Pacagnella, M. L. Costa, M. H. Sousa, and for the Brazilian Multicentre Study on Preterm Birth study group, "Brazilian multicentre study on preterm birth (emip): Prevalence and factors associated with spontaneous preterm birth," *PLOS ONE*, vol. 9, no. 10, pp. 1–12, 10 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0109069

[33] M. Esplin, T. Manuck, B. Christensen, J. Biggio, R. Bukowski, S. Parry, H. Zhang, M. Varner, W. Andrews, G. Saade, Y. Sadovsky, U. Reddy, and J. Ilekis, "Cluster analysis of spontaneous preterm birth phenotypes identifies potential associations among preterm birth mechanisms," *American Journal of Obstetrics and Gynecology*, vol. 212, pp. S107–S108, 06 2015.

[34] S. Deguen, N. Ahlers, M. Gilles, A. Danzon, M. Carayol, D. Zmirou-Navier, and W. Kihal-Talantikite, "Using a clustering approach to investigate socio-environmental inequality in preterm birth—a study conducted at fine spatial scale in paris (france)," *International journal of environmental research and public health*, vol. 15, no. 9, p. 1895, 2018.

[35] Datasus, "Sinasc - sistema de informações de nascidos vivos." [Online]. Available: http://www2.datasus.gov.br/DATASUS/index.php?area=060702

[36] "Base desidentificada do cadastro Único com marcação do bolsa família." [Online]. Available: https://aplicacoes.mds.gov.br/sagi/portal/index.php?grupo=212

[37] IBGE, "Estimativas da população." [Online]. Available: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html

[38] E. C. W. Group, "Fertility and ageing," *Human Reproduction Update*, vol. 11, no. 3, pp. 261–276, 05 2005. [Online]. Available: https://doi.org/10.1093/humupd/dmi006