Relatório IBGE - Pareamento de Dados PNAD COVID19

Junho 2020

Em um mundo contemporâneo, orientado a dados, a existência de evidências heterogêneas tem se tornado cada vez mais importante. Devido ao recente problema da sociedade com a pandemia global do COVID 19[1] e a dificuldade dos encontros presenciais, o Instituto Brasileiro de Geografia e Estatística (IBGE) encara a necessidade de adaptação dos seus processos. Uma das soluções autoriza entrevista via telefone[2], entretanto as bases de dados atuais não possuem os registros de telefones individualizados dos domicílios.

O pareamento probabilístico de registros [2] é uma metodologia que visa comparar pares de registros, atributo por atributo estimando uma probabilidade (m e u) destes atributos serem equivalentes. Uma nota é atribuída para cada probabilidade, para cada atributo. A probabilidade final de um par de registros é dada pela soma das notas de cada atributo. Ao fim do processo, é estimada uma nota de corte mínima para um par ser considerado verdadeiro, ou seja, o mesmo registro. Este trabalho foi realizado utilizando esta metodologia e inspirado em trabalhos recentes realizados [3] [4]. O processo de pareamento probabilístico é dividido em 5 principais etapas: Limpeza e Padronização, Estimativa de Parâmetros, Blocagem, Processamento e Conferência Clerical.

Visando aprimorar e coletar números de telefones para os domicílios selecionados como parte da amostra da PNAD COVID19, o IBGE realizou o pareamento probabilístico gerando um cruzamento de dados entre as bases PNAD e Foram utilizadas seis bases de dados como fonte fornecedora de telefones: Oi, ANATEL, VIVO, Ministério da Economia, Ministério da Saúde e SERCOMTEL. A tabela 1 ilustra a quantidade de registros contida em cada uma das bases de dados.

Para executar o pareamento probabilístico são necessárias algumas escolhas por parte do analista de dados. É preciso definir quais atributos serão utilizados em comum entre as duas bases de dados. Para este pareamento foram selecionados os atributos: nome completo, primeiro nome, último nome e nome do meio, data de nascimento, município de residência, UF, sexo, logradouro e número do logradouro.

[1] Disponível em `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`

[2] Disponível em: `https://respondendo.ibge.gov.br/coleta-por-telefone.html`

Tabela 1: Quantidade de registros por *dataset*

Dataset	**#n**
PNAD Contínua 1º Trim 2019	576.214
OI	5.070.869
ANATEL	197.350.786
VIVO	3.225.434
MINISTÉRIO ECONOMIA	1.178.420
SERCOMTEL	127.924
MINISTÉRIO DA SAÚDE	89.429.865
TOTAL	296.959.512

Para cada uma das bases de dados forem coletados, quando presentes, o nome, data de nascimento, município de residência, UF, sexo, logradouro, número do logradouro e números de telefone, sendo esses campos passados por um processo de limpeza. Para os nomes de uma pessoa, foram removidos valores numéricos, caracteres especiais, acentos, múltiplos espaços, espaços no início ou final dos nomes, tabulações e palavras não representativas para a tarefa de pareamento, como "AUSENTE", "DADOS NAO FORNECIDOS", "A CONFIRMAR", entre outros nomes inválidos. Preposições como "DE", "DA", "DAS" também foram removidas. Para os municípios de residência e logradouros, o mesmo processo foi realizado. Neste último, ainda foram padronizados termos como "ED", "COND" e "R", além de palavras como "STA", "NA SA" e algumas variações de nomes de ruas. Os municípios de residência, quando codificados por um identificador, foram mapeados para os seus nomes correspondentes, em forma de *string*. Além disso, os nomes foram mapeados por fonética através do algoritmo de *soundex*. Por exemplo, "WA" e "VA" possuem o mesmo som, então todos os nomes que possuíam as vogais "WA" foram alterados para "VA" sem prejuízos nas comparações. As datas de nascimento, por sua vez, foram reescritas no formato DIA/MÊS/ANO e o sexo foi representado por meio de variáveis numéricas, para todas as bases, concluindo o processo de padronização. Este processo influencia diretamente na qualidade do pareamento.

Por fim, para os números de telefone, foram removidos os caracteres especiais e letras, além de números inválidos, como "999999999" e números contendo uma quantidade de dígitos inferior a 7. Os números de DDDs, quando existentes, foram concatenados aos valores de telefone correspondentes. Outros números com sequências numerais, e.g: 1111, 2222, 3333, 4444 foram invalidados. Telefones "0" também foram invalidados. Telefones frequentes como "2147483647" que correspondem a centrais do ministério da saúde ou do ministério da economia foram excluídos da base de dados.

Na etapa de estimativa de parâmetros são definidos os valores estatísticos necessários para a etapa de comparação dos atributos nos registros: a probabilidade M de concordância e a probabilidade U de discordância A probabilidade M é definida quando um par é considerado verdadeiro e o atributo deste par concorda nos registros comparados, enquanto a probabilidade U é definida quando um par é considerado verdadeiro e o atributo deste par discorda nos registros. Assim, com base, respectivamente, na concordância ou discordância nos registros, a soma de probabilidade é incrementada ou decrementada para todo par considerado verdadeiro.

Em um cenário perfeito, o ideal é comparar todos os registros da base de dados, mas isto não é possível computacionalmente. O custo desta operação é de $296.959.512^2$ comparações. Assim, as estratégias de blocagem são utilizadas para permitir a comparação de registros mais prováveis de pertencerem à mesma pessoa. Foram utilizadas três estratégias de agrupamento: (i) registros que possuem logradouro, município de residência e data de nascimento correspondentes; (ii) registros que possuem primeiro nome, último nome e UF correspondentes e (iii) registros que possuem primeiro nome, logradouro e data de nascimento equivalentes. Para os atributos logradouro, primeiro nome e último nome, registros com um mesmo som também foram considerados na tarefa de agrupamento.

O software utilizado para a execução do pareamento probabilístico, foi o cPareia, uma evolução do PAREIA [1] que utiliza uma versão *standalone* para execução paralela em apenas um servidor para grandes e múltiplas bases de dados. O servidor utilizado foi um Dell Xeon com 20 núcleos de processamento e 180GB de ram com discos rígidos de tecnologia SSD.

Foram gerados 761.587.510 pares de registros com 1.756.887 pares considerados verdadeiros Infelizmente o número real de pares verdadeiros é desconhecido neste cenário. Para se obter este número seria necessário realizar a comparação entre todos os registros da base de dados o que é computacionalmente inviável. Como consequência é impossível estimar a porcentagem de pares verdadeiros encontrados, e.g, acurácia. A estimativa de erro, para o falso positivo, foi calculada através da conferência, de até 20%. Este número é alto devido a qualidade do dado de origem com uma grande quantidade de dados ausentes.

Dos 576.214 registros contidos na PNAD Contínua 2019, 512.351 registros foram encontrados telefones. Desses, 492.898 foram identificados nas entrevistas, enquanto 512.351 foram identificados no pareamento de registros. Esta amostra da PNAD Contínua de 2019 foi utilizada como entrada dos dados da PNAD COVID19. As bases de dados foram geradas e enviadas para os respectivos setores do IBGE sob ordem de sigilo dos dados identificados e estão sendo utilizadas para a realização das entrevistas censitárias da PNAD COVID19 através do contato telefone.

Referências

[1] Walter dos Santos. Algoritmo paralelo e eficiente para o problema de pareamento de dados, 2008.

[2] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[3] Augusto Afonso Guerra Junior, Ramon Gonçalves Pereira, Eli Iola Gurgel, Mariangela Cherchiglia, Leonardo Vinicius Dias, Juliano Ávila Núbia Santos, Afonso Reis, Francisco Assis Acurcio, and Wagner Meira Junior. Building the national database of health centred on the individual: Administrative and epidemiological record linkage - brazil, 2000-2015. In *International Journal of Population Data Science*, Nov 2018.

[4] Robespierre Pita, Clicia Pinto, Pedro Melo, Malu Silva, Marcos Barreto, and Davide Rasella. A spark-based workflow for probabilistic record linkage of healthcare data. In *EDBT/ICDT Workshops*, pages 17–26, 2015.